Sous la direction de Jean-Pierre Ramis et André Warusfel
Xavier Buff · Josselin Garnier · François Moulin · Monique Ramis · Jacques Sauloy

Mathématiques

Tout-en-un pour la Licence 3

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, membre de l'Académie des Sciences, Inscriptions et Belles-Lettres de Toulouse, professeur émérite à l'Institut de Mathématique de Toulouse (Université Paul Sabatier), a été directeur de l'Institut de Recherches Mathématiques Avancées de Strasbourg et de l'Institut de Mathématiques de Toulouse.

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

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

Josselin Garnier, ancien élève de l'École normale supérieure de la rue d'Ulm, professeur à l'Université Paris Diderot, Laboratoire de Probabilités et Modèles Aléatoires & Laboratoire Jacques-Louis Lions.

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.

Les éditions Dunod remercient Jean-Marie Monier, professeur de mathématiques en classes préparatoires au lycée La Martinière-Monplaisir (Lyon), pour sa relecture attentive de l'ouvrage.

Graphisme de couverture : Elizabeth Riba
Illustration de couverture : © morgenstjerne – Shutterstock.com

© Dunod, 2015, 2021
11 rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-083174-6

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 (350) qui

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

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

Table des matières

Préface iii

Avant-propos xi

I Algèbre

I.1 Arithmétique 3

1 Divisibilité dans un anneau commutatif 4
1.1 Anneaux euclidiens et anneaux principaux 4
1.2 Anneaux principaux et anneaux factoriels 7
1.3 Polynômes sur un anneau factoriel 12
2 Le groupe multiplicatif de l'anneau $\mathbb{Z}/n\mathbb{Z}$ 13
2.1 Rappels sur $(\mathbb{Z}/n\mathbb{Z})^*$ 13
2.2 Le groupe multiplicatif d'un corps fini 14
2.3 Le groupe multiplicatif de l'anneau $\mathbb{Z}/p^r\mathbb{Z}$ 15
2.4 Deux applications informatiques 16
3 Résidus quadratiques 21
3.1 Carrés dans un corps fini 21
3.2 Symbole de Legendre 22
3.3 Loi de réciprocité quadratique 24
3.4 Symbole de Jacobi 27
4 Sommes de deux carrés 29
4.1 Rappels sur l'anneau des entiers de Gauß 29
4.2 Sommes de deux carrés : théorème de Fermat et Euler 33
4.3 Sommes de trois et de quatre carrés 35
5 Nombres premiers, critères de primalité 37
5.1 Aspects pratiques 39
5.2 Mauvaises méthodes 42
5.3 Bonnes méthodes probabilistes : Solovay-Strassen 45
5.4 Bonnes méthodes déterministes 45
5.5 Deux classes spéciales de nombres premiers 50
6 Fractions continues 53
6.1 Compléments à l'algorithme d'Euclide 53
6.2 Développement en fraction continue dans $\mathbb{Q}$ et dans $K(X)$ 56
6.3 Les réduites d'une fraction continue sur un corps arbitraire 58
6.4 Développement en fraction continue d'un réel 61
6.5 Développement en fraction continue d'une série formelle 64
Exercices 66

I.2 Compléments sur les formes bilinéaires alternées et sur les formes quadratiques 75

1 Compléments sur les formes bilinéaires alternées 76
1.1 Sous-espaces isotropes, involutifs, Lagrangiens 79
1.2 Classification des formes bilinéaires alternées 80
1.3 Pfaffien 82
1.4 Introduction au groupe symplectique 86
2 Compléments sur les formes quadratiques 88
2.1 Rappels et compléments 90
2.2 Discriminant d'une forme quadratique 93
2.3 Classification des formes quadratiques sur un corps fini F_q 97
2.4 Espaces hyperboliques — Théorème de décomposition de Witt 99
2.5 Introduction à l'étude des formes quadratiques sur $\mathbb{Q}$ 111
Exercices 116

II Géométrie

II.1 Surfaces 127

1 Nappes paramétrées 128
1.1 Définitions 128
1.2 Nappes géométriques 131
1.3 Plan tangent, espace tangent 133
1.4 Position par rapport au plan tangent 137
2 Surfaces implicites 144
2.1 Définitions 144
2.2 Sous-variétés lisses 145
2.3 Espace et plan tangent 146
2.4 Intersection de deux surfaces 148
3 Exemples 151
3.1 Nappes réglées 151
3.2 Nappes de révolution 153
3.3 Quadriques 156
Exercices 165

III Analyse

III.1 Intégration 173

1 Initiation aux intégrales multiples 175
1.1 Intégration sur un pavé 176
1.2 Intégration sur un ensemble cubable 195
1.3 Intégrales itérées et théorème de Fubini 207
1.4 Formule de changement de variables 215
1.5 Intégrales multiples généralisées 225
2 L'intégrale de Henstock-Kurzweil 245
2.1 Intégrale de Henstock-Kurzweil sur un segment 246

2.2 Le théorème fondamental de l'analyse . . . 267
2.3 Intégrale de Henstock-Kurzweil sur un intervalle quelconque . . . 273
2.4 Le lemme de Henstock . . . 279
2.5 Lemme de Vitali et différentiabilité des intégrales indéfinies . . . 283
2.6 Fonctions absolument intégrables . . . 286
2.7 Les théorèmes de convergence . . . 294
2.8 Fonction définie par une intégrale : continuité et dérivabilité . . . 305
2.9 Intégrale de Henstock-Kurzweil des fonctions à valeurs dans un espace vectoriel de dimension finie . . . 307
3 Intégrale de Henstock-Kurzweil et mesure de Lebesgue . . . 310
3.1 Mesure de Lebesgue — Ensembles négligeables . . . 311
3.2 Ensembles et fonctions mesurables . . . 321
3.3 Espaces L^1 et L^2 . . . 334
4 Intégrales multiples au sens de Henstock-Kurzweil . . . 341
4.1 Définition et premières propriétés . . . 341
4.2 Théorème de Fubini . . . 345
4.3 Intégrales sur un ouvert de $\mathbb{R}^n$ . . . 347
4.4 Formule de changement de variables . . . 353
Exercices . . . 357

III.2 Introduction aux équations aux dérivées partielles 373
1 Généralités . . . 373
1.1 Définitions et premiers exemples . . . 373
1.2 Problèmes bien posés . . . 376
1.3 Équations linéaires . . . 382
2 Équations d'ordre 1 . . . 383
2.1 Équations de transport . . . 383
2.2 Méthode des caractéristiques . . . 387
2.3 Le problème de Cauchy en dimension 2 . . . 391
3 Équations linéaires d'ordre 2 . . . 397
3.1 Caractéristiques et classification des équations . . . 397
3.2 Réduction aux formes standards . . . 402
4 Équations linéaires d'ordre 2 à coefficients constants . . . 408
4.1 L'équation des ondes uni-dimensionnelle . . . 408
4.2 Méthode de séparation de variables . . . 414
4.3 L'équation des cordes . . . 415
4.4 L'équation de la chaleur uni-dimensionnelle . . . 422
Exercices . . . 425

III.3 Constructions de R et introduction aux nombres p-adiques 429
1 Constructions à partir d'écritures décimales . . . 431
1.1 Écritures décimales . . . 431
1.2 Construction de $\mathbb{R}$ à partir des écritures décimales . . . 432
1.3 Écritures décimales à gauche — Nombres décadiques . . . 439
2 Introduction aux nombres p-adiques . . . 445
2.1 Anneaux b-adiques — Corps p-adiques — Présentation algébrique . . . 445
2.2 Distances et topologie. Suites et séries . . . 450
2.3 Zéros d'un polynôme dans un corps p-adique . . . 470
2.4 Nombres automorphes — Nombres b-adiques et nombres p-adiques . . . 474

3 Coupures et sections commençantes 479
3.1 Construction de $\mathbb{R}$ par les sections commençantes 479
3.2 Un théorème d'unicité 489
4 Complétions de corps et construction de $\mathbb{R}$ 491
4.1 Complétion d'un corps normé 493
4.2 Comparaison entre les anneaux $\mathbb{C}[X]$ et $\mathbb{Q}_p$ 507
4.3 Pour aller plus loin : extensions, clôtures algébriques et complétions . . . 508
5 Nombres, observabilité et calculabilité 509
5.1 Qu'est-ce qu'un nombre ? Approximations rationnelles et approximations décimales 509
5.2 Les réels sont-ils réels ? Alan Turing et la calculabilité 514
Exercices 517

III.4 Polynômes orthogonaux 531

1 Introduction 531
2 Généralités 532
2.1 Introduction 532
2.2 Formules de récurrence et de Darboux-Christoffel 537
2.3 Les zéros des polynômes orthogonaux 544
2.4 Approximation et formules de quadrature 546
3 Polynômes orthogonaux classiques 555
3.1 Équations différentielles et polynômes orthogonaux 557
3.2 Formule d'Olinde Rodrigues 574
3.3 Polynômes de Jacobi et cas particuliers (Legendre et Tchebychev) 581
3.4 Polynômes d'Hermite 590
3.5 Polynômes de Laguerre 599
3.6 Séries génératrices 604
3.7 Propriétés de densité pour les polynômes d'Hermite et de Laguerre 608
3.8 Tables des polynômes orthogonaux classiques 611
4 Compléments 617
4.1 Déterminants de Hankel, approximants de Padé 617
4.2 Polynômes orthogonaux et fractions continues 641
Exercices 647

IV Probabilités

IV.1 Notions fondamentales sur les probabilités 665

1 Ensemble fondamental et événements 665
1.1 Ensemble fondamental 665
1.2 La notion d'événement 666
1.3 La notion de tribu 667
2 Probabilités 668
2.1 Propriétés élémentaires d'une probabilité 669
2.2 Probabilité uniforme sur un ensemble fini 672
2.3 Probabilités sur un ensemble dénombrable 674
2.4 Probabilités uniformes sur $\mathbb{R}$ (ou $\mathbb{R}^d$) 675
2.5 Probabilités de réunions d'ensembles : règle d'inclusion-exclusion 677

3 Probabilités conditionnelles 680
3.1 Intuition et définition 680
3.2 Formule de Bayes 682
3.3 Conditionnement multiple 685
4 Indépendance 686
4.1 Indépendance de deux événements 686
4.2 Indépendance de plusieurs événements 688
4.3 Construction d'un espace de probabilité 690
4.4 Probabilité de réunions d'événements indépendants 692
5 Problèmes et paradoxes 696
5.1 Un exemple classique : la ruine du joueur 696
5.2 Paradoxes 698
6 Complément : équations aux différences 701
Exercices 702

IV.2 Variables aléatoires discrètes 707

1 Lois et variables aléatoires discrètes 707
1.1 Définitions 707
1.2 Histogrammes 709
2 Quelques lois usuelles 710
2.1 Loi de Bernoulli 710
2.2 Loi binomiale et nombre de succès 711
2.3 Temps d'attente et loi géométrique 714
2.4 Échantillonnage et loi hypergéométrique 715
3 Espérance de variables aléatoires discrètes réelles 717
3.1 Définition de l'espérance 717
3.2 Propriétés élémentaires de l'espérance 718
3.3 Propriétés de transport 720
3.4 Quelques remarques 721
3.5 Variance 722
3.6 Espérances et variances pour des lois usuelles 723
4 Problèmes d'approximation 727
4.1 Approximation de la loi hypergéométrique 727
4.2 Approximation de la loi binomiale, loi de Poisson 729
5 Familles de variables aléatoires 730
5.1 Loi d'un vecteur aléatoire 730
5.2 Covariance et corrélation 733
5.3 Indépendance 735
5.4 Indépendance et covariance 736
5.5 Schéma succès-échec infini 739
6 Fonctions génératrices 741
6.1 Définition 741
6.2 Cas des vecteurs aléatoires 743
6.3 Indépendance et convolution 744
6.4 Fonctions génératrices de lois usuelles 745
6.5 Sommes aléatoires de v.a. indépendantes 748
Exercices 750

IV.3 Variables aléatoires à densité 755
1 Variables aléatoires réelles 755
1.1 Définition 755
1.2 Loi et fonction de répartition 756
2 Variables à densité 758
2.1 Densité de probabilité 758
2.2 Lois usuelles 759
2.3 Caractérisation des v.a. réelles à densité 762
3 Moments de variables aléatoires à densité 764
3.1 Espérance 764
3.2 Moments 766
3.3 Moments des lois usuelles 767
3.4 Inégalités célèbres 769
4 Vecteurs aléatoires 770
4.1 Définition et loi 770
4.2 Vecteur aléatoire à densité 771
5 Indépendance 773
5.1 Définition 773
5.2 Indépendance de variables aléatoires à densité 774
5.3 Covariance et variance 776
5.4 Somme de variables aléatoires à densité indépendantes 778
5.5 Vecteurs gaussiens 780
6 Simulation de variables aléatoires 786
6.1 Simulation d'une v.a. uniforme 786
6.2 Méthode de la fonction inverse 787
6.3 Simulation d'une v. a. discrète 788
6.4 Simulation d'une v. a. gaussienne 789
6.5 Méthode du rejet pour une loi uniforme 789
Exercices 792
IV.4 Théorèmes limites et estimation 797
1 Loi des grands nombres 797
1.1 Loi faible des grands nombres 797
1.2 Loi forte des grands nombres 798
2 Théorème de la limite centrale 800
2.1 Approximation normale de la loi binomiale 800
2.2 Énoncé du théorème 803
3 Estimation 809
3.1 But de l'estimation 809
3.2 Qualités d'un estimateur 811
3.3 Estimateurs usuels 812
4 Intervalles de confiance 816
4.1 Intervalle de confiance et estimation 816
4.2 Échantillons gaussiens 816
4.3 Échantillons non gaussiens 824
Exercices 828
Bibliographie 835
Indications 837
Index 853

Avant-propos

Cet ouvrage est le dernier d'une série de trois, conçue pour couvrir les programmes de mathématiques de la plupart des Licences scientifiques.
Il complète le précédent en traitant certains sujets habituellement enseignés au niveau de la deuxième année de Licence et 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, et à savoir « 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 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

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

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 dans nos ouvrages 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

Partie I

Arithmétique I.1

On attribue à Gauß (lui-même surnommé « le prince des mathématiciens ») la phrase selon laquelle « les mathématiques sont la reine des sciences et la théorie des nombres est la reine des mathématiques ». C'est sans conteste le domaine où l'activité des mathématiciens a été le plus purement motivée par la beauté et le plaisir. Trois mutations ont progressivement changé le caractère de l'arithmétique (c'est l'autre nom de la théorie des nombres) telle que l'avait connue Gauß :

1. Sous son influence, puis sous celle d'autres mathématiciens du XIXe siècle (surtout allemands !), de profondes techniques d'algèbre puis d'analyse (et même, au XXe siècle, de géométrie algébrique) ont renforcé l'arsenal de l'arithméticien.
2. Au XXe siècle, des problèmes de pure théorie des nombres ont rencontré des échos en physique.
3. Encore au XXe siècle, des applications financières et même militaires lui sont tombées dessus ! Il s'agit bien entendu des questions de sécurité de communication : cryptographie, authentification ...

Ce chapitre commencera donc par un peu d'outillage algébrique (section 1, consacrée à une approche abstraite de la théorie de la divisibilité et de la factorisation), complété par la section 2, intermédiaire entre l'algèbre et l'arithmétique, et qui comportera des applications informatiques ; et il se terminera par un pont avec la théorie des fonctions spéciales (section 6), qui jouent un rôle essentiel en physique ; cet aspect sera développé dans le module III.4 de ce volume sur les polynômes orthogonaux ; les aspects proprement arithmétiques des fractions continues pourront être approfondis dans le livre [MPA1]. Entre ces limites, les sections 3, 4 et 5 permettront de jouer avec des problèmes remontant à Fermat, Euler, Lagrange, Gauß, Legendre... et dont certains (critères de primalité) jouissent d'un intérêt renouvelé depuis quelques années.

Nous avons proposé dans la bibliographie quelques livres qui vous permettront d'approfondir les sujets abordés dans ce module et d'en découvrir d'autres. Les ouvrages [HW] et [Serre] sont de grands classiques que tout mathématicien, même amateur, devrait posséder (le second est toutefois nettement plus avancé que le premier). [Demazure] fait le lien avec les applications informatiques, que [Knuth2] (qui est beaucoup plus ancien) développe de manière détaillée. [Baker] et [Hindry] sont très riches, mais d'un abord plus difficile.

Comme des notions plus générales vont apparaître, précisons que nous réservons l'appellation de *nombre premier* aux entiers naturels premiers : $2, 3, 5, 7 \ldots$.

1 Divisibilité dans un anneau commutatif

Notre but est de généraliser autant que possible le *théorème fondamental de l'arithmétique* : existence pour tout entier d'une factorisation essentiellement unique en produit de nombres premiers. Le cadre est ici celui des anneaux commutatifs intègres (autrement, les difficultés techniques seraient trop grandes). Les éléments les plus caractéristiques de l'arithmétique d'un tel anneau sont ses *unités*, *i.e.* ses éléments inversibles (ce sont eux qui empêchent la décomposition d'être vraiment unique) ; et, bien entendu, ses éléments irréductibles et ses éléments premiers : nous verrons que ces deux termes ne désignent pas exactement la même chose, et c'est même là le cœur du problème !

1.1 Anneaux euclidiens et anneaux principaux

Résumons l'enchaînement logique des propriétés arithmétiques des anneaux intègres $\mathbb{Z}$ et $K[X]$ (K un corps commutatif), telles qu'on les a établies dans le livre [L1] :

1. Dans $\mathbb{Z}$ et dans $K[X]$, on dispose d'une *division euclidienne*.
2. Les anneaux $\mathbb{Z}$ et $K[X]$ sont principaux. On en déduit le théorème de Bézout, le lemme de Gauß et le lemme d'Euclide.
3. Dans $\mathbb{Z}$ et dans $K[X]$, tout élément admet une factorisation essentiellement unique en produit de facteurs irréductibles.

Nous allons maintenant formaliser (ici et dans la section 1.2) les liens logiques entre ces propriétés en vue d'applications plus générales.

Définition 1. Un anneau (commutatif intègre) A est dit *euclidien* s'il admet un *stathme euclidien*, c'est-à-dire une application $g : A \setminus \{0\} \to \mathbb{N}$ telle que, pour tout couple (x, y) d'éléments de A :

1. Si $x, y \neq 0$ et $x \mid y$, alors $g(x) \leqslant g(y)$. De manière équivalente, si $x, x' \neq 0$, alors $g(x) \leqslant g(xx')$.
2. Si $x \neq 0$, il existe $q, r \in A$ tels que $y = qx + r$ et : soit $r = 0$; soit $g(r) < g(x)$.

Les éléments q et r sont appelés *quotient* et *reste* de la *division euclidienne*. On ne requiert *pas* l'unicité de la division euclidienne.

L'exemple qui suit sera fondamental dans les applications arithmétiques de ce chapitre. On rappelle que $\mathbb{Z}[\mathrm{i}]$ désigne l'anneau des entiers de Gauß $a + b\mathrm{i}$, $a, b \in \mathbb{Z}$ et que la norme algébrique N est définie par $N(z) := z\overline{z}$ (on approfondira ces notions en 4.1).

Exercice 1.

Démontrer que, pour tout complexe $z \in \mathbb{C}$, il existe un entier de Gauß q tel que $N(z - q) < 1$, où N est la norme algébrique sur $\mathbb{C}$. En déduire que N est un stathme euclidien sur $\mathbb{Z}[\mathrm{i}]$.

Solution. Si $z = a + \mathrm{i}b$, notons a_0 l'entier le plus proche de a et b_0 l'entier le plus proche de b (s'il y a ambiguïté, c'est-à-dire si a ou b est un entier plus $1/2$, on prend la partie entière) ; on a donc $|a - a_0|, |b - b_0| \leqslant 1/2$, d'où, en posant $q := a_0 + b_0\mathrm{i}$, $N(z - q) \leqslant 1/4 + 1/4 < 1$.

Soient maintenant $x, y \in \mathbb{Z}[\mathrm{i}]$, $x \neq 0$. En appliquant ce qui précède à $z := \dfrac{y}{x}$, on trouve $q \in \mathbb{Z}[\mathrm{i}]$ tel que, si $r := y - qx$, on ait $N(r) < N(x)$ et $y = qx + r$: que r soit nul ou pas, c'est bien une division euclidienne.
Par ailleurs, si $x, x' \in \mathbb{Z}[\mathrm{i}] \setminus \{0\}$, on a $N(x') \geqslant 1$, d'où $N(x) \leqslant N(xx')$.

Théorème 1. Tout anneau euclidien est principal.

Démonstration. On rappelle (*cf.* le livre [L1], module « Groupes, anneaux, corps ») qu'un anneau principal est un anneau intègre dans lequel tout idéal est principal.
Soit I un idéal de l'anneau euclidien A. Si $I = 0$, il n'y a rien à démontrer. Sinon, soit $x \in I \setminus \{0\}$ tel que l'entier $g(x)$ est minimum. Alors $Ax \subset I$. Pour tout $y \in I$, dans la division euclidienne $y = qx + r$, on a $r \in I$, donc $r = 0$ car $g(r) < g(x)$ est impossible (minimalité), donc $y \in Ax$ et $I = Ax$. ∎

Exemples. L'application $x \mapsto |x|$ est un stathme euclidien sur $\mathbb{Z}$.
L'application $P \mapsto \deg P$ est un stathme euclidien sur $K[X]$ (K un corps). Ces anneaux sont donc principaux, ainsi que $\mathbb{Z}[\mathrm{i}]$. En revanche, l'anneau $\mathbb{Z}[X]$ n'est pas principal : si l'idéal $<2, X>$ était engendré par u, ce dernier diviserait 2 et X, donc vaudrait ± 1 ; mais 1 et -1 n'appartiennent pas à l'idéal engendré par 2 et X, car tout élément P de cet idéal est tel que $P(0)$ est pair.
Soient x, y deux éléments de l'anneau principal A. Tout générateur d de l'idéal $Ax + Ay$ est un diviseur commun à x et y (car $x, y \in Ad$); comme il s'écrit $d = ax + by$, tout diviseur commun à x et y divise d : c'est donc un pgcd de x et y.

Proposition 2 (Algorithme d'Euclide). Supposons A euclidien. L'algorithme suivant permet de calculer un pgcd de x et y. On admet l'existence d'une fonction **`reste(y,x)`** qui, si $x \neq 0$, rend le reste **`r`** d'une division euclidienne $y = qx + r$.
On suppose au départ que : $y \neq 0$, et que : $x = 0$ ou $g(x) < g(y)$ (on peut s'y ramener au prix d'une division euclidienne de y par x). Voici le corps de l'algorithme :

```
pgcd(x,y)
si x = 0
    rendre y
sinon
    (r := reste(y,x); rendre pgcd(x,r));
```

Démonstration. C'est un algorithme *récursif* (voir le module « Structures discrètes et récursivité » du livre [L2]). Dans le cas de sortie où $x = 0$, il n'y a pas de problème, on sait bien que y est un pgcd de x et y.
Si $x \neq 0$, et si $y = qx + r$, on sait qu'un pgcd de x et r est aussi un pgcd de x et y (même raisonnement que dans les versions de cet algorithme du livre [L1]). Pour calculer **`pgcd(x,y)`**, il suffit donc de calculer **`pgcd(x,r)`**. Est-ce plus simple ? *Oui*, car soit $r = 0$ (et le calcul est instantané), soit $g(r) < g(x) < g(y)$, et l'on est plus près de la *sortie* de l'algorithme qu'au premier appel de la fonction **`pgcd`**.

Si l'on préfère une version non récursive (boucle **`tant que`**), il suffit de calquer celle du livre [L1]. ■

Le résultat suivant est donné sous une forme concrète dans le cas d'un anneau euclidien. Voir [MPA1] pour une extension (considérablement enrichie) au cas des anneaux principaux.

Proposition 3 (Algorithme du pivot sur un anneau euclidien). Soit A un anneau euclidien et soit $M \in M_{m,n}(A)$ une matrice à coefficients dans A. On peut, par des permutations de lignes et de colonnes, et par des transvections (à coefficients dans A) sur les lignes et les colonnes, mettre M sous la forme :

$$\begin{pmatrix} d_1 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ 0 & d_2 & \dots & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & d_r & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \end{pmatrix} = \sum_{i=1}^{r} d_i E_{i,i} = \begin{pmatrix} D & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

pour un certain $r \in [\![0, \min(m,n)]\!]$ et des éléments $d_1, \dots, d_r$ de $A \setminus \{0\}$ tels que $d_1 \mid d_2 \mid \cdots \mid d_r$. Dans la dernière écriture, $D := \mathrm{Diag}(d_1, \dots, d_r)$, et $0_{p,q}$ désigne la matrice nulle de type (p,q).

Démonstration. Il s'agit bien sûr d'une adaptation de l'algorithme du pivot de Gauß sur un corps, mais en plus compliqué ; le lecteur est donc invité à se replonger dans l'algorithme classique (*cf.* le livre [L1], module sur les matrices). On peut supposer la matrice M non nulle. Pour simplifier les notations, nous noterons ici $\gamma(M)$ le minimum des $g(m_{i,j})$ pour $m_{i,j}$ non nul (les $g(m_{i,j})$ sont des entiers).

Pour commencer, on choisit un coefficient $m_{i,j}$ (le « pivot ») tel que $g(m_{i,j}) = \gamma(M)$. Par permutation de lignes et colonnes, on le ramène en position $(1,1)$. C'est donc maintenant $g(m_{1,1})$ qui est minimal. Par divisions euclidiennes $m_{1,j} = q_j m_{1,1} + r_j$, on peut définir des transvections sur les colonnes : $C_j \leftarrow C_j - q_j C_1$, à l'issue desquelles chaque $m_{1,j}$ $(j > 1)$ est nul ou tel que $g(m_{1,j}) < g(m_{1,1})$. De même, par des transvections sur les lignes, on peut obtenir la propriété que chaque $m_{i,1}$ $(i > 1)$ est nul ou tel que $g(m_{i,1}) < g(m_{1,1})$.

Si l'un des $m_{1,j}$ $(j > 1)$ ou l'un des $m_{i,1}$ $(i > 1)$ n'est pas nul, $\gamma(M)$ a diminué strictement et l'on recommence tout le processus ci-dessus. Cela ne peut se produire une infinité de fois puisque l'entier $\gamma(M)$ diminue strictement chaque fois. On peut donc ramener M à la forme $\begin{pmatrix} m_{1,1} & 0 \\ 0 & M' \end{pmatrix}$, où M' a une ligne et une colonne de moins que M (si M n'avait qu'une ligne ou une colonne, l'algorithme est déjà terminé).

Supposons que $m_{1,1}$ ne divise pas tous les coefficients de M'. Soit $m_{i,j}$ $(i,j > 1)$ non multiple de $m_{1,1}$; on a donc une division euclidienne $m_{i,j} = q m_{1,1} + r$, $g(r) < g(m_{1,1})$.

On opère la transvection $C_j \leftarrow C_j + C_1$, puis la transvection $L_i \leftarrow L_i - qL_1$ et maintenant $m_{i,j} = r$, donc $\gamma(M)$ a diminué strictement. En itérant l'ensemble des deux processus décrits ci-dessus, on peut ramener M à la forme $\begin{pmatrix} m_{1,1} & 0 \\ 0 & M' \end{pmatrix}$, où, de plus, $m_{1,1}$ divise tous les coefficients de M'. Ce $m_{1,1}$ est d_1.
On recommence maintenant l'ensemble de l'algorithme ci-dessus sur M'. ■

Le lecteur courageux pourra tenter la programmation de cet algorithme : il constatera que la structure ne peut en être substantiellement simplifiée.

Exemple. Nous allons étudier le système à inconnues entières $\begin{pmatrix} -10 & 14 \\ -8 & 10 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$, ou encore $\begin{cases} -10x + 14y = a, \\ -8x + 10y = b. \end{cases}$ $(x, y, a, b \in \mathbb{Z})$. On effectue successivement les opérations $L_1 \leftrightarrow L_2$, $C_2 \leftarrow C_2 + C_1$, $L_2 \leftarrow L_2 - L_1$, $C_1 \leftrightarrow C_2$, $C_2 \leftarrow C_2 + 4C_1$ et $L_2 \leftarrow L_2 - L_1$, ce qui donne la matrice $\begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}$. Du côté des systèmes, les transvections sur les colonnes se traduisent par les changements de variable $x' := x - y$ puis $y' := y - 4x'$. Le système final est $\begin{cases} 2y' = b, \\ 6x' = a - 2b. \end{cases}$
Il admet des solutions entières si, et seulement si, $2 \mid b$ et $6 \mid a - 2b$. On trouve alors les solutions du système de départ en inversant le changement de variables : $y := y' + 4x'$ et $x := x' + y = y' + 5x'$. Par exemple si $a := 10$ et $b := 2$, on trouve $(x', y') = (1, 1)$, puis $(x, y) = (6, 5)$.

1.2 Anneaux principaux et anneaux factoriels

Ces notions ont déjà reçu des applications dans le module sur les polynômes à plusieurs indéterminées du livre [L2]. Rappelons et complétons le vocabulaire sur la divisibilité dans un anneau *intègre*. Nous notons A^* le groupe des unités (éléments inversibles) de A.

1. On dit que x divise y, notation : $x \mid y$, s'il existe $a \in A$ tel que $y = ax$. Si $x = 0$, cela implique $y = 0$ et a est quelconque ; si $x \neq 0$, un tel a est unique (intégrité !), on le note $\dfrac{y}{x}$. On dit que x et y sont associés, ce que l'on notera $x \sim y$, si chacun divise l'autre. Si $x = 0$, cela implique $y = 0$ (et réciproquement) ; si x et y sont non nuls, cela équivaut à l'existence de $u \in A^*$ tel que $y = ux$, et donc $x = u^{-1}y$.
2. On dit que x est *irréductible* s'il est non inversible et s'il n'admet aucune décomposition non triviale ; autrement dit, pour toute décomposition $x = yz$, on a y inversible (et donc z associé à x) ou bien z inversible (et donc y associé à x).
3. On dit que x est *premier* s'il est non inversible et non nul et si l'idéal principal Ax est premier ; de manière équivalente, $x \neq 0$, $x \notin A^*$ et :
$$\forall y, z \in A\ ,\ x \mid yz \Rightarrow (x \mid y \quad \text{ou} \quad x \mid z)\,.$$
Il est bien évident que tout élément premier est irréductible.

Idéaux et divisibilité dans un anneau principal. Soit A un anneau principal qui n'est pas un corps (cette hypothèse sera désormais implicite). Outre l'idéal nul, ses idéaux sont de la forme $<x>$ avec $x \neq 0$. De plus, $<x> = <y>$ si, et seulement si, x et y sont associés. Il y a donc bijection entre les idéaux non nuls et les classes d'éléments non nuls de A (pour la relation « être associé »). De plus, $x \mid y$ si, et seulement si, $<y> \subset <x>$. Ainsi, le plus grand idéal, A, correspond à la plus petite classe (pour la relation « diviser »), A^* (la classe de 1); et les idéaux maximaux correspondent aux classes des éléments irréductibles. Les idéaux premiers non nuls de A sont donc les $<x>$, x irréductible, et ils sont maximaux et ces x sont premiers.

La situation n'est en général pas aussi simple. Commençons par quelques exemples motivants. Ceux-ci font appel aux propriétés de la norme algébrique des nombres complexes : $N(a + \mathrm{i}b) := a^2 + b^2$ (pour $a, b \in \mathbb{R}$); rappelons que $N(uv) = N(u)N(v)$.

Exemple. Soit $d \in \mathbb{Z}$, $d > 0$. Rappelons que, par convention, nous notons $\sqrt{-d} := \mathrm{i}\sqrt{d}$. Si $u \in \mathbb{Z}[\sqrt{-d}] \setminus \{0\}$, alors $N(u) \in \mathbb{N}^*$ (si $a, b \in \mathbb{Z}$ sont non tous deux nuls, $N(a + b\sqrt{-d}) = a^2 + db^2$ est un entier strictement positif). On en déduit que les éléments inversibles de l'anneau $\mathbb{Z}[\sqrt{-d}]$ sont les éléments u tels que $N(u) = 1$: en effet, si u a pour inverse v, on a $N(u)N(v) = N(uv) = 1$, d'où $N(u) = 1$ (c'est un entier positif); réciproquement, si $N(u) = 1$, le conjugué $\overline{u}$ est l'inverse de u. Pour des entiers, $a^2 + db^2 = 1$ n'est possible que si $a = \pm 1$ et $b = 0$, sauf dans le cas où $d = 1$, pour lequel on a aussi les solutions $a = 0$ et $b = \pm 1$. Les éléments inversibles de $\mathbb{Z}[\sqrt{-d}]$ sont donc ± 1 et $\pm \mathrm{i}$ si $d = 1$ et ± 1 si $d \geqslant 2$.
Montrons maintenant que le nombre premier $p \in \mathbb{N}$ est réductible dans $\mathbb{Z}[\sqrt{-d}]$ si, et seulement s'il est de la forme $a^2 + db^2$ avec $a, b \in \mathbb{Z}$. Supposons tout d'abord $p = uv$, avec $u, v \in \mathbb{Z}[\sqrt{-d}]$ non inversibles. Alors $p^2 = N(p) = N(u)N(v)$ et $N(u), N(v) > 1$, donc $N(u) = N(v) = p$. Comme $N(a + b\sqrt{-d}) = a^2 + db^2$, p est bien de la forme indiquée. Réciproquement, si $p = a^2 + db^2$ avec $a, b \in \mathbb{Z}$, on a $p = uv$, avec $u := a + b\sqrt{-d} \in \mathbb{Z}[\sqrt{-d}]$ et $v := \overline{u} \in \mathbb{Z}[\sqrt{-d}]$ tous deux non inversibles.

Exercice 2.

On prend maintenant $d := 5$. Montrer que 3 est irréductible dans l'anneau $\mathbb{Z}[\sqrt{-5}]$. Vérifier que 3 divise $(2 + \sqrt{-5})(2 - \sqrt{-5})$ mais qu'il ne divise ni $2 + \sqrt{-5}$ ni $2 - \sqrt{-5}$.
Solution. On vérifie facilement que $3 = a^2 + 5b^2$ est impossible avec $a, b \in \mathbb{Z}$. Ainsi, 3 est irréductible dans $\mathbb{Z}[\sqrt{-5}]$ (de même, du fait que $N(2+\sqrt{-5}) = N(2-\sqrt{-5}) = 9$, on peut déduire que $2+\sqrt{-5}$ et $2-\sqrt{-5}$ sont irréductibles). On a $(2+\sqrt{-5})(2-\sqrt{-5}) = 3^2$. Mais ni $\dfrac{2 + \sqrt{-5}}{3}$, ni $\dfrac{2 - \sqrt{-5}}{3}$, ne sont de la forme $a + b\sqrt{-5}$ avec $a, b \in \mathbb{Z}$, donc 3 ne divise ni $2 + \sqrt{-5}$ ni $2 - \sqrt{-5}$. Ainsi, 3 *est irréductible mais n'est pas premier*. C'est cette discordance qui est impossible dans un anneau factoriel.

Exercice 3.

Soit $d \in \mathbb{Z}$ qui n'est pas un carré. Décrire les éléments de $\mathbb{Z}$ qui sont premiers dans $\mathbb{Z}[\sqrt{d}]$.

Solution. Puisque, par construction, l'anneau $\mathbb{Z}[\sqrt{d}]$ est intègre, l'idéal $\{0\}$ est premier dans $\mathbb{Z}[\sqrt{d}]$; mais, par définition, l'élément 0 n'est pas considéré comme premier.
Soit $p \in \mathbb{N}^*$ ($-p$ engendre le même idéal) ; par convention, 1 n'est jamais premier (car l'anneau trivial n'est pas intègre), et l'on peut supposer $p \geqslant 2$. En vertu des résultats du module « Compléments d'algèbre » du livre [L2] (théorèmes d'isomorphisme pour les anneaux, et polynômes sur un anneau), l'anneau $\mathbb{Z}[\sqrt{d}]/<p>$ est isomorphe à l'anneau $\mathbb{Z}[X]/<X^2 - d, p>$, qui est lui-même isomorphe à l'anneau $\mathbb{F}_p[X]/<X^2 - \overline{d}>$, où $\overline{d}$ désigne la classe de d modulo p ; rappelons que l'on note $\mathbb{F}_p$ le corps $\mathbb{Z}/p\mathbb{Z}$. Cet anneau est intègre si, et seulement si, le polynôme $X^2 - \overline{d}$ est irréductible, *i.e.* (puisqu'il est de degré 2) s'il n'a pas de racines. En conclusion, $p \geqslant 2$ est premier dans $\mathbb{Z}[\sqrt{d}]$ si, et seulement si, d n'est pas un carré modulo p.

1.2.1 Divisibilité dans les anneaux factoriels

Les preuves des résultats qui vont suivre sont en tous points similaires à celles données dans le cas de $\mathbb{Z}$ et de $K[X]$, voire pour tout anneau principal, dans le livre [L1] : nous laissons au lecteur le soin d'en rédiger les détails.

Théorème et définition 4. Soit A un anneau intègre (commutatif) dans lequel tout élément non nul et non inversible est produit d'éléments irréductibles. Les propriétés suivantes sont alors équivalentes :
(i) La décomposition en produit d'irréductibles est unique à l'ordre des facteurs près et à inversible près.
(ii) Tout élément irréductible est premier.
On dit alors que l'anneau A est *factoriel*. Cette propriété équivaut encore à la suivante : tout élément non inversible de A est produit d'éléments *premiers*.

La propriété (i) signifie que, si $p_1 \cdots p_r = q_1 \cdots q_s$, les p_i et q_j étant irréductibles, alors $r = s$ et, quitte à réordonner les facteurs, q_i est associé à p_i pour tout $i \in [\![1, r]\!]$.

Point Méthode

Il sera plus commode de procéder comme suit. Dans l'ensemble des éléments irréductibles de A, on fixe un ensemble de représentants P. Autrement dit, tous les éléments de P sont irréductibles ; et, pour chaque $p \in A$ irréductible, il y a un unique $p' \in P$ tel que $p \sim p'$ (*i.e.* p et p' sont associés). Ce choix étant fait, les propriétés du théorème et de la définition se résument ainsi : tout $a \in A$ non nul admet une unique écriture $a = up_1^{r_1} \cdots p_k^{r_k}$ ($u \in A^*$, $k \in \mathbb{N}$ et les $r_i \in \mathbb{N}^*$, les $p_i \in P$).

Exemple. Si $A = \mathbb{Z}$, on prend pour P l'ensemble des nombres premiers (dans $\mathbb{N}$). Si $A = K[X]$, on prend pour P l'ensemble des polynômes unitaires irréductibles.

Théorème 5. Tout anneau principal est factoriel.

1.2.2 Valuations

Soit $a \in A$. On introduit les *valuations* v_p par la formule :

$$a = u \prod_{p \in P} p^{v_p(a)},$$

où $u \in A^*$ et où les $v_p(a) \in \mathbb{N}$ sont presque tous nuls.
On convient par ailleurs que $\forall p \in P\ ,\ v_p(0) := +\infty$. On a alors les formules :

$$v_p(ab) = v_p(a) + v_p(b) \qquad \text{et} \qquad v_p(a+b) \geqslant \min\big(v_p(a), v_p(b)\big).$$

On vérifie immédiatement que $a \mid b$ équivaut à : pour tout $p \in P$, $v_p(a) \leqslant v_p(b)$.
Ainsi, $a \sim b$ équivaut à : pour tout $p \in P$, $v_p(a) = v_p(b)$.
Il en découle que le *plus grand commun diviseur* $a \wedge b$ de a et b se calcule ainsi :

$$a \wedge b = \prod_{p \in P} p^{\min\left(v_p(a), v_p(b)\right)};$$

les pgcd de a et b sont les éléments associés à $a \wedge b$. Les diviseurs communs à a et b sont les diviseurs de $a \wedge b$. On voit que a et b sont premiers entre eux (ou étrangers) si, et seulement si, $a \wedge b = 1$.

Dire que les éléments a et b sont étrangers n'est pas équivalent à dire que les idéaux $<a>$ et $<b>$ sont étrangers (c'est-à-dire, selon le module « Compléments d'algèbre » du livre [L2], que $<a> + <b> = A$) : voir l'exercice I.1.8 de la page 69.

Proposition 6 (Lemme de Gauß). Si a et b sont étrangers et si $a \mid bc$, alors $a \mid c$.

Plus généralement, on peut calculer :

$$pgcd(a_1, \ldots, a_n) := a_1 \wedge \cdots \wedge a_n := \prod_{p \in P} p^{\min\left(v_p(a_1), \ldots, v_p(a_n)\right)}$$

et dire que les éléments $a_1, \ldots, a_n \in A$ sont premiers entre eux (ou étrangers) *dans leur ensemble* si ce pgcd est égal à 1. De même, on peut définir le ppcm par la formule : $ppcm(a_1, \ldots, a_n) := a_1 \vee \cdots \vee a_n := \prod_{p \in P} p^{\max\left(v_p(a_1), \ldots, v_p(a_n)\right)}$. Les multiples communs à $a_1, \ldots, a_n$ sont les multiples de leur ppcm.

Exercice 4.
On suppose que a et b sont étrangers et que ab est de la forme x^d pour un $d \geqslant 2$. Montrer qu'il existe $y, z \in A$ tels que $a \sim y^d$ et $b \sim z^d$.
Solution. Puisque a et b sont étrangers, pour tout $p \in P$, $v_p(a)$ ou $v_p(b)$ est nul ; mais comme $v_p(a) + v_p(b) = v_p(x^d) = dv_p(x)$, $v_p(a)$ et $v_p(b)$ sont tous deux multiples de d. Or, la condition « il existe $y \in A$ tel que $a \sim y^d$ » est clairement équivalente à : « pour tout $p \in P$, $v_p(a)$ est multiple de d ».

Extension des v_p au corps des fractions de A

Soit $\dfrac{a}{b}$ un élément du corps des fractions K de A. Quitte à simplifier le numérateur et le dénominateur par $a \wedge b$, on peut supposer que cette fraction est irréductible, c'est-à-dire que a et b sont premiers entre eux. Dans ce cas, pour toute égalité $\dfrac{a}{b} = \dfrac{c}{d}$, on a $c = am$ et $d = bm$ avec $m \in A$: cela découle du lemme de Gauß. En particulier, si $\dfrac{c}{d}$ est également irréductible, alors a et c sont associés, ainsi que b et d.
On peut étendre les valuations v_p au corps des fractions de A en posant :

$$\forall p \in P \ , \ \forall a, b \in A \setminus \{0\} \ , \ v_p\left(\frac{a}{b}\right) := v_p(a) - v_p(b).$$

Le lecteur vérifiera que cette définition n'est pas ambiguë et que les règles sur $v_p(ab)$ et $v_p(a+b)$ restent valides. De plus, tout élément $x \in K^*$ admet une unique écriture :

$$x = u \prod_{p \in P} p^{v_p(x)},$$

où $u \in A^*$ et où les $v_p(x) \in \mathbb{Z}$ sont presque tous nuls.

Exercice 5.
Démontrer que $\sqrt{2}$ est irrationnel.
Solution. Si $x := \sqrt{2}$ était rationnel, de l'égalité $x^2 = 2$ on tirerait $2v_2(x) = 1$, ce qui est impossible puisque $v_2(x) \in \mathbb{Z}$.

Remarque. Notons P^* le sous-groupe de K^* engendré par P, autrement dit :

$$P^* := \{x \in K^* \mid x = \prod_{p \in P} p^{v_p(x)}\}.$$

Il résulte des conventions précédentes que *le* pgcd et *le* ppcm d'éléments de $A \setminus \{0\}$ sont des éléments de P^*.

La propriété suivante, bien que facile à démontrer, est très utile :

Proposition 7. Pour que $x \in K$ soit élément de A, il faut, et il suffit, que $v_p(x)$ soit positif ou nul pour tout $p \in P$.

Exercice 6.
Démontrer que $x = \sqrt{2} + \sqrt{3}$ est irrationnel.
Solution. Si x était rationnel, de l'égalité $(x^2-5)^2 = 24$, *i.e.* $x^4 - 10x^2 + 1 = 0$, on déduirait que $x^2 = 10 - x^{-2}$, donc que $v_p(x^2) \geqslant \min(v_p(10), -v_p(x^2))$, donc que $v_p(x) \geqslant 0$ pour tout $p \in P$, et x serait entier. Mais c'est impossible car $1{,}4 + 1{,}7 < x < 1{,}5 + 1{,}8$.

1.3 Polynômes sur un anneau factoriel

Soit A un anneau factoriel, de corps des fractions K. On choisit un ensemble de représentants P de l'ensemble des irréductibles de A (pour la relation « être associé »).

Définition 2. Un polynôme non nul $F \in A[X]$ est dit *primitif* si ses coefficients sont premiers entre eux dans leur ensemble.

Proposition et définition 8. Soit $F \in K[X]$ un polynôme non nul. Il existe alors un unique $c(F) \in P^*$ et un unique polynôme primitif $\tilde{F} \in A[X]$ tels que $F = c(F)\tilde{F}$. L'élément $c(F)$ est appelé *contenu* de F.

Démonstration. Écrivons $F = \sum \frac{a_i}{b_i} X^i$, et soit b le ppcm des dénominateurs des b_i. Alors $bF \in A[X]$, et l'on prend pour $c(F)$ le pgcd des coefficients de bF divisé par b. Les propriétés indiquées (unicité, primitivité) sont alors immédiates. ∎

Il est clair que $F \in A[X]$ si, et seulement si, $c(F) \in A$.

Théorème 9 (Gauß).
(i) Le produit de deux polynômes primitifs est primitif.
(ii) Si $F, G \in K[X]$ sont non nuls, $c(FG) = c(F)c(G)$.

Démonstration. Soient $F, G \in A[X]$. Si $FG \in A[X]$ n'est pas primitif, il existe $p \in P$ qui divise tous les coefficients de FG. Dans l'anneau intègre $A/<p>[X]$, le produit des images $\overline{F}$ et $\overline{G}$ est nul, donc l'un des deux, par exemple $\overline{F}$, est nul ; donc p divise tous les coefficients de F, qui n'est donc pas primitif : on a prouvé (i) par contraposée.
L'assertion (ii) est alors immédiate en invoquant la proposition 8. ∎

Lemme 10. Les éléments suivants sont premiers dans $A[X]$: d'une part, les irréductibles de A, d'autre part, les polynômes primitifs de $A[X]$ qui sont irréductibles dans $K[X]$.

Démonstration. En ce qui concerne les irréductibles de A, cela résulte de la preuve du théorème 9. Soit F un polynôme primitif de $A[X]$ irréductible dans $K[X]$. Supposons que $F \mid GH$ dans $A[X]$. Comme F divise GH dans $K[X]$, il divise (par exemple) G dans $K[X]$: on a donc $FE = G$ avec $E \in K[X]$. Alors $c(E) = c(G)$ (car $c(F) = 1$), donc $c(E) \in A$, donc $E \in A[X]$ et F divise G dans $A[X]$. ∎

Théorème 11. L'anneau $A[X]$ est factoriel. Ses irréductibles sont, d'une part, les irréductibles de A, d'autre part, les polynômes primitifs de $A[X]$ qui sont irréductibles dans $K[X]$.

Démonstration. Soit F un élément non inversible de $A[X]$. S'il est constant, c'est un

élément non inversible de A, donc, de manière essentiellement unique, un produit d'irréductibles de A ; et il n'est pas multiple d'un polynôme non constant. Si F est non constant, il est produit d'irréductibles de $K[X]$, ce que l'on écrit $F = F_1 \cdots F_r$. On a donc $c(F) = c(F_1) \cdots c(F_r)$, d'où $\tilde{F} = \tilde{F}_1 \cdots \tilde{F}_r$. Finalement, $F = c(F)\tilde{F}_1 \cdots \tilde{F}_r$, où $c(F) \in A$ et les $\tilde{F}_i$ primitifs dans $A[X]$ et irréductibles dans $K[X]$. Le lemme précédent et la remarque qui suit le théorème 4 de la page 9 permettent alors de conclure. ■

Nous pouvons maintenant compléter la théorie du résultant (module sur les polynômes à plusieurs indéterminées du livre [L2]).

Corollaire 12. Soient $P, Q \in A[X]$. Si l'anneau A est factoriel, pour que $\mathrm{Res}(P, Q)$ s'annule, il faut, et il suffit, que P et Q aient un facteur commun non constant dans $A[X]$.

2 Le groupe multiplicatif de l'anneau $\mathbb{Z}/n\mathbb{Z}$

Cette section est surtout destinée à fournir des outils pour la suivante ; cependant, on y trouvera deux applications informatiques intéressantes en 2.4.

2.1 Rappels sur $(\mathbb{Z}/n\mathbb{Z})^*$

Pour tout $n \geqslant 1$ entier, on note $(\mathbb{Z}/n\mathbb{Z})^*$ le groupe multiplicatif de l'anneau $\mathbb{Z}/n\mathbb{Z}$, c'est-à-dire, en tant qu'ensemble :

$$(\mathbb{Z}/n\mathbb{Z})^* = \{\overline{a} \mid a \in \{0, \ldots, n-1\}\ ,\ a \wedge n = 1\},$$

où l'on a noté $\overline{a} \in \mathbb{Z}/n\mathbb{Z}$ la classe de a et $a \wedge n$ le pgcd de a et de n (choisi positif). L'indicatrice d'Euler $\varphi(n)$ est alors définie par la formule :

$$\varphi(n) := \mathrm{card}\,(\mathbb{Z}/n\mathbb{Z})^* = \mathrm{card}\,\{a \in \{0, \ldots, n-1\}\ ,\ a \wedge n = 1\}.$$

Exemple. Si $n = p^r$, p premier, $r \geqslant 1$, les non inversibles de $\mathbb{Z}/p^r\mathbb{Z}$ sont les $\overline{a}$ tels que $p \mid a$, *i.e.* les $\overline{pb}$ avec $0 \leqslant b < p^{r-1}$. Il y en a donc p^{r-1} et $\varphi(p^r) = p^r - p^{r-1}$.

Si $n = p_1^{r_1} \cdots p_k^{r_k}$ (décomposition en facteurs premiers), on déduit du lemme chinois l'isomorphisme d'anneaux puis l'isomorphisme de groupes :

$$\mathbb{Z}/n\mathbb{Z} \simeq \mathbb{Z}/p_1^{r_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_k^{r_k}\mathbb{Z} \Longrightarrow (\mathbb{Z}/n\mathbb{Z})^* \simeq (\mathbb{Z}/p_1^{r_1}\mathbb{Z})^* \times \cdots \times (\mathbb{Z}/p_k^{r_k}\mathbb{Z})^*,$$

puisque pour deux anneaux A, B quelconques $(A \times B)^* = A^* \times B^*$. On en tire les formules :

$$\varphi(n) = \varphi(p_1^{r_1}) \cdots \varphi(p_k^{r_k}) = (p_1^{r_1} - p_1^{r_1-1}) \cdots (p_k^{r_k} - p_k^{r_k-1}) = n(1 - 1/p_1) \cdots (1 - 1/p_k).$$

Exercice 7.
À quelle condition a-t-on $\varphi(n) = 1$? $\varphi(n) = 2$? $\varphi(n)$ impair ?
Solution. Puisque $\varphi(n) = \varphi(p_1^{r_1}) \cdots \varphi(p_k^{r_k})$, on ne peut avoir $\varphi(n) = 1$ que si $n = 1$ ou 2. De plus, sauf dans le cas de $\varphi(2)$, chaque facteur $\varphi(p_i^{r_i})$ est pair ; donc on ne peut avoir $\varphi(n) = 2$ que si $n = 3$ ou 4 et $\varphi(n)$ ne peut être impair que si $n = 1$ ou 2.

Plus généralement, on voit que la fonction arithmétique φ est *multiplicative*, ce qui, en arithmétique, exprime la propriété suivante :

$$\forall a, b \in \mathbb{N}\ ,\ a \wedge b = 1 \Longrightarrow \varphi(ab) = \varphi(a)\varphi(b).$$

Puisque le groupe $(\mathbb{Z}/n\mathbb{Z})^*$ a $\varphi(n)$ éléments, on déduit du théorème de Lagrange le théorème d'Euler :

$$\forall a \in \mathbb{Z}\ ,\ a \wedge n = 1 \Longrightarrow a^{\varphi(n)} \equiv 1 \pmod{n}.$$

Le petit théorème de Fermat en est un cas particulier ; si p est premier :

$$\forall a \in \mathbb{Z} \setminus p\mathbb{Z}\ ,\ a^{p-1} \equiv 1 \pmod{p},\ \text{d'où l'on déduit :}\ \forall a \in \mathbb{Z}\ ,\ a^p \equiv a \pmod{p}.$$

La relation $\forall a \in \mathbb{Z} \setminus p\mathbb{Z}\ ,\ a^{p-1} \equiv 1 \pmod{p}$ entraîne réciproquement que tout non multiple de p est premier avec p, donc que p est premier. Cependant, la deuxième relation, qui est conséquence immédiate de la première, n'implique pas la primalité de p : les *nombres de Carmichael*, comme $561 = 3 \times 11 \times 17$ la possèdent (proposition 39 de la page 43).

Application aux groupes cycliques

On sait que tout groupe cyclique d'ordre n est isomorphe à $(\mathbb{Z}/n\mathbb{Z}, +)$ et que les générateurs du groupe $(\mathbb{Z}/n\mathbb{Z}, +)$ sont les éléments de $(\mathbb{Z}/n\mathbb{Z})^*$ (section 2 du module « Groupes, anneaux, corps » du livre [L1]). On en déduit que tout groupe cyclique d'ordre n admet exactement $\varphi(n)$ générateurs.

Exemple. Il découlera de la section 2.2 que $(\mathbb{Z}/5\mathbb{Z})^*$ et $(\mathbb{Z}/7\mathbb{Z})^*$ sont cycliques. Ils ont respectivement $\varphi(5) = 4$ et $\varphi(7) = 6$ éléments. Ils sont donc respectivement isomorphes à $\mathbb{Z}/4\mathbb{Z}$ et $\mathbb{Z}/6\mathbb{Z}$. Le groupe $(\mathbb{Z}/5\mathbb{Z})^*$ a $\varphi(4) = 2$ générateurs (les classes de 2 et de 3) et le groupe $(\mathbb{Z}/7\mathbb{Z})^*$ en a $\varphi(6) = 2$ (les classes de 3 et de 5).

Proposition 13. Pour tout $n \in \mathbb{N}^*$, on a :

$$\sum_{d|n} \varphi(d) = n \qquad \text{(somme étendue aux diviseurs entiers naturels de } n).$$

Démonstration. Notons $E := \{0, \ldots, n-1\}$ et, pour $d \mid n$, $E_d := \{a \in E \mid a \wedge n = d\}$, de sorte que E est la réunion disjointe des E_d. Soit $d \mid n$ et soit $n' := n/d$. L'application $a' \mapsto da'$ définit une bijection de $\{0, \ldots, n'-1\}$ sur E_d. On a donc $\operatorname{card} E_d = \varphi(n/d)$ et n est la somme des $\varphi(n/d)$ pour $d \mid n$. Comme l'application $d \mapsto n/d$ permute l'ensemble des diviseurs de n, la formule voulue en découle. ∎

Une autre démonstration est proposée dans l'exercice I.1.23 de la page 70.

2.2 Le groupe multiplicatif d'un corps fini

Nous commencerons par un résultat un peu plus général que nécessaire.

Proposition 14. Soient K un corps commutatif et G un sous-groupe fini de K^*. Alors G est cyclique.

Démonstration. Notons $n := \operatorname{card} G$ et, pour tout diviseur d de n, $\psi(d)$ le nombre d'éléments de G qui sont d'ordre exactement d. Nous voulons démontrer que $\psi(n) \neq 0$.

D'après le théorème de Lagrange, l'ordre de tout élément de G est un diviseur de n, et donc $\sum_{d|n} \psi(d) = n$.

Soit $d \mid n$ tel que $\psi(d) \neq 0$ et soit $x \in G$ un élément d'ordre d. Le sous-groupe $<x>$ qu'il engendre a exactement d éléments, qui vérifient donc $x^d = 1$. En particulier, ils sont les racines du polynôme $X^d - 1$, qui en a au plus d *puisque le corps K est commutatif*. Ainsi $<x>$ est égal à l'ensemble de toutes les racines de ce polynôme. Tous les éléments d'ordre d de G sont donc dans $<x>$ et en sont des générateurs, d'où, en vertu de ce qui a été dit plus haut sur les groupes cycliques, l'égalité $\psi(d) = \varphi(d)$.

On a donc $\psi(d) = 0$ ou $\psi(d) = \varphi(d)$ pour tout $d \mid n$. Comme $\sum_{d|n} \psi(d) = \sum_{d|n} \varphi(d) = n$, on a en fait $\psi(d) = \varphi(d)$ pour tout $d \mid n$, d'où $\psi(n) = \varphi(n) \neq 0$. ∎

Ce résultat peut être en défaut si l'on ne suppose pas le corps commutatif (exercice I.1.24 de la page 70).

Corollaire 15. Soit K un corps fini commutatif[1]. Alors K^* est cyclique.

Exemple. Soit p un nombre premier. Un entier $a \in \mathbb{Z}$ dont la classe dans $\mathbb{Z}/p\mathbb{Z}$ est un générateur de $(\mathbb{Z}/p\mathbb{Z})^*$ est appelé *racine primitive modulo p*. Ces nombres sont utiles pour les tests de primalité (voir la section 5). On en trouve des tables, par exemple sur le site :

`http://fr.wikipedia.org/wiki/Racine_primitive_modulo_n`

2.3 Le groupe multiplicatif de l'anneau $\mathbb{Z}/p^r\mathbb{Z}$

Dans toute cette section, p désignera un nombre premier impair. Pour le cas des groupes $(\mathbb{Z}/2^r\mathbb{Z})^*$, des informations partielles sont proposées en exercice ; le cas général est traité dans [Demazure].

Lemme 16. L'élément $\overline{1+p} \in (\mathbb{Z}/p^r\mathbb{Z})^*$ est d'ordre p^{r-1}.

Démonstration. On va démontrer par récurrence sur i que $(1+p)^{p^i} = 1 + p^{i+1}x_i$, où $x_i \in \mathbb{Z} \setminus p\mathbb{Z}$. Pour $i = 0$, on a $x_0 = 1$. Supposons la relation vérifiée au rang $i \geqslant 1$. Alors :

$$(1+p)^{p^{i+1}} = \left(1 + p^{i+1}x_i\right)^p = 1 + p^{i+2}x_{i+1},$$

où :

$$x_{i+1} = x_i + \sum_{k=2}^{p-1} \binom{p}{k} p^{k(i+1)-(i+2)} x_i^k + p^{p(i+1)-(i+2)} x_i^p \equiv x_i \pmod{p},$$

car $p \mid \binom{p}{k}$ pour $0 < k < p$ et que $p(i+1) - (i+2) > 1$ puisque $p > 2$. La relation étant établie, on voit que $\left(\overline{1+p}\right)^{p^{r-2}} \neq \overline{1}$ (prendre $i = r-2$) et que $\left(\overline{1+p}\right)^{p^{r-1}} = \overline{1}$ (prendre $i = r-1$). ∎

[1] En fait, tout corps fini est commutatif (théorème de Wedderburn, voir le livre [MPA1]).

Théorème 17. Pour p premier impair, le groupe $(\mathbb{Z}/p^r\mathbb{Z})^*$ est cyclique.

Démonstration. Le morphisme surjectif d'anneaux $\mathbb{Z}/p^r\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$ induit un morphisme de groupes $(\mathbb{Z}/p^r\mathbb{Z})^* \to (\mathbb{Z}/p\mathbb{Z})^*$ qui est lui aussi surjectif, puisqu'à la source comme au but on trouve les classes des éléments de $\mathbb{Z} \setminus p\mathbb{Z}$. Soit x_0 un générateur de $(\mathbb{Z}/p\mathbb{Z})^*$ (corollaire 15 de la page précédente) et soit $y_0 \in (\mathbb{Z}/p^r\mathbb{Z})^*$ un antécédent de x_0. Puisque $y_0^k = 1 \Rightarrow x_0^k = 1$, l'ordre $p-1$ de x_0 dans $(\mathbb{Z}/p\mathbb{Z})^*$ divise l'ordre de y_0 dans $(\mathbb{Z}/p^r\mathbb{Z})^*$. Il existe donc une puissance z_0 de y_0 dont l'ordre est $p-1$. Alors l'ordre de $z_0\,\overline{1+p} \in (\mathbb{Z}/p^r\mathbb{Z})^*$ est $(p-1)p^{r-1} = \varphi(p^r)$ (exercice I.1.25 de la page 70) et $z_0\,\overline{1+p}$ est un générateur de $(\mathbb{Z}/p^r\mathbb{Z})^*$. ∎

Exercice 8.
Trouver un générateur de $(\mathbb{Z}/125\mathbb{Z})^*$.
Solution. Puisque (par exemple) la classe de 3 engendre $(\mathbb{Z}/5\mathbb{Z})^*$, on cherche l'ordre de $\overline{3} \in (\mathbb{Z}/125\mathbb{Z})^*$, qui doit être un multiple de 4 et un diviseur de $\varphi(125) = 100$. On constate que le reste de 3^{20} modulo 125 est 26 et l'on a donc directement un générateur.

2.4 Deux applications informatiques

2.4.1 Première application : générateurs pseudo-aléatoires

L'un des pères fondateurs de l'informatique, John von Neumann, a dit : *« Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin. »*[2]. Nous tenterons donc plus modestement de *simuler* l'aléatoire et parlerons de « générateur pseudo-aléatoire ». Dans ce but, nous décrirons des moyens de produire de longues suites de nombres variés (en utilisant l'arithmétique modulaire, et en particulier la structure du groupe $(\mathbb{Z}/m\mathbb{Z})^*$). Il y a pour cela une théorie riche et compliquée (et amusante), que nous effleurerons à peine. Pour une étude approfondie, voir [Knuth2] qui traite de plus des tests permettant d'évaluer la qualité d'un générateur (son aptitude à simuler l'aléatoire). Vu de l'extérieur, un générateur pseudo-aléatoire est un programme qui rend un nombre chaque fois qu'on l'appelle. Vu de l'intérieur, ce programme calcule en fait les termes d'une suite numérique (x_n). Chaque fois qu'on l'appelle, il renvoie un terme ; d'un appel sur l'autre, il passe au terme suivant. Autrement dit, entre deux appels, le programme garde trace du rang du dernier terme rendu : l'indice n est une *variable rémanente*, et il faut donc savoir programmer de telles variables. Le plus souvent, il s'agit de suites définies par récurrence : $x_{n+1} = f(x_n)$. Dans ce cas, la variable rémanente contient la dernière valeur de x_n qui a été rendue. Dans notre approche élémentaire, nous ne traiterons pas ce problème : nous considérerons simplement des programmes qui calculent $f(x)$, donc chaque nouveau terme en fonction du précédent. Il existe deux types principaux de générateurs aléatoires : ceux qui rendent des nombres réels entre 0 et 1 ; et ceux qui rendent des entiers entre 0 et $m-1$, l'entier $m \geqslant 2$ ayant été passé en argument. Nous ne nous occuperons que du deuxième type.

[2] « Quiconque envisage des méthodes arithmétiques pour produire des nombres aléatoires est, bien entendu, en état de péché. »

Générateurs linéaires congruentiels

Nous supposons fixé le *module* $m \geqslant 2$ (argument d'appel). Au lieu de considérer une suite d'entiers compris entre 0 et $m-1$, nous considérerons une suite d'éléments de $\mathbb{Z}/m\mathbb{Z}$. Le format général de cette suite est donné par la relation de récurrence :

$$\forall n \in \mathbb{N}\ ,\ x_{n+1} := ax_n + b.$$

Un générateur pseudo-aléatoire basé sur ce principe est appelé *linéaire congruentiel*. Il faut, pour le spécifier choisir le module m, la « graine », ou « semence », (en anglais : the seed) $x_0 \in \mathbb{Z}/m\mathbb{Z}$ et les coefficients $a, b \in \mathbb{Z}/m\mathbb{Z}$. On ne prendra évidemment pas $a := \overline{0}$ car la suite serait constante à partir du rang $n = 1$ et n'aurait pas du tout l'air aléatoire. On ne prendra pas non plus en général $a := 1$ (nous employons cette notation simplifiée pour $\overline{1}$), car on aurait $x_n = x_0 + nb$, et l'on repérerait trop facilement une régularité. La théorie des générateurs linéaires congruentiels n'est vraiment développée que pour le cas d'un module premier ou primaire (*i.e.* puissance d'un nombre premier).

Cas d'un module premier

Supposons m premier. Puisque l'on a exclu le cas où $a = 1$, l'élément $1 - a$ est inversible dans $\mathbb{Z}/m\mathbb{Z}$ et l'on va pouvoir appliquer la méthode vue en terminale pour les suites arithmético-géométriques. On cherche un « point fixe » $u = au + b$, d'où $u = b(1-a)^{-1}$. La relation de récurrence devient alors : $(x_{n+1} - u) := a(x_n - u)$, et l'on en déduit $x_n - u = a^n(x_0 - u)$, donc :

$$\forall n \in \mathbb{N}\ ,\ x_n = a^n(x_0 - u) + u.$$

Finalement, à un décalage près de valeur u, tout se passe comme si la suite était géométrique. Pour la suite de ce paragraphe, nous supposons donc que $b = u = \overline{0}$ et donc que $x_n = a^n x_0$. On prendra évidemment $x_0 \neq \overline{0}$, sinon la suite serait constante et n'aurait pas du tout l'air aléatoire. Finalement, on est donc conduit à examiner la suite des $a^n x_0$. Pour savoir si elle a l'air aléatoire, on trouvera dans [Knuth2] des tests empiriques. On peut déjà se demander si elle ne boucle pas trop vite. En effet, le petit théorème de Fermat nous dit que $a^{m-1} = 1$ et donc que, si $n = q(m-1) + r$, on a $a^n = (a^{m-1})^q a^r = a^r$, d'où $x_n = x_r$: cela signifie que la suite se reproduit avec une période $(m-1)$. Mais ce n'est pas nécessairement *la* période de cette suite.

Exemple. Prenons $m := 7$. Tout $a \in (\mathbb{Z}/7\mathbb{Z})^*$ vérifie $a^6 = \overline{1}$, donc la suite des $a^n x_0$ se reproduit à coup sûr avec une période 6. Mais si l'on prend par exemple $a := \overline{2}$, on a $a^3 = \overline{1}$ et la période tombe à 3. Il vaut mieux prendre $a := \overline{3}$ qui est tel que $a^0 = \overline{1}$, $a^1 = \overline{3}$, $a^2 = \overline{2}$, $a^3 = \overline{6}$, $a^4 = \overline{4}$, $a^5 = \overline{5}$ sont distincts et qui fournit donc une suite aussi longue que possible.

Dans tous les cas, il est raisonnable de choisir pour a un générateur du groupe cyclique $(\mathbb{Z}/m\mathbb{Z})^*$.

Cas d'un module primaire

Supposons m *primaire*, c'est-à-dire puissance d'un nombre premier : $n = p^r$, où p est premier et $r \geqslant 2$. Dans ce cas, il n'est pas aussi facile d'étudier une suite arithmético-géométrique générale (car $a \neq 1$ n'entraîne pas que $1 - a$ est inversible), aussi prendrons nous d'emblée $b := \overline{0}$. Et, bien entendu, on suppose toujours $a \neq \overline{0}, \overline{1}$ et $x_0 \neq \overline{0}$.

Si a est la classe d'un multiple de p, alors a^r est la classe d'un multiple de p^r, donc $a^r = \overline{0}$. Dans ce cas, la suite est stationnaire en $\overline{0}$ et donc très peu aléatoire. Nous supposons donc que a est la classe d'un entier non multiple de p. Mais un tel entier n'a pas de diviseur commun avec $p^r = m$, donc a est inversible : on a encore $a \in (\mathbb{Z}/m\mathbb{Z})^*$. la situation est analogue à celle décrite plus haut : la suite aura pour longueur l'ordre de a. Si p est impair, le groupe étant cyclique, on choisira donc pour a un générateur.

Exemple. Avec $m := 27$, on a $\varphi(m) = 18$. Les générateurs de $(\mathbb{Z}/27\mathbb{Z})^*$ sont $\overline{2}$, $\overline{5}$, $\overline{11}$, $\overline{14}$, $\overline{20}$ et $\overline{23}$ (en fait, les puissances de $\overline{2}$ avec un exposant premier à 18).

2.4.2 Deuxième application : cryptographie à clé publique

La méthode RSA de cryptographie à clé publique a été inventée en 1978 par Ron Rivest, Adi Shamir et Leonard Adleman du MIT. Sa mise en œuvre suppose le choix d'un *module* n, d'une *clé publique* (e, n) et d'une *clé secrète* (d, n). Les propriétés requises sont les suivantes :

1. L'entier naturel n est assez grand pour que l'on puisse facilement *coder* (voir la remarque ci-dessous) tout message sous la forme d'une suite d'entiers de $\{0, \ldots, n-1\}$.
2. Les entiers naturels d et e vérifient la propriété :

$$\forall a \in \mathbb{Z}\ ,\ a^{de} \equiv a \pmod{n}.$$

Nous distinguons ici le *codage* du *cryptage*. Le codage est simplement la mise du texte sous une forme adaptée à l'algorithme. Par exemple, chaque caractère du message à transmettre est représenté par un octet et le message est tronçonné en paquets d'octets de tailles telles que l'on puisse les représenter par un entier $M \in \{0, \ldots, n-1\}$. Le problème du cryptage est alors, pour l'émetteur du message, de remplacer M par un entier $M' := C(M) \in \{0, \ldots, n-1\}$ (la lettre C est ici pour « cryptage »), que le destinataire du message saura reconvertir en $M := D(M')$ (la lettre D est ici pour « décryptage »).

La clé publique (e, n) et la clé secrète (d, n) étant données, le principe est le suivant. Leur détenteur diffuse publiquement (!) la clé publique (e, n) et garde par devers lui (!) la clé secrète (d, n). Chaque fois que quelqu'un veut lui envoyer un message secret $M \in \{0, \ldots, n-1\}$:

1. Il crypte M selon la formule $M' := C(M) := M^e \pmod{n}$.
2. Le destinataire décrypte M' selon la formule $M := D(M') := (M')^d \pmod{n}$.

Ici, par abus de notation, $M^e \pmod{n}$ désigne le reste de la division de M^e par n, et similairement pour $(M')^d \pmod{n}$. En fait, ces calculs peuvent s'effectuer relativement vite par exponentiation rapide dans $\mathbb{Z}/n\mathbb{Z}$. La méthode RSA est correcte en vertu du calcul :

$$D(C(M)) = M^{de} \pmod{n} = M.$$

Cette relation est certainement vérifiée si $de \equiv 1 \pmod{\varphi(n)}$ (comme les choix ci-dessous l'impliquent) en vertu du théorème d'Euler (page 14) et si de plus M est premier avec n. Pour le cas général, voir l'exercice I.1.31 de la page 70.
Il reste à voir comment on détermine les clés publique et secrète :

1. On choisit deux grands nombres premiers distincts p et q, et l'on pose $n := pq$.
2. On choisit un entier e premier avec $n' := \varphi(n) = (p-1)(q-1)$ ni trop grand ni trop petit.
3. On détermine un entier d tel que $de \equiv 1 \pmod{n'}$ à l'aide de l'algorithme d'Euclide étendu.

Exemple. On choisit dans une table deux « grands » nombres premiers. Pour pouvoir effectuer les calculs à la main, nous nous contenterons de $p = 41$ et $q = 61$, d'où $n = 2501$ et $n' = 2400$. Les seuls facteurs premiers de n' étant 2, 3 et 5, on peut prendre (par exemple) $d = 7$. L'algorithme d'Euclide permet de trouver $e = 343$ (par miracle on est tombé sur l'égalité $n' = 7^4 - 1$ et donc ici $e = d^3$). Pour le codage, on décide d'accepter lettres (codées de 1 à 26) et chiffres (codés de 27 à 36). Donc chaque caractère isolé d'un texte alphanumérique est représenté par un nombre compris entre 1 et 36 (on ne tient pas compte de la différence majuscule-minuscule et l'on ne s'occupe pas des signes typographiques, etc.). Pour exploiter un peu mieux le grand nombre de codes disponibles, on décide enfin de découper le texte en paquets de deux caractères consécutifs. Quand deux tels caractères ont respectivement reçu les codes a_1, a_2 (compris entre 1 et 36) on attribue au paquet le code $a := 36a_1 + a_2 - 36$. On a appliqué à $(a_1 - 1, a_2 - 1)$ la bijection $(x_1, x_2) \mapsto 36x_1 + x_2$ de $\{0, \ldots, 35\}^2$ sur $\{0, \ldots, 36^2 - 1\}$. On voit donc que a est compris entre 1 et 1296, et que l'on peut retrouver a_1 et a_2 à partir de a en faisant la division euclidienne de $a-1$ par 36 : on a en effet $a-1 = 36(a_1-1)+(a_2-1)$ avec de plus la condition $0 \leqslant a_2 - 1 \leqslant 35$, donc le quotient et le reste sont respectivement $a_1 - 1$ et $a_2 - 1$. Pour transmettre par exemple le paquet « 5t », on code le caractère '5' par l'entier $a_1 = 31$, le caractère 't' par l'entier $a_2 = 20$ et donc le paquet par $a = 36 \times 31 + 20 - 36 = 1100$. C'est en fait l'entier M dans la description générale donnée plus haut. Puis on calcule $C(M)$, c'est-à-dire le reste de la division euclidienne de 1100^{343} par 2501. On n'est pas obligé pour cela de calculer 1100^{343}, qui est très gros . On peut effectuer tous les calculs modulo 2501, *i.e.* poser $a := \overline{1100} \in \mathbb{Z}/2501\mathbb{Z}$ et exécuter l'algorithme d'exponentiation dichotomique dans $\mathbb{Z}/2501\mathbb{Z}$. On pose donc $b_0 := b := \overline{1100}$, puis :

$$\begin{array}{llllll}
b_1 = b^2 & = b_0 \times b_0 & = \overline{1100} \times \overline{1100} & = \overline{1210000} & = \overline{2017}, \\
b_2 = b^5 & = b_1 \times b_1 \times b & = \overline{2017} \times \overline{2017} \times \overline{1100} & = \overline{4475117900} & = \overline{1069}, \\
b_3 = b^{10} & = b_2 \times b_2 & = \overline{1069} \times \overline{1069} & = \overline{1142761} & = \overline{2305}, \\
b_4 = b^{21} & = b_3 \times b_3 \times b & = \overline{2305} \times \overline{2305} \times \overline{1100} & = \overline{5844327500} & = \overline{704}, \\
b_5 = b^{42} & = b_4 \times b_4 & = \overline{704} \times \overline{704} & = \overline{495616} & = \overline{418}, \\
b_6 = b^{85} & = b_5 \times b_5 \times b & = \overline{418} \times \overline{418} \times \overline{1100} & = \overline{192196400} & = \overline{2053}, \\
b_7 = b^{171} & = b_6 \times b_6 \times b & = \overline{2053} \times \overline{2053} \times \overline{1100} & = \overline{4636289900} & = \overline{1126}, \\
b_8 = b^{343} & = b_7 \times b_7 \times b & = \overline{1126} \times \overline{1126} \times \overline{1100} & = \overline{1394663600} & = \overline{958}
\end{array}$$

et l'on transmet donc 958 (on a donc au total 13 multiplications de nombres de quatre

chiffres et divisions euclidiennes de nombres de dix chiffres par 2501, ce qui est faisable sur calculette ...). Pour décrypter ce nombre, il faut calculer le reste de la division euclidienne de 958^7 par 2501 :

$$\overline{958}\times\overline{958}\times\overline{958} = \overline{879217912} = \overline{1366}, \quad \text{puis} \quad \overline{1366}\times\overline{1366}\times\overline{958} = \overline{1787585848} = \overline{1100}.$$

On retrouve bien le code de départ. Pour le décoder, on effectue la division euclidienne $1099 = 36 \times 30 + 19$, d'où $a_1 = 30 + 1 = 31$ et $a_2 = 19 + 1 = 20$, ce qui correspond aux caractères '5' et 't' donc au paquet « 5t ».

Fiabilité de RSA

Cette méthode est praticable parce qu'il est « plutôt facile » de trouver des grands nombres premiers (voir la section 5). Si l'on savait aussi facilement factoriser n, la méthode de cryptage serait facile à casser. Dans l'article original de Rivest, Shamir et Adleman sur RSA (1978), la table suivante est donnée ; elle indique le nombre $c(n)$ d'opérations nécessaires pour factoriser un entier de n chiffres et le temps $t(n)$ correspondant, avec le meilleur algorithme de factorisation connu à l'époque et la technologie du moment :

n	50	75	100	200	300	500
$c(n)$	$1{,}4.10^{10}$	$9{,}0.10^{12}$	$2{,}3.10^{15}$	$1{,}2.10^{23}$	$1{,}5.10^{29}$	$1{,}3.10^{39}$
$t(n)$	3,9 heures	104 jours	74 ans	$3{,}8.10^{9}$ ans	$34{,}9.10^{15}$ ans	$34{,}2.10^{25}$ ans

Si l'on se fie à la « loi de Moore », selon laquelle la vitesse des machines double tous les dix-huit mois, il faudrait en 2014 plus de deux siècles pour factoriser un nombre de 200 chiffres. La situation est bien différente. Dans le livre « Mathématiques et Technologie » de Christiane Rousseau et Yvan Saint-Aubin (Springer, collection SUMAT) on trouve (au chapitre 7, consacré à RSA) des informations plus récentes ; par exemple, des nombres de 130 chiffres ont été factorisés en 1996, de 155 chiffres en 1999, de 200 chiffres en 2005. Le site `http://en.wikipedia.org/wiki/Integer_factorization` indique la factorisation en 2009 d'un entier de 232 chiffres en deux ans à l'aide de centaines de machines ; et, en 2019, la factorisation en 2009 d'un entier de 240 chiffres ayant coûté neuf cents années (cumulées) de calcul. Notons pour finir que factoriser n n'est pas la seule attaque possible contre RSA, mais cela est une autre histoire ...

Exercice 9.

On tente d'attaquer RSA en factorisant n de la manière suivante : on tire au hasard un entier $m \in \{1, \ldots, n\}$ et l'on calcule (avec Bézout) le pgcd de m et n. Que vaut cette méthode ?

Solution. Le coût du calcul du pgcd n'est pas très élevé (voir [Knuth2]). Mais la probabilité d'obtenir un résultat est faible : en effet, il y a échec si $m \wedge n = 1$, donc avec probabilité $\dfrac{\varphi(n)}{n} = (1-1/p)(1-1/q)$. Si p et q sont grands, la probabilité de succès $1/p+1/q-1/pq$ est très faible. Il faudrait un très grand nombre d'essais pour avoir une chance de succès ; voir le calcul de l'espérance du temps de premier succès dans une suite de tirages de Bernoulli qui est ici :

$$\frac{1}{\frac{1}{p}+\frac{1}{q}-\frac{1}{pq}} = \frac{pq}{p+q-1}.$$

3 Résidus quadratiques

3.1 Carrés dans un corps fini

Soit K un corps commutatif fini ; notons p sa caractéristique et q son cardinal. Nos applications concerneront uniquement le corps $\mathbb{Z}/p\mathbb{Z}$, (des corps finis plus généraux sont étudiés dans [MPA1]). Quelques propriétés élémentaires de K sont démontrées à l'exercice I.1.32 de la page 71.

L'application $x \mapsto x^2$ définit un endomorphisme du groupe commutatif K^*, dont le noyau est $\{x \in K^* \mid x^2 = 1\}$, c'est-à-dire l'ensemble $\{+1, -1\}$ des racines du polynôme $X^2 - 1 = (X-1)(X+1)$. Si K est de caractéristique $p = 2$, on a $+1 = -1$ et le noyau est trivial ; l'endomorphisme est alors injectif donc bijectif (puisque K est fini). C'est un exemple d'« automorphisme de Frobenius », voir l'exercice I.1.32.

Nous supposerons désormais que la caractéristique p de K est un nombre premier impair. Le noyau $\{+1, -1\}$ de notre endomorphisme ayant 2 éléments, son image $K^{*2} := \{x^2 \mid x \in K^*\}$ est un sous-groupe d'indice 2 de K^*, donc a $(q-1)/2$ éléments ; on invoque le théorème d'isomorphisme $G/\operatorname{Ker} f \simeq \operatorname{Im} f$ (premier théorème d'isomorphisme pour les groupes, du module « Compléments d'algèbre » du livre [L2]).

D'après le théorème de Lagrange, tout $x \in K^*$ vérifie $x^{q-1} = 1$, donc $(x^2)^{(q-1)/2} = 1$. Donc les éléments de K^{*2} sont tous racines du polynôme $X^{(q-1)/2} - 1$. Comme celui-ci admet au plus $(q-1)/2$ racines, K^{*2} est exactement l'ensemble de ces racines.

Proposition 18.
L'endomorphisme $x \mapsto x^{(q-1)/2}$ de K^* a pour noyau K^{*2} et pour image $\{+1, -1\}$.

Démonstration. L'assertion sur le noyau découle de la discussion précédente. Comme $(x^{(q-1)/2})^2 = x^{q-1} = 1$, on a $x^{(q-1)/2} = \pm 1$. Comme l'image a 2 éléments (encore en vertu du premier théorème d'isomorphisme), c'est $\{+1, -1\}$. ∎

Exercice 10.
Démontrer la proposition en utilisant seulement le fait que K^* est cyclique.
Solution. Cela revient à poser $n := q-1$, qui est pair : $n = 2m$, et à établir que dans $\mathbb{Z}/n\mathbb{Z}$ les endomorphismes $x \mapsto 2x$ et $x \mapsto mx$ ont chacun pour noyau l'image de l'autre, le premier noyau ayant 2 éléments et le deuxième m éléments.

Corollaire 19. Pour tout $a \in K$, il existe $x, y \in K$ tels que $a = x^2 + y^2$.

Démonstration. L'ensemble K^2 des x^2, $x \in K$, a $(q-1)/2 + 1 = (q+1)/2$ éléments (on compte ici le zéro). L'ensemble $a - K^2$ des $a - y^2$, $y \in K$, a donc également $(q+1)/2$ éléments. Comme $\operatorname{card} K^2 + \operatorname{card}(a - K^2) = q + 1 > \operatorname{card} K$, ces ensembles ne sont pas disjoints. ∎

Exemple. Les carrés de $(\mathbb{Z}/5\mathbb{Z})^*$ sont $\overline{1}$ et $\overline{4}$, ce sont également les éléments x tels que $x^{(q-1)/2} = x^2 = 1$ (ici abréviation pour $\overline{1}$); alors que les non carrés $\overline{2}$ et $\overline{3}$ ont pour carré -1. Les carrés de $(\mathbb{Z}/7\mathbb{Z})^*$ sont $\overline{1}$, $\overline{2}$ et $\overline{4}$, ce sont également les éléments x tels que $x^{(q-1)/2} = x^3 = 1$; alors que les non carrés $\overline{3}$, $\overline{5}$ et $\overline{6}$ ont pour cube -1.

3.2 Symbole de Legendre

Définition 3. Soit p un nombre premier impair.
(i) On dit que $a \in \mathbb{Z} \setminus p\mathbb{Z}$ est *résidu quadratique modulo* p s'il existe $b \in \mathbb{Z}$ tel que $a \equiv b^2 \pmod{p}$, autrement dit, si $\overline{a} \in (\mathbb{Z}/p\mathbb{Z})^*$ est un carré.
(ii) Soit $a \in \mathbb{Z}$ quelconque. Le *symbole de Legendre* $\left(\dfrac{a}{p}\right) \in \mathbb{Z}$ est défini par les formules suivantes :

$$\left(\frac{a}{p}\right) := \begin{cases} 0 \text{ si } a \in p\mathbb{Z}, \\ +1 \text{ si } a \notin p\mathbb{Z} \text{ et } a \text{ résidu quadratique modulo } p, \\ -1 \text{ si } a \notin p\mathbb{Z} \text{ et } a \text{ non résidu quadratique modulo } p. \end{cases}$$

Il est clair que $\left(\dfrac{a}{p}\right)$ ne dépend que de la classe de a modulo p ; si $x \in \mathbb{Z}/p\mathbb{Z}$ désigne cette classe, on notera donc encore $\left(\dfrac{x}{p}\right) := \left(\dfrac{a}{p}\right)$.

Exemple. Pour que le nombre premier naturel $p \in \mathbb{N}^*$ reste premier dans $\mathbb{Z}[\sqrt{d}]$, il faut, et il suffit, que $\left(\dfrac{d}{p}\right) = -1$ (exercice 3 de la page 8).

Théorème 20. Soient p premier impair et $a \in \mathbb{Z} \setminus p\mathbb{Z}$. Alors :

$$\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p}.$$

Démonstration. Soit $x \in (\mathbb{Z}/p\mathbb{Z})^*$ la classe de a. D'après la proposition 18 de la page précédente, $\left(\dfrac{x}{p}\right)$ vaut 1 si $x^{(p-1)/2} = 1$ et -1 si $x^{(p-1)/2} = -1$, autrement dit, $x^{(p-1)/2}$ est la classe de $\left(\dfrac{x}{p}\right)$ modulo p, ce qui équivaut à la relation à démontrer. ∎

Corollaire 21. Le symbole de Legendre est multiplicatif :

$$\forall a, b \in \mathbb{Z}\ ,\ \left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right).$$

Démonstration. Si $a, b \notin p\mathbb{Z}$ cela découle du théorème, sinon c'est immédiat. ∎

Corollaire 22 (Euler). On a $\left(\dfrac{-1}{p}\right) = (-1)^{(p-1)/2}$. Par conséquent, -1 est résidu quadratique, respectivement non résidu quadratique modulo p si, et seulement si, $p \equiv 1 \pmod 4$, respectivement $p \equiv -1 \pmod 4$.

Démonstration. Pour les nombres ± 1, la congruence modulo $p \geqslant 3$ équivaut à l'égalité. Par ailleurs, la congruence $p \equiv \pm 1 \pmod 4$ traduit la parité de $(p-1)/2$. ■

Le corollaire dit que, pour que l'on puisse résoudre la congruence $a^2 + 1 \equiv 0 \pmod p$, il faut et il suffit que p soit de la forme $4k+1$; mais il ne fournit pas de solution.

Exercice 11.
Si $p = 4k+1$, montrer que $a := (2k)!$ est solution. On rappelle le théorème de Wilson (module II.1 du livre [L1]) : si p est premier, $(p-1)! \equiv -1 \pmod p$.

Solution. Ici, $(p-1)! = (2k)!A$, où $A := \prod_{i=1}^{2k}(2k+i) \equiv \prod_{i=1}^{2k}(2k+i-p) \pmod p$ et $\prod_{i=1}^{2k}(2k+i-p) = \prod_{i=1}^{2k}(i-2k-1) = a$, d'où $(p-1)! \equiv a^2 \pmod p$, d'où $a^2 \equiv -1 \pmod p$ en vertu du théorème de Wilson.

Pour calculer $\left(\dfrac{2}{p}\right)$, nous ferons appel à un lemme dû à Gauß et qui nous servira aussi pour démontrer la loi de réciprocité quadratique. Écrivons $p = 2m+1$. Pour tout $a \in \mathbb{Z} \setminus p\mathbb{Z}$, le reste b de la division euclidienne de a par p appartient à l'un des deux intervalles $E' := \{1, \ldots, m\}$, $E'' := \{m+1, \ldots, 2m\}$. Notons $r(a) := b$ si $b \in E'$ et $r(a) := b - p$ si $b \in E''$. L'entier $r(a)$ est caractérisé par les relations : $a \equiv r(a) \pmod p$ et $1 \leqslant |r(a)| \leqslant m$.

Lemme 23 (Gauß). Soient $a \in \mathbb{Z} \setminus p\mathbb{Z}$ et soit $\ell := \operatorname{card}\{i = 1, \ldots, m \mid r(ia) < 0\}$. Alors :

$$\left(\frac{a}{p}\right) = (-1)^{\ell}.$$

Démonstration. Le produit $\prod_{i=1}^{m}(ai)$ est égal à $m!a^m$, donc congru modulo p à $m!\left(\dfrac{a}{p}\right)$ d'après le théorème 20 (puisque $m = (p-1)/2$). Mais $\prod_{i=1}^{m}(ai)$ est également congru modulo p au produit $\prod_{i=1}^{m} r(ai)$, qui est égal à $(-1)^{\ell}\prod_{i=1}^{m}|r(ai)|$.

On remarque d'autre part que l'application $i \mapsto |r(ai)|$ de E' dans lui-même est injective ; en effet, puisque p ne divise pas a :

$$r(ai) \equiv \pm r(aj) \pmod p \iff ai \equiv \pm aj \pmod p \iff i \equiv \pm j \pmod p \iff i = j,$$

pour $i, j \in E'$. Cette application est donc une permutation de E' et $\prod_{i=1}^{m} |r(ai)| = m!$.

Des deux congruences ci-dessus, on déduit :

$$m!\left(\frac{a}{p}\right) \equiv (-1)^{\ell} m! \pmod{p} \Longrightarrow \left(\frac{a}{p}\right) \equiv (-1)^{\ell} \pmod{p},$$

puisque $m!$ est inversible modulo p. ■

Exercice 12.

Calculer $\left(\frac{5}{7}\right)$ par cette méthode.

Solution. Ici, $m = 3$. Les restes des divisions des $5i$ par 7 pour $i = 1, \ldots, 6$ sont $5, 3, 1, 6, 4, 2$. Les $r(ai)$ sont $-2, 3, 1, -1, -3, 2$. Parmi les $m =$ premiers, *i.e.* ic les trois premiers, un seul est négatif, donc $\ell = 1$ et $\left(\frac{5}{7}\right) = -1$.

On le vérifie par la formule d'Euler : $5^{(7-1)/2} = 125 = 7 \times 18 - 1$, d'où $\left(\frac{5}{7}\right) = -1$.

On prend maintenant $a := 2$. On voit que $r(2i) = 2i$ si, et seulement si, $2i \leqslant m$, *i.e.* si $i \leqslant \lfloor m/2 \rfloor$ (partie entière de $m/2$). De même, $r(2i) = 2i - p$ si, et seulement si, $m < 2i \leqslant 2m$, *i.e.* si $\lfloor m/2 \rfloor < i \leqslant m$. Avec les notations du lemme, on a donc $\ell = m - \lfloor m/2 \rfloor = \lfloor (m+1)/2 \rfloor$ (le lecteur vérifiera sans peine que cette dernière égalité est vraie pour tout m).

Corollaire 24 (Gauß). On a $\left(\frac{2}{p}\right) = (-1)^{\lfloor (p+1)/4 \rfloor} = (-1)^{(p^2-1)/8}$ et les équivalences :

$$\left(\frac{2}{p}\right) = 1 \Longleftrightarrow p \equiv \pm 1 \pmod{8}, \quad \left(\frac{2}{p}\right) = -1 \Longleftrightarrow p \equiv \pm 3 \pmod{8}.$$

Démonstration. Puisque $p = 2m + 1$, on a $(m+1)/2 = (p+1)/4$ et la première égalité résulte de l'étude précédente. Comme $(p^2 - 1)/8 = (m^2 + m)/2$, il reste à vérifier que cet entier et $\lfloor (m+1)/2 \rfloor$ ont même parité. Si $m = 2k$, leur différence vaut $2k^2$; si $m = 2k-1$, elle vaut $2k^2 - 2k$. La suite est alors immédiate par examen des quatre cas possibles. ■

3.3 Loi de réciprocité quadratique

Dans chacun des deux exemples traités complètement ($a = -1$ et $a = 2$), on voit que la condition $\left(\frac{a}{p}\right) = +1$, respectivement $\left(\frac{a}{p}\right) = -1$, organise l'ensemble des nombres premiers impairs en classes de congruence : modulo 4 dans le premier cas, modulo 8 dans le second cas. Voici un autre exemple, dans lequel nous répartissons les nombres premiers impairs (et non multiples de 3) inférieurs à 100 selon la valeur de $\left(\frac{3}{p}\right)$:

$\left(\frac{3}{p}\right) = +1$	11	13	23	37	47	59	61	71	73	83	97	
$\left(\frac{3}{p}\right) = -1$	5	7	17	19	29	31	41	43	53	67	79	89

On constate que chacune des deux lignes s'organise en classes modulo 12 : les classes $\pm\overline{1}$ dans la première ligne, les classes $\pm\overline{5}$ dans la deuxième.

Un peu d'Histoire

C'est ce genre d'observation qui conduisit Euler à formuler la loi de réciprocité quadratique, que démontra Gauß dans ses extraordinaires *Disquisitiones Arithmeticae*. Ce dernier y attachait tant d'importance qu'il en donna huit démonstrations. Cette importance fut confirmée par l'histoire ultérieure de la théorie des nombres : les « lois de réciprocité généralisées » donnèrent naissance à la théorie du corps de classes au début du XXe siècle, puis à une impressionnante extension non abélienne, le « programme de Langlands » qui est au cœur d'une grande partie des mathématiques contemporaines.

Théorème 25 (Loi de réciprocité quadratique, Gauß). Soient p, q deux nombres premiers impairs distincts. Alors :

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}.$$

Notons que, chacun des trois facteurs présents valant ± 1, on peut aussi bien écrire que le produit de deux quelconques d'entre eux est égal au troisième ou encore que le produit des trois vaut $+1$. De plus, $\frac{p-1}{2}\frac{q-1}{2}$ étant impair si, et seulement si, $p, q \equiv -1 \pmod 4$, la loi de réciprocité se traduit par :

$$\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right) \qquad \text{sauf si } p, q \equiv -1 \pmod 4 \quad \text{auquel cas } \left(\frac{p}{q}\right) = -\left(\frac{q}{p}\right).$$

Avant d'en donner la preuve, montrons comment ce théorème « explique » l'observation ci-dessus. Il nous dit que :

$$\left(\frac{3}{p}\right)\left(\frac{p}{3}\right) = (-1)^{\frac{p-1}{2}\frac{3-1}{2}} = (-1)^{\frac{p-1}{2}} \Longrightarrow \left(\frac{3}{p}\right) = (-1)^{\frac{p-1}{2}}\left(\frac{p}{3}\right).$$

Le premier facteur $(-1)^{\frac{p-1}{2}}$ dépend de la classe de p modulo 4, le deuxième $\left(\frac{p}{3}\right)$ de sa classe modulo 3 : le produit dépend donc bien de la classe modulo 12.

Exercice 13.

Calculer $\left(\frac{3}{p}\right)$ dans tous les cas.

Solution. Le facteur $(-1)^{\frac{p-1}{2}}$ vaut $+1$ si $p \equiv 1 \pmod 4$ et -1 si $p \equiv -1 \pmod 4$. Le facteur $\left(\frac{p}{3}\right) \equiv p^{(3-1)/2} \pmod 3$ vaut $+1$ si $p \equiv 1 \pmod 3$ et -1 si $p \equiv -1 \pmod 3$.

Ainsi $\left(\dfrac{3}{p}\right)$ vaut $+1$ si $p \equiv \pm 1 \pmod{12}$ et -1 dans les autres cas : c'est bien ce que nous avons observé.

Démonstration de la loi de réciprocité quadratique

Elle relève de la « géométrie des nombres » . Nous écrivons $p = 2k+1$, $q = 2l+1$ et :

$$m := \operatorname{card}\{i \in \{1,\dots,k\} \mid \exists r \in \{-k,\dots,-1\} \ : \ iq \equiv r \pmod{p}\},$$

$$n := \operatorname{card}\{j \in \{1,\dots,l\} \mid \exists s \in \{-l,\dots,-1\} \ : \ jp \equiv s \pmod{q}\},$$

de sorte que, d'après le lemme 23, $\left(\dfrac{q}{p}\right) = (-1)^m$ et $\left(\dfrac{p}{q}\right) = (-1)^n$.

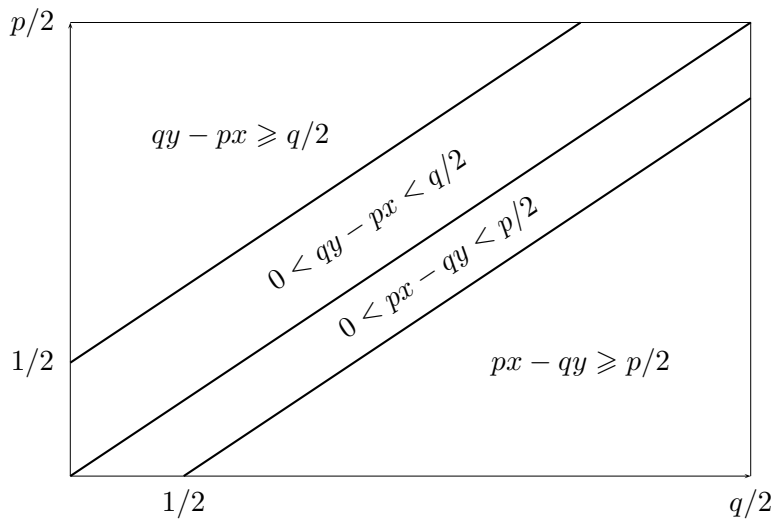
Si $i \in \{1,\dots,k\}$ appartient à l'ensemble à m éléments ci-dessus, il existe un unique entier j tel que $iq = jp + r$ et cet entier vérifie $1 \leqslant jp - iq \leqslant k$. Réciproquement, l'existence d'un tel j garantit que i est dans cet ensemble, qui est donc en bijection avec l'ensemble des couples (i,j) d'entiers tels que $1 \leqslant i \leqslant k$, ce qui équivaut à $0 < i < p/2$; et $1 \leqslant jp - iq \leqslant k$, ce qui équivaut à $0 < jp - iq < p/2$. On déduit d'ailleurs de ces conditions que $0 < j < \frac{1}{p}(\frac{p}{2} + q\frac{p}{2}) = l+1$, donc que $0 < j < q/2$. Cela donne une nouvelle formule pour m :

$$m = \operatorname{card}\{(i,j) \in \mathbb{N}^2 \mid 0 < i < p/2, 0 < j < q/2, 0 < jp - iq < p/2\}.$$

Symétriquement :

$$n = \operatorname{card}\{(i,j) \in \mathbb{N}^2 \mid 0 < i < p/2, 0 < j < q/2, 0 < iq - jp < q/2\}.$$

Le rectangle ouvert de $\mathbb{R}^2$ défini par les inégalités $0 < i < p/2$, $0 < j < q/2$ contient au total kl points entiers. L'égalité à démontrer est $(-1)^m(-1)^n = (-1)^{kl}$; elle équivaut à la parité de $kl - (m+n)$, c'est-à-dire du nombre de points entiers du rectangle qui ne vérifient aucune des deux conditions $0 < jp - iq < p/2$, $0 < iq - jp < q/2$.



Comme, pour des points entiers, $jp - iq \geqslant p/2 \iff jp - iq \geqslant k+1$, etc, ceux de ces points qui sont situés hors des deux bandes jouxtant la première diagonale forment les deux triangles respectivement définis par les triples inégalités :

$$\big(jp - iq \geqslant k+1, i \geqslant 1, j \leqslant l\big) \text{ et } \big(iq - jp \geqslant l+1, j \geqslant 1, i \leqslant k\big).$$

Nous allons établir une bijection entre ces deux ensembles de points entiers, ce qui prouvera que leur réunion a un nombre pair d'éléments, donc que $kl - (m+n)$ est pair.

Pour cela nous introduisons l'involution $(i,j) \mapsto (i',j')$ de $\mathbb{R}^2$ (en fait une symétrie axiale) définie par $i + i' = k+1$, $j + j' = l+1$. Cette involution permute $\mathbb{Z}^2$ et transforme la première triple inégalité ci-dessus en la deuxième, et réciproquement ; en effet, il est évident que :

$$i \geqslant 1 \iff i' \leqslant k, \quad j \leqslant l \iff j' \geqslant 1, \quad jp - iq \geqslant k+1 \iff i'q - j'p \geqslant l+1,$$

cette dernière équivalence résultant de l'égalité :

$$(jp - iq) - (i'q - j'p) = (l+1)p - (k+1)q = k - l. \qquad \blacksquare$$

La loi de réciprocité quadratique peut fournir un moyen rapide de calculer les symboles de Legendre (la méthode sera encore perfectionnée plus loin grâce au symbole de Jacobi).

Exemple. On veut calculer $\left(\dfrac{37}{97}\right)$. Outre la loi de réciprocité, on utilisera la multiplicativité, le fait que $\left(\dfrac{a}{p}\right)$ ne dépend que de la classe de a modulo p et les valeurs spéciales $\left(\dfrac{-1}{p}\right)$ et $\left(\dfrac{2}{p}\right)$:

$$\left(\frac{37}{97}\right) = (-1)^{\frac{37-1}{2}\frac{97-1}{2}}\left(\frac{97}{37}\right) = \left(\frac{97}{37}\right) = \left(\frac{60}{37}\right) = \left(\frac{4}{37}\right)\left(\frac{3}{37}\right)\left(\frac{5}{37}\right).$$

On a bien entendu $\left(\dfrac{4}{37}\right) = +1$ (tout carré est un résidu quadratique). On a déjà vu que $\left(\dfrac{3}{37}\right) = +1$. Il reste à calculer :

$$\left(\frac{5}{37}\right) = \left(\frac{37}{5}\right) = \left(\frac{2}{5}\right) = -1,$$

pour conclure que $\left(\dfrac{37}{97}\right) = -1$.

3.4 Symbole de Jacobi

Le symbole de Jacobi est une généralisation du symbole de Legendre qui facilite les calculs, mais qui a de plus sa propre importance dans les tests de primalité, ce dont nous donnerons un exemple à la section 5. Soient $a \in \mathbb{Z}$ quelconque et $n = p_1^{r_1} \cdots p_k^{r_k}$ la décomposition

en facteurs premiers d'un entier naturel impair n. Le symbole de Jacobi est défini par la formule :

$$\left(\frac{a}{n}\right) := \prod_{i=1}^{k} \left(\frac{a}{p_i}\right)^{r_i}.$$

De manière équivalente : $\left(\frac{a}{1}\right) = 1$, $\left(\frac{a}{p}\right)$ est identique au symbole de Legendre quand p est premier impair, et $\left(\frac{a}{nn'}\right) = \left(\frac{a}{n}\right)\left(\frac{a}{n'}\right)$ quels que soient n, n' impairs.
C'est donc un élément de $\{-1, 0, +1\}$ qui s'annule si, et seulement si, a et n ont un diviseur commun non trivial. Des propriétés analogues du symbole de Legendre, on déduit que $\left(\frac{ab}{n}\right) = \left(\frac{a}{n}\right)\left(\frac{b}{n}\right)$ et que $a \equiv a' \pmod{n} \Rightarrow \left(\frac{a}{n}\right) = \left(\frac{a'}{n}\right)$.

Proposition 26.
(i) On a les formules :

$$\left(\frac{-1}{n}\right) = (-1)^{(n-1)/2} \qquad \text{et} \qquad \left(\frac{2}{n}\right) = (-1)^{(n^2-1)/8}.$$

(ii) Si m et n sont tous deux impairs, on a la formule de réciprocité :

$$\left(\frac{m}{n}\right)\left(\frac{n}{m}\right) = (-1)^{\frac{m-1}{2}\frac{n-1}{2}}.$$

Démonstration. (i) Chaque formule est vraie lorsque n est premier. Il suffit donc de montrer que, si elle est vraie pour n et n', elle l'est encore pour nn'. Comme le membre gauche est multiplicatif en n, il suffit de vérifier que le membre droit l'est. Cela équivaut à la parité des expressions $(nn'-1)/2 - (n-1)/2 - (n'-1)/2$ et $(n^2n'^2-1)/8 - (n^2-1)/8 - (n'^2-1)/8$. Remplaçant n par $2k+1$ et n' par $2k'+1$, on voit que ces expressions valent respectivement $2kk'$ et $2(k^2+k)(k'^2+k')$, ce qui achève la preuve.

(ii) On utilise d'abord la multiplicativité en n des deux fonctions $\left(\frac{m}{n}\right)$ et $\left(\frac{n}{m}\right)$ (à m fixé) pour se ramener au cas où n est premier. Pour cela, comme en (i), il suffit de vérifier la parité de $\frac{m-1}{2}\frac{nn'-1}{2} - \frac{m-1}{2}\frac{n-1}{2} - \frac{m-1}{2}\frac{n'-1}{2}$ lorsque m, n, n' sont impairs. On écrit donc $n = 2k+1$ et $n' = 2k'+1$, et l'on trouve $(m-1)kk'$ qui est bien pair puisque m est impair. De la même manière on se ramène ensuite au cas où m est premier. On n'a plus alors qu'à invoquer la loi de réciprocité quadratique pour le symbole de Legendre. ■

Lien du symbole de Jacobi avec les résidus quadratiques

Si m est résidu quadratique (c'est-à-dire congru à un carré) modulo n, on voit facilement que $\left(\frac{m}{n}\right) = 1$. La réciproque est évidemment vraie si n est premier, mais pas en général.

Exercice 14.
Vérifier que $\left(\frac{20}{21}\right) = 1$ mais que 20 n'est pas résidu quadratique modulo 21.

Solution. En utilisant toutes les multiplicativités :

$$\left(\frac{20}{21}\right) = \left(\frac{4}{21}\right)\left(\frac{5}{21}\right) = \left(\frac{5}{21}\right) = \left(\frac{5}{3}\right)\left(\frac{5}{7}\right) = 1,$$

par calcul direct de $\left(\frac{5}{3}\right) = \left(\frac{2}{3}\right) = -1$ et d'après un calcul antérieur de $\left(\frac{5}{7}\right) = -1$. Cependant $\left(\frac{20}{3}\right) = \left(\frac{2}{3}\right) = -1$, donc 20 n'est pas résidu quadratique modulo 3 et *a fortiori* il ne l'est pas modulo 21. De même $\left(\frac{20}{7}\right) = \left(\frac{-1}{7}\right) = -1$.

De la même manière, on n'a pas nécessairement $\left(\frac{m}{n}\right) \equiv m^{(n-1)/2} \pmod{n}$. En fait, $m^{(n-1)/2}$ n'est même pas nécessairement congru à ± 1 modulo n, comme on le voit par exemple en prenant $m = 2$ et $n = 15$. Cependant, les cas d'égalité jouent un rôle crucial dans le test de primalité de Solovay-Strassen, que nous étudierons à la section 5, en vertu du résultat suivant :

Théorème 27. Si l'entier impair n est tel que $\left(\frac{m}{n}\right) \equiv m^{(n-1)/2} \pmod{n}$ pour tout entier m tel que $m \wedge n = 1$, alors n est premier.

Démonstration. En élevant la relation au carré, on voir que tout élément $x \in (\mathbb{Z}/n\mathbb{Z})^*$ vérifie $x^{n-1} = 1$ (autrement dit, n est un nombre de Carmichael, voir la proposition 39 de la page 43). Si p^2 divisait n, alors la composante $(\mathbb{Z}/p^r\mathbb{Z})^*$ de $(\mathbb{Z}/n\mathbb{Z})^*$, qui a $p^{r-1}(p-1)$ éléments avec $r - 1 \geqslant 2$, admettrait un élément d'ordre p ; or p ne divise par $n - 1$, contredisant l'égalité ci-dessus.

Donc $n = p_1 \cdots p_k$ (premiers distincts). Si $k \geqslant 1$, à l'aide du lemme chinois, on peut trouver $a \in \mathbb{Z}$ tel que $a \equiv 1 \pmod{p_2 \cdots p_k}$ et a non résidu modulo p_1.

Alors $\left(\frac{a}{p_1}\right) = -1$, $\left(\frac{a}{p_2 \cdots p_k}\right) = 1$, donc $\left(\frac{a}{n}\right) = -1$, donc (par l'hypothèse faite sur n) $a^{(n-1)/2} \equiv -1 \pmod{n}$, donc $a^{(n-1)/2} \equiv -1 \pmod{p_2 \cdots p_k}$, ce qui est impossible puisque $a \equiv 1 \pmod{p_2 \cdots p_k}$. ∎

4 Sommes de deux carrés

4.1 Rappels sur l'anneau des entiers de Gauß

C'est le sous-anneau de $\mathbb{C}$ engendré par i :

$$\mathbb{Z}[\mathrm{i}] = \{a + b\mathrm{i} \mid a, b \in \mathbb{Z}\}.$$

Le morphisme $P \mapsto P(\mathrm{i})$ de $\mathbb{Z}[X]$ dans $\mathbb{C}$ a pour image $\mathbb{Z}[\mathrm{i}]$ et pour noyau $<X^2+1>$, d'où un isomorphisme $\mathbb{Z}[X]/<X^2+1> \simeq \mathbb{Z}[\mathrm{i}]$. Son corps des fractions est $\mathbb{Q}(i) = \{a + b\mathrm{i} \mid a, b \in \mathbb{Q}\}$, c'est-à-dire le sous-corps de $\mathbb{C}$ engendré par i, une $\mathbb{Q}$-algèbre de dimension 2.
La *norme algébrique* de $z = a + b\mathrm{i} \in \mathbb{Z}[\mathrm{i}]$ est l'entier naturel $N(z) := z\overline{z} = a^2 + b^2$. Elle est multiplicative[3] : $N(uv) = N(u)N(v)$. On en déduit d'abord que le groupe des unités de $\mathbb{Z}[\mathrm{i}]$ est $\mathbb{Z}[\mathrm{i}]^* = \{u \in \mathbb{Z}[\mathrm{i}] \mid N(u) = 1\}$. En effet, si $u \in \mathbb{Z}[\mathrm{i}]$ est tel que $N(u) = 1$ alors $u^{-1} = \overline{u} \in \mathbb{Z}[\mathrm{i}]$, donc $u \in \mathbb{Z}[\mathrm{i}]^*$; et réciproquement, si $u \in \mathbb{Z}[\mathrm{i}]^*$, il existe $v \in \mathbb{Z}[\mathrm{i}]$ tel que $uv = 1$, d'où $N(u)N(v) = N(1) = 1$, d'où $N(u) = 1$ puisque $N(u), N(v) \in \mathbb{N}$. On en déduit ensuite :

$$\mathbb{Z}[\mathrm{i}]^* = \{+1, -1, +\mathrm{i}, -\mathrm{i}\},$$

puisque, notant $u = a + b\mathrm{i}$, on doit résoudre l'équation $a^2 + b^2 = 1$ avec $a, b \in \mathbb{Z}$, ce qui implique $(a, b) = (\pm 1, 0)$ ou $(0, \pm 1)$.
L'application $N : \mathbb{Z}[\mathrm{i}] \setminus \{0\} \to \mathbb{N}$ est un stathme euclidien. L'anneau $\mathbb{Z}[\mathrm{i}]$ est donc euclidien, principal et factoriel. Notre prochaine tâche sera donc de déterminer les éléments irréductibles (et *ipso facto* premiers) de $\mathbb{Z}[\mathrm{i}]$. Une étape préliminaire est de déterminer comment se comportent les nombres premiers (*i.e.* les entiers naturels premiers) quand on les considère comme des éléments de $\mathbb{Z}[\mathrm{i}]$. La terminologie que nous emploierons dans ce cas particulier (voir le théorème ci-dessous) est celle de la théorie des nombres algébriques.

Lemme 28. Soit K un corps de caractéristique différente de 2.
Le trinôme $X^2 + aX + b \in K[X]$ est irréductible si, et seulement si, $a^2 - 4b$ n'est pas un carré dans K.

Démonstration. Un polynôme de degré 2 est réductible si, et seulement s'il a une racine ; or $x^2 + ax + b = 0 \Leftrightarrow (2x + a)^2 = a^2 - 4b$. ■

Lemme 29. Soit p un nombre premier. Alors $\mathbb{Z}[\mathrm{i}]/<p> \simeq \mathbb{F}_p[X]/<X^2+1>$.

Démonstration. D'après le deuxième théorème d'isomorphisme pour les anneaux, ces deux anneaux sont isomorphes à $\mathbb{Z}[X]/<p, X^2+1>$. ■

Théorème 30. Soit p un nombre premier. Alors :
(i) Si $p = 2$, p est *ramifié* dans $\mathbb{Z}[\mathrm{i}]$, *i.e.* il y est divisible par un carré non trivial ; plus précisément : $p \sim (1 + \mathrm{i})^2$.
(ii) Si $p \equiv -1 \pmod 4$, p est *inerte* dans $\mathbb{Z}[\mathrm{i}]$, *i.e.* il y est premier.
(iii) Si $p \equiv 1 \pmod 4$, p est *décomposé* dans $\mathbb{Z}[\mathrm{i}]$, *i.e.* il y est produit d'éléments premiers distincts ; plus précisément : $p = \pi\pi'$ où $\pi, \pi' \in \mathbb{Z}[\mathrm{i}]$ sont irréductibles, conjugués et non équivalents.

Démonstration. (i) On a $p = (1 + \mathrm{i})(1 - \mathrm{i}) = -\mathrm{i}(1 + \mathrm{i})^2$. Nous savons que $1 + \mathrm{i}$ n'est

[3] Ici au sens de l'algèbre, à ne pas confondre avec la notion de fonction arithmétique multiplicative définie page 14.

pas inversible. En fait, il est irréductible :

$$\begin{aligned} 1+\mathrm{i} = uv &\Longrightarrow 2 = N(1+\mathrm{i}) = N(u)N(v) \\ &\Longrightarrow N(u) = 1 \text{ ou } N(v) = 1 \\ &\Longrightarrow u \in \mathbb{Z}[\mathrm{i}]^* \text{ ou } v \in \mathbb{Z}[\mathrm{i}]^*. \end{aligned}$$

(ii) Si $p \equiv -1 \pmod 4$, on sait que $\left(\dfrac{-1}{p}\right) = -1$, donc d'après le lemme 28 de la page précédente, $\mathbb{F}_p[X]/<X^2+1>$ est intègre, donc d'après le lemme 29, $\mathbb{Z}[\mathrm{i}]/<p>$ l'est aussi, *i.e.* p est premier dans $\mathbb{Z}[\mathrm{i}]$.

(iii) Si $p \equiv 1 \pmod 4$, on sait que $\left(\dfrac{-1}{p}\right) = +1$, donc qu'il existe $a \in \mathbb{Z}$ tel que $p \mid a^2+1 = (a+\mathrm{i})(a-\mathrm{i})$. Comme p ne divise ni $a+\mathrm{i}$ ni $a-\mathrm{i}$, il n'est pas premier (on arriverait à la même conclusion en invoquant les deux lemmes).
Il est donc *réductible* (c'est le point crucial !) : $p = \pi\pi'$ où $\pi, \pi' \in \mathbb{Z}[\mathrm{i}]$ sont non inversibles, *i.e.* $N(\pi), N(\pi') > 1$. D'autre part :

$$p = \pi\pi' \Longrightarrow p^2 = N(p) = N(\pi)N(\pi') \Longrightarrow N(\pi) = N(\pi') = p.$$

Comme ci-dessus on en déduit que π et π' sont irréductibles (si $\pi = uv$, en prenant les normes on trouve $p = N(u)N(v)$ donc $N(u) = 1$ ou $N(v) = 1$). Puisqu'ils ont même module et que leur produit est réel positif, les complexes π et π' sont conjugués.
Comme p n'est pas un carré, π et π' ne sont ni réels ni imaginaires purs, leur quotient n'est donc pas dans $\{+1, -1, +\mathrm{i}, -\mathrm{i}\}$, *i.e.* ils ne sont pas associés. ■

Remarque. On a déterminé dans chaque cas *la* décomposition de p en produit de facteurs premiers, qui est essentiellement unique puisque $\mathbb{Z}[\mathrm{i}]$ est factoriel.

Corollaire 31. (i) Si $p \equiv 1 \pmod 4$, p est somme de deux carrés (d'entiers).
(ii) Si $p \equiv -1 \pmod 4$, et si $p \mid a^2+b^2$, $a, b \in \mathbb{Z}$, alors $p \mid a$ et $p \mid b$.

Démonstration. (i) Avec les notations du théorème, il suffit de prendre $a + b\mathrm{i} := \pi$.
(ii) Si $p \mid a^2+b^2 = (a+b\mathrm{i})(a-b\mathrm{i})$, alors ($p$ étant premier) $p \mid a+b\mathrm{i}$ ou $p \mid a-b\mathrm{i}$. ■

On donnera plus loin un algorithme permettant d'écrire le nombre premier $p \equiv 1 \pmod 4$ sous forme de somme de deux carrés.

Les irréductibles de $\mathbb{Z}[\mathrm{i}]$

Nous sommes maintenant en mesure de les déterminer. Soit $\pi \in \mathbb{Z}[\mathrm{i}]$ irréductible. Il engendre donc un idéal $<\pi>$ qui est premier puisque $\mathbb{Z}[\mathrm{i}]$ est factoriel, et même maximal puisque $\mathbb{Z}[\mathrm{i}]$ est principal. L'idéal $<\pi> \cap \mathbb{Z}$ de $\mathbb{Z}$ est premier. Il n'est pas trivial puisqu'il contient $\pi\overline{\pi} = N(\pi) \in \mathbb{N}^*$. Il existe donc un nombre premier p tel que $<\pi> \cap \mathbb{Z} = p\mathbb{Z}$.

Puisque $p \in <\pi>$, on a $\pi \mid p$ dans $\mathbb{Z}[\mathrm{i}]$ et l'on peut écrire $p = \pi\pi'$ avec $\pi' \in \mathbb{Z}[\mathrm{i}]$. Comme $p^2 = N(p) = N(\pi)N(\pi')$, $N(\pi), N(\pi') \in \mathbb{N}$, $N(\pi) > 1$ (par hypothèse, π étant irréductible n'est pas inversible), on a soit $N(\pi) = N(\pi') = p$ soit $N(\pi) = p^2, N(\pi') = 1$.

- Cas où $N(\pi) = N(\pi') = p$: donc p est décomposé. Ou bien $p = 2$, $\pi, \pi' \sim 1 + \mathrm{i}$ (la décomposition de 2 a déjà été trouvée) ; ou bien $p \equiv 1 \pmod 4$ et $\pi, \pi' = \overline{\pi}$ forment à équivalence près la décomposition trouvée au théorème 30.
- Cas où $N(\pi) = p^2, N(\pi') = 1$: alors π' est inversible, donc $p \sim \pi$ dans $\mathbb{Z}[\mathrm{i}]$; comme p est inerte, on voit que $p \equiv -1 \pmod 4$.

Cette discussion a permis de démontrer :

Théorème 32. Les irréductibles de $\mathbb{Z}[\mathrm{i}]$ sont :

- $\pm 1 \pm \mathrm{i}$ (qui sont équivalents entre eux) ;
- les $p\varepsilon$, $\varepsilon \in \mathbb{Z}[\mathrm{i}]^*$, p étant un nombre premier inerte *i.e.* $p \equiv -1 \pmod 4$;
- les $\pi, \overline{\pi}$ tels que $\pi\overline{\pi}$ est un nombre premier décomposé $p \equiv 1 \pmod 4$ (et π, $\overline{\pi}$ ne sont pas équivalents entre eux).

Comme expliqué dans le Point Méthode de la section 1.2 (page 9), il est commode de faire le choix d'un ensemble de représentants des irréductibles de $\mathbb{Z}[\mathrm{i}]$; nous prendrons $1 + \mathrm{i}$, les nombres premiers inertes $q \equiv -1 \pmod 4$ et, pour chaque nombre premier décomposé $p \equiv +1 \pmod 4$, l'unique paire $\{\pi, \overline{\pi}\}$ de $\mathbb{Z}[\mathrm{i}]$ telle que $p = \pi\overline{\pi}$ et $\pi \in \mathbb{N} + \mathbb{N}\mathrm{i}$.

Exercice 15.

Pour chacun des éléments premiers $\pi \in \mathbb{Z}[\mathrm{i}]$ suivants : $1 + \mathrm{i}$, 3, $2 + \mathrm{i}$, décrire le corps résiduel $\mathbb{Z}[\mathrm{i}]/<\pi>$.

Solution.

On utilisera systématiquement l'isomorphisme : $\mathbb{Z}[\mathrm{i}]/<a+b\mathrm{i}> \simeq \mathbb{Z}[X]/<X^2+1, a+bX>$.
On trouve ainsi :

$$\mathbb{Z}[\mathrm{i}]/<1+\mathrm{i}> \simeq \mathbb{Z}[X]/<X^2+1, 1+X> \simeq \mathbb{Z}/<(-1)^2+1> = \mathbb{F}_2.$$

$$\mathbb{Z}[\mathrm{i}]/<2+\mathrm{i}> \simeq \mathbb{Z}[X]/<X^2+1, 2+X> \simeq \mathbb{Z}/<(-2)^2+1> = \mathbb{F}_5.$$

$$\mathbb{Z}[\mathrm{i}]/<3> \simeq \mathbb{Z}[X]/<X^2+1, 3> \simeq \mathbb{F}_3[X]/<X^2+1> = \mathbb{F}_9.$$

Pour ce dernier cas, on a trouvé un corps qui est une $\mathbb{F}_3$-algèbre engendrée par un élément x tel que $x^2 = -1$. C'est donc un $\mathbb{F}_3$-espace vectoriel de dimension 2, donc un corps à 9 éléments. En fait, voir le livre [MPA1], il existe un unique tel corps (à isomorphisme près), sa notation est $\mathbb{F}_9$.

Pour le cas général, voir l'exercice I.1.39 de la page 71.

4.2 Sommes de deux carrés : théorème de Fermat / Euler

Ce théorème est dû à Fermat, mais c'est Euler qui en donna la première démonstration publiée.

Théorème 33. Soient n un entier naturel non nul, $n = 2^r p_1^{s_1} \cdots p_k^{s_k} q_1^{t_1} \cdots q_l^{t_l}$ sa décomposition en facteurs premiers, où $p_i \equiv 1 \pmod 4$ pour $i = 1, \ldots, k$ et $q_j \equiv -1 \pmod 4$ pour $j = 1, \ldots, l$. Alors, pour que n soit une somme de deux carrés (d'entiers), il faut, et il suffit, que $t_1, \ldots, t_l$ soient pairs.

Démonstration. Montrons que la condition est nécessaire (c'est la partie « élémentaire » de la preuve). Écrivons $n = a^2 + b^2$, $a, b \in \mathbb{Z}$. Si t_j est impair pour un $j = 1, \ldots, l$ alors $q_j \mid a^2 + b^2$, donc, d'après le corollaire 31 de la page 31, q_j divise a et b, donc q_j^2 divise $n = a^2 + b^2$. On écrit $n = q_j^2 n'$, $a = q_j a'$ et $b = q_j b'$: on a donc $n' = a'^2 + b'^2 = 2^r p_1^{s_1} \cdots p_k^{s_k} q_1^{t_1} \cdots q_j^{t_j - 2} \cdots q_l^{t_l}$ et l'on peut recommencer. Comme une telle *descente infinie* est impossible dans $\mathbb{N}$, l'hypothèse de départ est fausse.

Montrons que la condition est suffisante. Puisque $2 = 1^2 + 1^2$ est une somme de deux carrés, que chaque $p_i \equiv 1 \pmod 4$ l'est (d'après le corollaire 31 ; c'est le seul point non élémentaire de la démonstration !) et que $q_1^{t_1} \cdots q_l^{t_l}$ est un carré, il suffit de vérifier que le produit de deux sommes de deux carrés est une somme de deux carrés. C'est immédiat d'après l'une ou l'autre des formules :

$$(a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2 = (ac + bd)^2 + (ad - bc)^2,$$

qui ne font que traduire les égalités $N(u)N(v) = N(uv) = N(u\overline{v})$. ■

Remarque. La méthode de *descente infinie* utilisée dans la première partie de la démonstration a été inventée par Fermat. Elle s'apparente à une récurrence forte. Nous la retrouverons dans la suite de ce module.

Le résultat peut être précisé : pour un entier n vérifiant les conditions du théorème, on peut expliciter *toutes* ses écritures comme somme de deux carrés. Nous reprenons le choix d'un ensemble de représentants des irréductibles de $\mathbb{Z}[\mathrm{i}]$ décrit à la suite du théorème 32 de la page ci-contre. La décomposition de $n = 2^r p_1^{s_1} \cdots p_k^{s_k} q_1^{t_1} \cdots q_l^{t_l}$ en facteurs premiers dans $\mathbb{Z}[\mathrm{i}]$ s'écrit alors :

$$n = (-\mathrm{i})^r (1 + \mathrm{i})^{2r} \pi_1^{s_1} \overline{\pi_1}^{s_1} \cdots \pi_k^{s_k} \overline{\pi_k}^{s_k} q_1^{t_1} \cdots q_l^{t_l}.$$

Écrire $n = a^2 + b^2$ revient à écrire $n = N(u)$, où u est équivalent à $a + b\mathrm{i}$ ou à $a - b\mathrm{i}$. Il y a donc en général, c'est-à-dire si n n'est pas un carré, huit choix possibles de u pour chaque écriture de n en somme de deux carrés ; on considère ainsi tous les couples $(\pm a, \pm b)$ et $(\pm b, \pm a)$ comme équivalents. Comme diviseur de n, l'entier de Gauß u devra avoir la forme :

$$u = \varepsilon (1 + \mathrm{i})^{r'} \pi_1^{s'_1} \overline{\pi_1}^{s''_1} \cdots \pi_k^{s'_k} \overline{\pi_k}^{s''_k} q_1^{t'_1} \cdots q_l^{t'_l}, \quad \varepsilon \in \mathbb{Z}[\mathrm{i}]^*.$$

Pour que l'on ait $p = N(u)$, il faut, et il suffit, que l'on ait :

$$r' = r, \qquad s'_i + s''_i = s_i, \quad i = 1, \dots, k \qquad \text{et} \qquad 2t'_j = t_j, \quad j = 1, \dots, l.$$

Il y a donc essentiellement $\prod_{i=1}^{k}(1 + s_i)$ choix possibles. Cependant, ces choix ne donnent pas tous lieu à des écritures vraiment différentes (par exemple u et $\overline{u}$ apparaissent tous deux) et le décompte exact est un peu plus compliqué ; voir les exercices I.1.41 et I.1.42.

Exemple. Soit $n = 325 = 5^2.13 = (2+\mathrm{i})^2(2-\mathrm{i})^2(3+2\mathrm{i})(3-2\mathrm{i})$. Les valeurs possibles de u (à inversible près) sont :

$$(2+\mathrm{i})^2(2-\mathrm{i})^0(3+2\mathrm{i}) = 1+18\mathrm{i}, \qquad (2+\mathrm{i})^1(2-\mathrm{i})^1(3+2\mathrm{i}) = 15+10\mathrm{i},$$

$$(2+\mathrm{i})^0(2-\mathrm{i})^2(3+2\mathrm{i}) = 17-6\mathrm{i}, \qquad (2+\mathrm{i})^2(2-\mathrm{i})^0(3-2\mathrm{i}) = 17+6\mathrm{i},$$

$$(2+\mathrm{i})^1(2-\mathrm{i})^1(3-2\mathrm{i}) = 15-10\mathrm{i}, \qquad (2+\mathrm{i})^0(2-\mathrm{i})^2(3-2\mathrm{i}) = 1-18\mathrm{i},$$

ce qui donne les représentations :

$$325 = 1^2 + 18^2 = 15^2 + 10^2 = 17^2 + 6^2.$$

Remarque. Si le nombre $r_2(n)$ des couples $(a, b) \in \mathbb{Z}^2$ tels que $a^2 + b^2 = n$ est une fonction arithmétique irrégulière, son *ordre moyen* $\dfrac{r_2(1) + \cdots + r_2(n)}{n}$ se comporte de manière plus régulière. En effet, la somme qui figure au numérateur est le nombre de points entiers $(a, b) \in \mathbb{Z}^2$ du disque fermé de centre $(0, 0)$ et de rayon $\sqrt{n}$. En mettant chaque tel point au centre d'un carré d'aire unité $[a - 1/2, a + 1/2] \times [b - 1/2, b + 1/2]$, on pave une surface qui est très proche du disque, d'où l'on peut déduire (rigoureusement, au prix d'un petit effort, voir le livre [MPA1]) que cette somme est équivalente à πn, donc que $\dfrac{r_2(1) + \cdots + r_2(n)}{n} \to \pi$.

L'algorithme de Cornacchia

De manière générale, il permet de résoudre l'équation $n = x^2 + dy^2$ sous certaines conditions (voir par exemple le livre de Bernadette Perrin-Riou *Algèbre, arithmétique et Maple*, éditions Cassini). Dans le cas simple de l'équation $x^2 + y^2 = p$, il s'identifie pratiquement à la démonstration par descente infinie de l'existence de x et y. On part de $a \in \mathbb{Z}$ tel que $a^2 \equiv -1 \pmod{p}$ (*cf.* le corollaire 22 de la page 23 et l'exercice 11 qui le suit). Donc $a^2 + 1 = mp$, $m \in \mathbb{N}$. Il existe un unique $a' \equiv \pm a$ tel que $1 \leqslant a' \leqslant (p-1)/2$, et l'on peut aussi bien supposer que c'est a ; donc que $mp \leqslant (p-1)^2/4 + 1 \leqslant p(p-1)/4$, donc que $1 \leqslant m \leqslant (p-1)/4$. Si $m = 1$, l'algorithme est terminé. Sinon, on pose $x_0 := a$, $y_0 := 1$ et $m_0 := m$.

À la $i^{ème}$ étape de l'algorithme, on a $x_i^2 + y_i^2 = m_i p$ avec $x_i, y_i \in \mathbb{N}$ et $m_i \in \mathbb{N}^*$. On prend $x'_i \equiv x_i \pmod{m_i}$ et $y'_i \equiv y_i \pmod{m_i}$ tels que $|x'_i|, |y'_i| \leqslant m_i/2$ (donc x'_i, y'_i

peuvent être négatifs). On a donc : ${x_i'}^2 + {y_i'}^2 \equiv x_i^2 + y_i^2 = m_i p \pmod{m_i}$, donc ${x_i'}^2 + {y_i'}^2 = m_i m_i'$ et $m_i' \in \mathbb{N}$ est tel que $m_i' \leqslant m_i/2$. Alors $(x_i^2+y_i^2)({x_i'}^2+{y_i'}^2) = m_i^2 m_i' p$. Mais $(x_i^2+y_i^2)({x_i'}^2+{y_i'}^2) = X_i^2+Y_i^2$, où $X_i = x_i x_i' + y_i y_i' \equiv x_i^2+y_i^2 = m_i p \pmod{m_i}$ et $Y_i = x_i y_i' - y_i x_i' \equiv x_i y_i - y_i x_i = 0 \pmod{m_i}$, d'où $X_i, Y_i \equiv 0 \pmod{m_i}$.
On pose $x_{i+1} := |X_i|/m_i$ et $y_{i+1} := |Y_i|/m_i$. On a $x_{i+1}^2 + y_{i+1}^2 = m_{i+1} p$, où $m_{i+1} := m_i'$.
Comme $1 \leqslant m_{i+1} \leqslant m_i/2$, on arrive en un nombre fini d'étapes au « cas de sortie » $m_i = 1$. On peut même majorer *a priori* le nombre d'étapes : puisque $m_0 \leqslant (p-1)/4$ et que $m_{i+1} \leqslant m_i/2$, l'algorithme arrive à $m_i = 1$ avec $i \leqslant \log_2 \frac{p-1}{4}$: le coût est donc *linéaire* en la taille de p (qui est de l'ordre de $\log_2 p$).

Exemple. Prenons $p = 29 = 4k+1$ avec $k = 7$. Avec Maple ou Caml on trouve que $(2k)! = 14! \equiv 12 \pmod{29}$ et l'on prend donc $a := 12$, d'où $a^2+1 = 145 = 5\times 29$, *i.e.* $m = 5$. On sait à l'avance qu'il y aura arrêt avec $i \leqslant \log_2 7 < 3$, *i.e.* $i \leqslant 2$.
L'algorithme démarre avec $x_0 = 12$, $y_0 = 1$ et $m_0 = 5$. On a donc $x_0' = 2$, $y_0' = 1$ (restes modulo m_0) et $m_0' = 1$. On calcule $X_0 = x_0 x_0' + y_0 y_0' = 25$ et $Y_0 = x_0 y_0' - x_0' y_0 = 10$, puis $x_1 := X_0/5 = 5$, $y_1 := Y_0/5 = 2$ et $m_1 := m_0' = 1$.
L'algorithme se termine victorieusement avec $5^2 + 2^2 = 29$.

Exercice 16.
Tenter $p = 53$ (aide : $26! \equiv 23 \pmod{53}$).
Solution. Grâce à l'aide, on prend $a := 23$, d'où $m := (23^2+1)/53 = 10$. On part donc de $x_0 := 23$, $y_0 := 1$ et $m_0 := 10$. L'arrêt aura lieu avec $i \leqslant \log_2 13 < 4$, *i.e.* $i \leqslant 3$.
On a $x_0' = 3$, $y_0' = 1$, $m_0' = 1$; puis $X = 70$, $Y = 20$, $x_1 = 7$, $y_1 = 2$ et $m_1 = 1$.

4.3 Sommes de trois et de quatre carrés

Signalons sans démonstration le théorème suivant, dû à Legendre et à Gauß :

Théorème 34. Un entier naturel $n \in \mathbb{N}$ est somme de trois carrés si, et seulement s'il n'est pas de la forme $4^a(8b-1)$, $a, b \in \mathbb{N}$.

Une démonstration complète en est donnée dans le livre de Serre [Serre, Appendice du chapitre 4].

Théorème 35 (Lagrange). Tout entier naturel est somme de quatre carrés.

Démonstration. La formule suivante a été établie dans les exercices à la fin du module sur les nombres complexes du livre [L1] :

$$\begin{aligned}(a_1^2+b_1^2+c_1^2+d_1^2)(a_2^2+b_2^2+c_2^2+d_2^2) \\ = \quad (a_1a_2-b_1b_2-c_1c_2-d_1d_2)^2 + (a_1b_2+b_1a_2+c_1d_2-d_1c_2)^2 \\ + (a_1c_2+c_1a_2+d_1b_2-b_1d_2)^2 + (a_1d_2+d_1a_2+b_1c_2-c_1b_2)^2.\end{aligned}$$

Elle est liée à la théorie des quaternions (voir les exercices à la fin du module « Le corps des nombres complexes » du livre [L1], ainsi que le chapitre sur les groupes classiques du livre [L2]) mais peut bien entendu être vérifiée directement ! Elle entraîne que le produit de deux sommes de quatre carrés est une somme de quatre carrés. Il suffit donc de démontrer que tout nombre premier est somme de quatre carrés. Une démonstration basée sur la théorie des *anneaux euclidiens non commutatifs* est donnée dans [Hindry, chap. 3] ; au même chapitre de ce livre, on trouve une preuve basée sur la *géométrie des nombres*. Nous en donnons ici une preuve dans l'esprit de la méthode de descente infinie, donc dans l'esprit de Fermat, Euler et Lagrange. Comme elle ressemble à celle qui sous-tend l'algorithme de Cornacchia, nous laissons au lecteur le soin de contrôler une partie des détails.

On fixe un nombre premier p que l'on peut évidemment supposer impair. Du corollaire 19 de la page 21, on déduit l'existence de $x, y \in \mathbb{Z}$ tels que $x^2 + y^2 \equiv -1 \pmod{p}$. On écrit donc $x^2 + y^2 + 1 = \ell p$, avec $\ell \in \mathbb{N}^*$. On peut supposer de plus que $0 \leqslant x, y \leqslant (p-1)/2$, ce qui implique $1 \leqslant \ell \leqslant (p-1)/2$.

Supposons plus généralement donnés $x, y, z, t, \ell \in \mathbb{N}$ tels que $x^2 + y^2 + z^2 + t^2 = \ell p$, avec $\ell \in \mathbb{N}^*$ aussi petit que possible. On sait que $\ell < p$. Si $\ell = 1$, la preuve est terminée.

Si ℓ était pair, les entiers x, y, z, t pourraient être regroupés par paires de même parité ; par exemple $x \equiv y, z \equiv t \pmod{2}$. On aurait alors l'égalité :

$$\left(\frac{x+y}{2}\right)^2 + \left(\frac{x-y}{2}\right)^2 + \left(\frac{z+t}{2}\right)^2 + \left(\frac{z-t}{2}\right)^2 = (\ell/2)p,$$

contredisant la minimalité de ℓ.

On peut donc supposer que ℓ est impair. Supposons que $\ell \geqslant 3$. On introduit x', y', z', t' respectivement congrus à x, y, z, t modulo ℓ et tous de valeur absolue strictement inférieure à $\ell/2$. On en déduit que $x'^2 + y'^2 + z'^2 + t'^2 = \ell\ell'$, avec $1 \leqslant \ell' < \ell$. On applique l'identité donnée plus haut, en remplaçant a_1, b_1, c_1, d_1 par x, y, z, t et a_2, b_2, c_2, d_2 par $x', -y', -z', -t'$:

$$X = xx' + yy' + zz' + tt', \qquad Y = -xy' + yx' - zt' + tz',$$
$$Z = -xz' + zx' - ty' + yt', \qquad T = -xt' + tx' - yz' + zy',$$

pour écrire :

$$(x^2 + y^2 + z^2 + t^2)(x'^2 + y'^2 + z'^2 + t'^2) = X^2 + Y^2 + Z^2 + T^2 = \ell'\ell^2 p, \text{ où}$$

et l'on vérifie grâce aux congruences que $X, Y, Z, T \equiv 0 \pmod{\ell}$. Posant donc :

$$x_1 := \frac{X}{\ell},\ y_1 := \frac{Y}{\ell},\ z_1 := \frac{Z}{\ell},\ t_1 := \frac{T}{\ell},$$

on voit que $x_1^2 + y_1^2 + z_1^2 + t_1^2 = \ell' p$, contredisant la minimalité de ℓ. ∎

Un peu d'Histoire

Waring a conjecturé en 1770 qu'un théorème analogue valait pour tous les exposants : tout entier naturel est somme de quatre carrés, de neuf cubes, de dix-neuf bicarrés, etc. Hilbert démontra ce théorème en 1909 puis Hardy et Littlewood en donnèrent une démonstration radicalement différente en 1920 ; voir [HW, chap. XX].

5 Nombres premiers, critères de primalité

Dans ses *Disquisitiones arithmeticae*, parus en 1801[4], Gauß écrivait :

> *Le problème où l'on se propose de distinguer les nombres premiers des nombres composés, et de décomposer ceux-ci en leurs facteurs premiers, est connu comme un des plus importants et des plus utiles de toute l'Arithmétique ; tout le monde sait qu'il a été l'objet des recherches des géomètres tant anciens que modernes, et il serait inutile de donner des détails à cet égard. (...) En outre, la dignité de la science semble demander que l'on recherche avec soin tous les secours nécessaires pour parvenir à la solution d'un problème si élégant et si célèbre.*

Un de ces « anciens géomètres », Euclide, a démontré dans ses *Éléments* qu'il y avait une infinité de nombres premiers :

Théorème 36 (Euclide).
L'ensemble $P := \{2, 3, 5, \dots\}$ des nombres premiers est infini.

Démonstration. Soient $p_1, \dots, p_k$ des éléments de P deux à deux distincts. L'entier $N := p_1 \cdots p_k + 1$ est strictement supérieur à 1, donc divisible par (au moins) un nombre premier p, qui ne peut être l'un des p_i car sinon p diviserait $N - p_1 \cdots p_k = 1$. L'inclusion $\{p_1, \dots, p_k\} \subset P$ est donc stricte et $\operatorname{card} P > k$ quel que soit k. ■

Nous noterons $(p_k)_{k \geqslant 1}$ la suite strictement croissante de tous les nombres premiers ; donc $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, etc. Soit par ailleurs pour tout entier naturel non nul $n \in \mathbb{N}^*$:

$$\pi(n) := \operatorname{card}\,(P \cap [\![1, n]\!])$$

le nombre des nombres premiers $p \leqslant n$. On a donc, pour $k, n \in \mathbb{N}^*$:

$$p_k \leqslant n \iff k \leqslant \pi(n).$$

Le théorème d'Euclide peut alors se traduire comme suit :

Corollaire 37.

$$\lim_{k \to +\infty} p_k = +\infty \quad \text{et} \quad \lim_{n \to +\infty} \pi(n) = +\infty.$$

Exercice 17.
Soit $(x_k)_{k \geqslant 1}$ la suite définie par $x_1 := 2$ et $x_{k+1} := x_1 \cdots x_k + 1$.
Montrer que $p_k \leqslant x_k$ pour tout k et estimer l'ordre de grandeur de x_k.

[4] Nos citations sont tirées du paragraphe 329, à la page 416 de l'édition Gabay.

Solution. Avec les notations ci-dessus, la démonstration par Euclide du théorème permet de conclure que $p_{k+1} \leqslant N = p_1 \cdots p_k + 1$ (puisque tout facteur premier de N est distinct de $p_1, \ldots, p_k$ donc au moins égal à p_{k+1}). La majoration $p_k \leqslant x_k$ découle alors d'une récurrence évidente.
Des relations $x_{k+1} := x_1 \cdots x_k + 1$ et $x_k := x_1 \cdots x_{k-1} + 1$ on déduit que $x_{k+1} = x_k^2 - x_k + 1$ pour tout $k \geqslant 1$. Comme visiblement $x_k \to +\infty$, on en déduit que $\ln\, x_{k+1} = 2 \ln\, x_k + \mathrm{o}(1)$, d'où la convergence de la suite des $2^{-k} \ln\, x_k$ (considérer la série des différences successives), d'où l'existence d'une constante $C > 0$ telle que $\ln\, x_k \sim C.2^k$: la *taille* (nombre de chiffres dans une base arbitraire) de x_k double à chaque étape, ce majorant est très grand.
En fait, de la relation $x_{k+1} = x_k^2 - x_k + 1$, on peut déduire que $x_{k+1} - 1/2 > (x_k - 1/2)^2$, d'où $x_k > (3/2)^{2^{k-1}}$ pour tout $k \geqslant 1$.

L'étude des nombres premiers a largement tourné autour de deux thèmes : comment en produire de grands (nous rencontrerons plus loin les nombres premiers « de Fermat » et « de Mersenne »); et décrire leur répartition. Euler, Legendre, Gauß, Riemann[5] ont pressenti le *théorème des nombres premiers* :

$$\pi(n) \sim \frac{n}{\ln\, n} \text{ lorsque } n \to +\infty.$$

Ce théorème fut démontré en 1896 indépendamment par Jacques Hadamard et par Charles-Jean de La Vallée Poussin. On pourra trouver une démonstration de ce théorème difficile dans [Baker] et dans [Hindry]; et dans le livre [MPA1] une estimation plus faible — mais instructive — due à Tchebychef :

$$\text{il existe } A, B > 0 \text{ tels que } \forall n \geqslant 2\ ,\ A\frac{n}{\ln\, n} \leqslant \pi(n) \leqslant B\frac{n}{\ln\, n}\cdot$$

La preuve de la majoration est proposée en exercice à la fin de ce chapitre; mais nous aurons besoin un peu plus loin d'une estimation qui s'en rapproche :

Théorème 38 (Legendre, 1808).
Lorsque n tend vers $+\infty$, $\pi(n)$ est négligeable devant n :

$$\pi(n) = \mathrm{o}(n), \quad n \to +\infty.$$

Démonstration. Pour tout entier n, nous noterons Π_n le produit des nombres premiers $p \leqslant n$. Soient $n \geqslant 2$ fixé et $k := \pi(n)$, de sorte que $p_k \leqslant n < p_{k+1}$ et que $\Pi_n = p_1 \cdots p_k$. L'entier $\dfrac{\Pi_{2n}}{\Pi_n}$ est le produit des nombres premiers p tels que $n < p \leqslant 2n$. C'est donc un diviseur de $(n+1) \cdots 2n = \dfrac{(2n)!}{n!} = n!\binom{2n}{n}$ et c'est par ailleurs un nombre premier avec $n!$

[5] Dans son célèbre article de 1859, « Ueber die Anzahl des Primzahlen unter einer gegebenen Grösse », soit « Sur le nombre de nombres premiers inférieurs à une grandeur donnée », Riemann a relié l'étude du comportement asymptotique de $\pi(n)$ à celui des zéros de la fonction Zeta. C'est dans cette article qu'il formula la conjecture connue sous le nom d'« Hypothèse de Riemann », qui est l'objet de l'un des sept « Prix du millénaire » proposés par la Fondation Clay.

puisque tous ses facteurs premiers sont strictement plus grands que n. D'après le lemme de Gauß, $\dfrac{\Pi_{2n}}{\Pi_n}$ est donc un diviseur de $\binom{2n}{n}$. En particulier :

$$\frac{\Pi_{2n}}{\Pi_n} \leqslant \binom{2n}{n} \leqslant 2^{2n},$$

cette dernière majoration venant de ce que $\binom{2n}{n}$ est l'un des termes du développement binomial de $(1+1)^{2n}$.
Si l'on trouve $C > 0$ tel que $\Pi_n \leqslant C.4^n$, en multipliant par l'inégalité $\Pi_{2n}/\Pi_n \leqslant 4^n$ on en déduira que $\Pi_{2n} \leqslant C.4^{2n}$. Or $\Pi_1 = p_1 = 2$, et la constante $C = 1/2$ convient pour $n = 1$. Il en découle par récurrence que $\Pi_n \leqslant (1/2).4^n$ pour tout entier n qui est une puissance de 2.
Soit maintenant n quelconque et soit n' une puissance de 2 telle que $n'/2 < n \leqslant n'$ (donc la plus petite puissance de 2 majorant n). Il est clair que $\Pi_n \leqslant \Pi_{n'}$, d'où :

$$\Pi_n \leqslant \Pi_{n'} \leqslant (1/2).4^{n'} < (1/2).4^{2n} = (1/2).16^n.$$

Mais par ailleurs, la suite strictement croissante des nombres premiers est strictement minorée par celle des entiers : $p_i > i$ pour tout i ; donc $\Pi_n = p_1 \cdots p_k > k!$, d'où les majorations :

$$k! < (1/2).16^n \Longrightarrow \ln\,(\pi(n))! < n \ln\, 16.$$

Puisque $\pi(n)$ tend vers l'infini quand n tend vers l'infini, $\ln\,(\pi(n))! \sim \pi(n) \ln\, \pi(n)$ et l'on conclut que $\pi(n)$ est négligeable devant $\ln\,(\pi(n))!$ donc devant n. ■

Il y a donc relativement peu de nombres premiers, même s'ils sont en nombre infini. On pourra trouver une étude plus approfondie de la répartition des nombres premiers, par des méthodes élémentaires dans le chapitre d'arithmétique de [MPA1] ; mais surtout, par des moyens plus sophistiqués dans [Baker] et dans [Hindry].
Dans la suite de cette section, nous allons parler de méthodes pour tester la primalité de grands entiers. Le problème de la factorisation de nombres composés est évidemment apparenté, mais il est beaucoup plus difficile et nous ne l'aborderons pas (sinon à l'occasion de remarques historiques). Comme la *faisabilité* des méthodes proposées est un critère crucial, notre présentation revêt parfois un caractère pragmatique et heuristique. Le lecteur appréciera sans doute dans cette ambiance terre-à-terre le nombre d'assertions commençant par « on ignore actuellement … » ; il reste de la place pour le rêve. La section se termine par l'étude des nombres premiers de Fermat et de Mersenne.

5.1 Aspects pratiques

Citons à nouveau Gauß :

> *(…) toutes les méthodes proposées jusqu'à présent sont restreintes à des cas très particuliers, ou sont si longues et si pénibles, que même pour ceux de ces nombres qui ne dépassent pas les limites des Tables dont on est redevable à quelques mathématiciens, c'est-à-dire, pour les nombres à l'égard desquels ces méthodes sont inutiles, elles fatiguent la patience du calculateur le plus exercé, et qu'elles ne sont pour ainsi dire pas applicables à de plus grands*

nombres. (...) Aussi nous ne doutons pas que les deux méthodes suivantes[6]*, dont nous pouvons affirmer la brièveté et l'efficacité d'après une longue expérience, ne plaisent aux amateurs de l'Arithmétique. Au reste, il est dans la nature du problème, que les méthodes,* quelles qu'elles soient*, deviennent d'autant plus longues, que les nombres auxquels on les applique sont plus considérables ; cependant, pour les méthodes suivantes, les difficultés ne s'accroissent qu'avec beaucoup de lenteur, et les nombres de sept, de huit, et même d'un plus grand nombre de chiffres, ont toujours été traités, surtout par la seconde, avec un succès très-heureux, et avec toute la célérité que l'on peut attendre pour de si grands nombres, qui, suivant les méthodes connues jusqu'à présent, exigeraient un travail intolérable, même pour le calculateur le plus infatigable.*

5.1.1 Quelques étapes historiques

En 1640, Fermat se demande si tous les nombres $F_r := 2^{2^r} + 1$ sont premiers. Dans une lettre à Mersenne, il constate que c'est vrai pour $r = 1, 2, 3, 4$ mais s'interroge sur le cas $r = 5$. Pourtant, nous verrons que $F_5 = 2^{32} + 1 = 4294967297$ ne l'est pas, et cela par une méthode que Fermat aurait pu appliquer sans peine. À la même époque, Mersenne se demande si tous les nombres $M_p := 2^p - 1$, avec p premier, sont premiers ; c'est bien le cas pour $p = 2, 3, 5, 7$, mais $M_{11} = 2047 = 23 \times 89$ est composé. Euler prouvera en 1772 que, par exemple, $M_{31} = 2^{31} - 1 = 2147483647$ est bien premier. Cependant, en 1878, E. Lucas prouve que M_{67} est composé, mais *sans pouvoir en donner des facteurs*. Ce n'est qu'en 1903 que Cole obtient la décomposition :

$$M_{67} = 2^{67} - 1 = 193707721 \times 761838257287.$$

Quelle qu'ait été la difficulté de *trouver* une telle décomposition, la *certification* ci-dessus est rapide, puisqu'elle comporte seulement le calcul de M_{67} et la multiplication au membre droit. Nous reviendrons sur cette notion de certification (théorème 44 de la page 47). Enfin Lucas put prouver en 1876 que M_{127} (un nombre d'une quarantaine de chiffres décimaux) est premier. C'est le record absolu et définitif pour des calculs effectués à la main[7].

Les nombres de Fermat et les nombres de Mersenne ne sont donc pas tous premiers. Mais ils se prêtent à des tests particuliers qui ont permis d'obtenir avec peu de moyens de spectaculaires résultats. Outre les exemples ci-dessus, mentionnons la preuve par Lucas à l'aide du test de Pépin (voir le théorème 46, page 51) que :

$$F_6 = 2^{64} + 1 = 18446744073709551617$$

est composé ; mais ce n'est qu'en 1882 que Landry (à l'âge de quatre-vingts ans !) donna sans explication[8] la décomposition : $F_6 = 274177 \times 67280421310721$.

[6]En réalité, toute la suite de cette citation concerne des méthodes de *factorisation* ce qui est un problème plus difficile que tester la primalité. Cependant ce qu'en dit Gauß s'applique parfaitement à notre sujet.

[7]Le record absolu et définitif pour l'ère pré-électronique fut établi par Ferrier, qui prouva en 1951 avec une calculatrice de bureau que $\dfrac{2^{148}+1}{17}$ est premier.

[8]L'arithméticien brésilien Paulo Ribenboim a depuis proposé une reconstruction possible du chemin suivi par

Landry avait d'ailleurs déjà donné en 1869 (donc à près de soixante-dix ans) la décomposition :

$$M_{59} = 179951 \times 3203431780337.$$

5.1.2 L'ère électronique

L'avènement des ordinateurs a évidemment permis de tester de beaucoup plus grands nombres[9]. En 1980, le test de Lucas permettait de prouver la primalité de M_{44497}, un nombre de Mersenne ayant 13395 chiffres : c'était alors le record. En 1985, on a testé la primalité de M_{216091} (plus de 55000 chiffres), en 1992 celle de M_{756839} (plus de 227000 chiffres) et en 1992 celle de M_{859433} (plus de 258000 chiffres). Nous reviendrons sur ces records au 5.5.2.

Mais l'événement le plus marquant de l'ère informatique, c'est qu'en 1977, avec l'invention de la méthode de cryptage RSA, les nombres premiers et la factorisation sont devenus *utiles*. Comme l'a dit Ribenboim : « *Poor number theory, the Queen, relegated (or raised ?) to be the object of a courtship inspired by necessity, not by awe* ». Nous avons déjà évoqué page 20 la façon dont Ron Rivest, Adi Shamir et Leonard Adleman (les inventeurs de RSA) reliaient la fiabilité de RSA à la difficulté de la factorisation, et comment ils évaluaient cette dernière. Les chiffres donnés devraient sensibiliser le lecteur au caractère critique des discussions qui vont suivre.

5.1.3 Quelques estimations de coûts

Rappelons quelques résultats de complexité algorithmique élémentaire (voir le module correspondant du livre [L1], que l'on peut compléter par [Demazure]). Nous noterons $|n|_b$ la taille d'un entier écrit en base b, voire $|n|$ si le choix de la base n'a pas d'importance : en effet, toutes ces mesures de taille sont essentiellement proportionnelles à $\log n$, donc équivalentes en ce qui concerne nos estimations, ce qui se note :

$$|n| = \Theta(\log n),$$

où, comme d'habitude, $u_n = \Theta(v_n)$ signifie que $u_n = \mathrm{O}(v_n)$ et $v_n = \mathrm{O}(u_n)$. Gauß prenait pour taille le nombre de chiffres décimaux ($b = 10$), un informaticien pourrait prendre le nombre de bits ($b = 2$) ou d'octets ($b = 2^8$). Notons que, dans tous les cas, la *valeur* de n est *exponentielle* en la *taille* de n : si n s'écrit avec 100 chiffres en base 10, il faut de l'ordre de 10^{100} étapes pour énumérer les entiers de 1 à n.

Voici trois estimations fondamentales en arithmétique algorithmique :

1. Le coût du calcul du pgcd $m \wedge n$, $m < n$, est $\mathrm{O}\left(|m|^3\right)$: voir [Demazure] pour une preuve.
2. Le coût du calcul de $a^m \pmod n$ (exponentiation dichotomique et calculs dans $\mathbb{Z}/n\mathbb{Z}$) est $\mathrm{O}\left(|m|\,|n|^2\right)$. En effet, selon le module d'algorithmique du livre [L1], une multiplication modulo n a pour coût $\mathrm{O}\left(|n|^2\right)$ et l'exponentiation dichotomique avec exposant m comporte $\mathrm{O}\left(|m|\right)$ telles multiplications.

Landry ; voir son livre « The Little Book of Bigger Primes »(Springer, 2004) ou son texte plus ancien « Les records des nombres premiers » sur `http://www.numdam.org/item/?id=SPHM_1987___8_A1_0`

[9] Mais l'amélioration des algorithmes y est également pour beaucoup, comme il ressort de la discussion sur la fiabilité de RSA page 20.

3. Le coût du calcul du symbole de Jacobi $\left(\frac{m}{n}\right)$, $1 \leqslant m \leqslant n$, est $\mathrm{O}\left(|m|^3\right)$ par une méthode proche de l'algorithme d'Euclide, voir encore [Demazure], ainsi que l'exercice I.1.53 de la page 72. Une méthode sans division due à Lehmer permet même un coût en $\mathrm{O}\left(|m|^2\right)$.

5.2 Mauvaises méthodes

Soit donc un « grand » entier naturel n, dont on veut tester la primalité. Nous allons estimer le coût de divers algorithmes en fonction de la taille de la donnée n. Nous conviendrons ici que $|n| = |n|_{10}$.

5.2.1 Méthode de l'école primaire

Si $n = ab$ est une décomposition non triviale avec $a \leqslant b$, on a bien entendu $a \leqslant \lfloor\sqrt{n}\rfloor$. L'algorithme le plus primitif est donc le suivant : pour d de 2 à $\lfloor\sqrt{n}\rfloor$, effectuer la division euclidienne de n par d ; si l'on trouve une fois un reste nul, n est composé (et l'on en a trouvé un facteur) ; sinon, il est premier.

Dans le pire des cas, on effectue environ $\sqrt{n} \approx 10^{|n|/2}$ divisions euclidiennes : le coût est donc exponentiel en $|n|$ (pour un nombre de 10 chiffres : cent mille divisions ; pour un nombre de 100 chiffres : 10^{50} divisions), et cela sans même prendre en compte le coût variable des divisions selon la taille des nombres en jeu. Ce test est pratiquement infaisable pour de grands nombres ; son seul avantage est que, s'il déclare qu'un nombre n'est pas premier, il en fournit une décomposition non triviale. C'est donc l'amorce d'un algorithme de factorisation (d'ailleurs tout aussi mauvais).

Ce test admet une variante « probabiliste » : essayer au hasard des nombres d entre 2 et la partie entière $\lfloor\sqrt{n}\rfloor$ en estimant qu'un grand nombre d'échecs (divisions à reste non nul) est un signe probable de primalité. Mais cette méthode n'est pas raisonnable. Par exemple pour un nombre de la forme $n = pq$, où p et q sont de grands nombres premiers distincts, la seule valeur de succès serait $d = \min(p, q)$ et la probabilité de tomber dessus est très faible, de l'ordre de $1/\sqrt{n}$.

5.2.2 Le crible d'Ératosthène

Cette méthode date du troisième siècle avant notre ère ! Elle fournit la liste de tous les nombres premiers jusqu'à n. On écrit d'abord la liste de tous les entiers de 2 à n. Puis, successivement pour chaque d de 2 à $\lfloor\sqrt{n}\rfloor$:

- Si d est barré, on passe à l'entier suivant.
- Si d n'est pas barré, on barre tous ses multiples propres dans la liste : $2d$, $3d$, …, $\lfloor(n/d)\rfloor d$; puis on passe à l'entier suivant.

À la fin, les nombres non barrés sont les nombres premiers $p \leqslant n$. La raison pour s'arrêter avant $\lfloor\sqrt{n}\rfloor$ est que, si l'entier $d \leqslant n$ est composé, son plus petit facteur premier est au plus égal à $\sqrt{n}$. Voici une implémentation possible.

```
Crible d'Ératosthène
Initialiser à vrai un tableau de booléens T[2..n];
d <- 2;
Tant que d*d <= n
    (si T[d] alors pour k de 2 à n/d faire T[k*d] <- faux;
    d <- d+1);;
```

Le coût total de cet algorithme est au moins égal au nombre d'entiers barrés, donc à $n - \pi(n)$, qui est équivalent à n d'après le théorème 38. C'est donc encore un coût exponentiel en la taille de n. Cependant, l'avantage de cet algorithme est qu'il produit la suite des nombres premiers (ou tout au moins un fragment).

5.2.3 Test basé sur le théorème de Wilson

Rappelons que, selon le théorème de Wilson (module II.1 du livre [L1]), l'entier naturel n est premier si, et seulement si, $(n-1)! \equiv -1 \pmod{n}$. Le test correspondant consiste donc à calculer $(n-1)!$ modulo n. Même si l'on effectue les calculs dans $\mathbb{Z}/n\mathbb{Z}$, donc à taille de multiplicandes bornée, il faut effectuer $(n-2)$ multiplications. C'est donc encore un coût exponentiel en la taille de n. De plus, dans le cas où le test conclut à la non-primalité, il ne fournit aucun facteur.

5.2.4 Tests basés sur le petit théorème de Fermat

L'une des formes du petit théorème de Fermat dit que, si n est premier, pour tout $a \in \mathbb{Z}$ on a $a^n \equiv a \pmod{n}$. Comme cette relation ne dépend en réalité que de la classe de a modulo n, et comme d'autre part elle est trivialement vérifiée pour $a = 0, 1$, on est conduit à tester si elle est vérifiée pour les $(n-2)$ valeurs $a = 2, \ldots, n-1$. Pour chacune de ces valeurs, le coût du calcul de $a^n \pmod{n}$ est polynomial en la taille de n (paragraphe 5.1.3). Mais il faut effectuer $(n-2)$ tels tests, d'où encore un coût exponentiel en la taille de n. De plus, comme pour le cas précédent, si le test conclut à la non-primalité, il ne fournit aucun facteur.

En fait, la situation est même pire que cela, car le petit théorème de Fermat n'énonce qu'une condition nécessaire : si n est un *nombre de Carmichael* (proposition ci-dessous), le test ne permet pas de trancher. C'est seulement si l'une des congruences est infirmée que l'on peut conclure avec certitude à la non-primalité. Or, il a été prouvé en 1992 par Alford, Granville et Pomerance qu'il y a une infinité de nombres de Carmichael.

Proposition 39. Soit $n > 1$ non premier. Les propriétés suivantes (qui caractérisent les *nombres de Carmichael*) sont équivalentes :
(i) $\forall a \in \mathbb{Z}\ ,\ a^n \equiv a \pmod{n}$.
(ii) $\forall a \in \mathbb{Z}\ ,\ a \wedge n = 1 \Rightarrow a^{n-1} \equiv 1 \pmod{n}$.
(iii) $n = p_1 \cdots p_k$, où les p_i sont premiers distincts, où chaque $p_i - 1$ divise $n - 1$ et où $k \geqslant 3$.

Démonstration. L'implication (i) $\Rightarrow$ (ii) est évidente puisque a est inversible modulo n. Supposons (ii) vérifié. En l'appliquant à $a := -1$, on voit que n est impair. Si p^r est un facteur primaire de n, on sait que $(\mathbb{Z}/p^r\mathbb{Z})^*$ est cyclique d'ordre $p^{r-1}(p-1)$ (théorème 17

de la page 16) et que c'est un facteur de $(\mathbb{Z}/n\mathbb{Z})^*$ (lemme chinois), donc $p^{r-1}(p-1)$ divise $n-1$. Puisque p divise n, cela entraîne que $r=1$ (donc n est quadratfrei); et que $(p-1)$ divise $n-1$, ce qui achève de prouver (iii). (On rapprochera cet argument de l'exercice I.1.29 de la page 70).
Supposons (iii) vérifié. On va montrer que chaque p_i divise a^n-a, ce qui donnera la conclusion voulue puisque n est quadratfrei. Si p_i divise a, c'est clair. Sinon, par le petit théorème de Fermat, p_i divise $a^{p_i-1}-1$, donc $a^{n-1}-1$ puisque p_i-1 divise $n-1$, donc bien entendu a^n-a. On ne peut avoir $k=1$ puisque n n'est pas premier. Si $n=pq$, la condition $p-1 \mid pq-1$ jointe au fait évident que $p-1 \mid pq-q$, entraîne $p-1 \mid q-1$; symétriquement, $q-1 \mid p-1$, donc $p=q$, contradiction. On a donc $k \geqslant 3$. ■

Exemples. Le plus petit nombre de Carmichael est $561 = 3\times 11\times 17$. Citons encore 1105 (qui est une somme de deux carrés); 1729 (qui est de deux manières une somme de deux cubes); 41041 (qui a quatre facteurs premiers).

Une variante de la méthode précédente consiste à tester si, pour tout a tel que a et n sont premiers entre eux, $a^{n-1} \equiv 1 \pmod{n}$. L'algorithme se déroule ainsi : pour $a = 2,\ldots,n-1$, on calcule le pgcd de a et n (facilement et à faible coût avec l'algorithme d'Euclide). Si ce pgcd n'est pas 1, c'est un facteur non trivial de n qui est donc composé (avec certificat !). Sinon, on calcule $a^{n-1} \pmod{n}$. Cette méthode a les mêmes défauts que la précédente : elle est extrêmement couteuse; et elle ne permet pas de conclure pour les nombres de Carmichael.

Variante un peu améliorée

Cependant, une légère variante *probabiliste* founit un *test de composition* acceptable : on choisit a « au hasard »; s'il n'est pas premier avec n (algorithme d'Euclide), on a un facteur. S'il est premier avec n et que $a^{n-1} \not\equiv 1 \pmod{n}$, on peut à coup sûr déclarer n composé. Un tel a est appelé *témoin de Fermat* et l'acceptabilité du test vient de ce que l'on peut démontrer qu'un entier composé a « généralement beaucoup » de témoins de Fermat : « généralement » parce que selon un théorème de Knödel (1953), la fréquence des nombres de Carmichael tend vers 0 à l'infini; et « beaucoup » en vertu du résultat suivant.

Proposition 40. Si n est composé et n'est pas un nombre de Carmichael, au moins la moitié des éléments de $(\mathbb{Z}/n\mathbb{Z})^*$ sont des témoins de Fermat de composition de n.

Démonstration. Si n n'est pas un nombre de Carmichael, le sous-groupe de $(\mathbb{Z}/n\mathbb{Z})^*$ formé des x tels que $x^{n-1}=1$ est au moins d'indice 2. ■

Si l'on tire au hasard un grand nombre de fois des éléments de $(\mathbb{Z}/n\mathbb{Z})^*$, et si n est composé et pas de Carmichael, on peut rendre la probabilité de trouver un témoin de Fermat de composition très proche de 1. Mais cet argument montrera toute son efficacité dans le test de Solovay-Strassen.

Exemple. Pour savoir si $F_5 := 2^{32}+1 = 4294967297$ est premier, on peut calculer le reste de la division euclidienne de $3^{2^{32}}$ par F_5. Après trente-deux multiplications à taille

bornée et une division euclidienne, on trouve le reste 3029026160, ce qui montre que F_5 est composé et que 3 est un témoin de Fermat de composition de F_5. Mais on ne peut en déduire un facteur de F_5.

5.3 Bonnes méthodes probabilistes : Solovay-Strassen

Selon Euler, pour tout nombre premier impair p et pour tout $m \in \mathbb{Z}$, on a la congruence $\left(\frac{m}{p}\right) \equiv m^{(p-1)/2} \pmod{p}$ (théorème 20 de la page 22).

Le théorème 27 de la page 29 dit que réciproquement, si l'entier impair n est tel que $\left(\frac{m}{n}\right) \equiv m^{(n-1)/2} \pmod{n}$ pour tout entier m tel que $m \wedge n = 1$, alors n est premier ; mais il faut ici utiliser le symbole de Jacobi. Bien que le calcul de $\left(\frac{m}{n}\right)$ et de $m^{(n-1)/2} \pmod{n}$ soit aisé (paragraphe 5.1.3 de la page 41), le nombre de valeurs de m nécessaires pour appliquer ce critère est exponentiel en $|n|$.

On tente donc de tester la primalité de n de la manière suivante : on tire au hasard m entre 2 et $n-1$. Si $m \wedge n \neq 1$, on déclare bien entendu n composé. Pour la suite nous supposons que $m \wedge n = 1$. Si $\left(\frac{m}{n}\right) \not\equiv m^{(n-1)/2} \pmod{n}$, on déclare n composé ; sinon on déclare n *peut-être* premier. Ces conclusions sont bien entendu correctes d'après ce qui précède. Cette méthode est-elle raisonnable ? La question est de savoir si, dans les cas où n n'est pas premier, on a une chance raisonnable de trouver un m qui en témoigne. Si n a peu de diviseurs (par exemple $n = pq$, comme dans RSA), on a peu de chances de tomber sur un m tel que $m \wedge n \neq 1$.

Mais la seconde partie du test est plus probante. En effet, les applications $x \mapsto \left(\frac{x}{n}\right)$ et $x \mapsto x^{(n-1)/2}$ étant des morphismes sur le groupe commutatif $(\mathbb{Z}/n\mathbb{Z})^*$, le lieu de leur coïncidence est un sous-groupe G. Si n n'est pas premier, d'après le théorème rappelé ci-dessus, c'est un sous-groupe propre donc d'indice au moins 2. Par conséquent, on a au moins une chance sur deux de trouver un *témoin* de non-primalité : c'est un test très raisonnable !

En pratique, on exécutera cent fois (par exemple) le test. Si une fois la réponse est une inégalité, on aura la certitude que n est composé. Si la réponse est cent fois une égalité, on estimera que n est « très probablement » premier.

Le coût de ce test est aussi raisonnable parce que les deux membres peuvent être calculés rapidement (paragraphe 5.1.3 de la page 41). En revanche, en cas d'inégalité, le test ne fournit pas un facteur de n. Voir [Demazure, Hindry] pour des compléments, en particulier pour une mise en œuvre efficace. On y trouvera aussi une variante bien meilleure, le test de Miller-Rabin.

5.4 Bonnes méthodes déterministes

Les bonnes méthodes déterministes *générales* sont d'un niveau plus avancé que celui choisi pour ce chapitre, aussi n'en donnerons-nous que pour deux classes importantes, les nombres de Fermat et les nombres de Mersenne, à la section suivante. Nous allons ici fixer le cadre théorique.

5.4.1 PRIMES est dans NP

S'il peut être difficile de *trouver* une preuve qu'en entier comme M_{59} est composé (voir page 41), il est facile d'écrire une telle preuve : celle-ci peut en effet se réduire à vérifier la relation $M_{59} = 179951 \times 32034317800337$. Plus généralement, pour certifier que n est composé, il suffit d'effectuer une multiplication $a \times b$ (avec $a, b > 1$) dont le résultat soit n. Le coût d'une telle multiplication (voir le chapitre d'algorithmique du livre [L1]) est pratiquement proportionnel au produit des tailles (nombres de chiffres dans une base quelconque) de a et b, donc c'est un $\mathrm{O}(|n|^2)$. Nous dirons que (a, b) est un *certificat* de composition et que le coût de la certification correspondante est $\mathrm{O}(|n|^2)$. Bien entendu, ce qui est difficile, c'est de trouver un tel certificat, mais ensuite la certification est rapide.

De manière générale, en théorie abstraite de la complexité des algorithmes, on dit qu'un problème pour lequel on peut certifier une solution en temps polynomial en la taille des données est *dans la classe NP*. Ce terme signifie « non deterministic polynomial » car une « machine non déterministe » pourrait essayer un certificat (a, b) au hasard et, avec de la chance, résoudre le problème en temps polynomial. Si l'on connaît un algorithme (déterministe !) capable de trouver un certificat en temps polynomial, on dit que le problème est *dans la classe P*. De l'argument précédent résulte que chercher si un nombre est composé est un problème de la classe NP ; mais on ignore actuellement[10] s'il est dans la classe P.

Vaughan Pratt a démontré en 1975 que le problème de la primalité est dans la classe NP ; autrement dit, si n est premier, il en existe un certificat fournissant une preuve « courte » (*i.e.* de longueur polynomiale en $|n|$) de cette primalité. Le lecteur consciencieux prendra la peine de réfléchir un moment pour se rendre compte que cela ne résulte aucunement de l'assertion correspondante ci-dessus concernant la propriété de composition ! La méthode de Vaughan Pratt repose sur un critère de primalité extrêmement utile, qui est une réciproque au petit théorème de Fermat.

Théorème 41. Pour que l'entier naturel impair n soit premier, il faut, et il suffit, qu'il existe $a \in \mathbb{Z}$ tel que $a^{n-1} \equiv 1 \pmod{n}$ et, pour tout facteur premier q de $n-1$, on ait $a^{(n-1)/q} \not\equiv 1 \pmod{n}$.

Démonstration. Si n est premier, il suffit de prendre pour a une racine primitive modulo n, autrement dit un entier tel que $\overline{a}$ soit un générateur du groupe cyclique $(\mathbb{Z}/n\mathbb{Z})^*$ (théorème 15 de la page 15).

Réciproquement, la condition $a^{n-1} \equiv 1 \pmod{n}$ assure que l'ordre de $\overline{a} \in (\mathbb{Z}/n\mathbb{Z})^*$ divise $n-1$ et les autres relations que cet ordre est égal à $n-1$, de sorte que le sous-groupe engendré par $\overline{a}$ est d'ordre $n-1$, d'où $\varphi(n) \geqslant n-1$, ce qui entraîne que n est premier. ∎

[10] En 2020. On peut démontrer rigoureusement le fait, intuitivement évident, que tout problème de la classe P est dans la classe NP. On ignore actuellement (début 2021) si la réciproque est vraie : c'est un très célèbre problème d'informatique théorique, ou de logique, intitulé « P = NP ? ». C'est aussi, comme l'Hypothèse de Riemann, l'objet de l'un des sept « Prix du millénaire » de la fondation Clay.

Corollaire 42. Pour que l'entier naturel impair n soit premier, il faut, et il suffit, qu'il existe $a \in \mathbb{Z}$, $1 < a < n$, tel que $a^{(n-1)/2} \equiv -1 \pmod{n}$ et, pour tout facteur premier impair q de $n-1$, on ait $a^{(n-1)/q} \not\equiv 1 \pmod{n}$.

Démonstration. La réduction à l'intervalle $1 < a < n$ est immédiate ; et, si n est premier, la relation $x^{n-1} = 1$ dans le corps $\mathbb{Z}/n\mathbb{Z}$ entraîne $x^{(n-1)/2} = \pm 1$, le cas $x^{(n-1)/2} = 1$ étant de plus exclu par hypothèse. ∎

Le théorème et son corollaire fournissent un test de primalité de n très efficace lorsque l'on connaît la décomposition de $n-1$ en facteurs premiers, comme nous le verrons à la section suivante pour les nombres de Fermat. Mais nous l'utiliserons ici de manière différente.

Lemme 43. Soit p un nombre premier impair. Il existe alors une suite strictement décroissante finie $p =: p_0 > p_1 > \cdots > p_k$ de nombres premiers impairs tels que chaque p_j, $1 \leqslant j \leqslant k$, soit diviseur de l'un des $p_i - 1$, $0 \leqslant i < j$, et que tout diviseur premier impair de l'un des $p_i - 1$, $0 \leqslant i \leqslant k$, soit l'un des p_j, $i < j \leqslant k$ (ce qui implique que $p_k - 1$ n'a pas de facteur premier impair). On a alors :

$$\sum_{i=0}^{k} \log_2 p_i < (\log_2 p)^2.$$

Démonstration. L'ensemble des p_i est construit récursivement en partant de $p_0 := p$, auquel on adjoint les diviseurs premiers impairs de $p_0 - 1$; puis en recommençant pour chacun des p_i ainsi apparus, etc. Le processus se termine puisque tous les nombres premiers qui apparaissent sont inférieurs à p, ils sont donc en nombre fini.
Pour prouver l'inégalité, notons $q_1, \ldots, q_l$ les facteurs premiers impairs de $p_0 - 1$ (donc la « première vague » d'adjonctions). On raisonne par récurrence forte, ce qui veut dire que l'on suppose que :

$$\log_2 p_1 + \cdots + \log_2 p_k < (\log_2 q_1)^2 + \cdots + (\log_2 q_l)^2.$$

Mais $q_1 \cdots q_l$ divise $(p-1)/2$, donc $\log_2 q_1 + \cdots + \log_2 q_l \leqslant \log_2 (p-1)/2 < \log_2 p - 1$, d'où l'on tire facilement que :

$$(\log_2 q_1)^2 + \cdots + (\log_2 q_l)^2 < (\log_2 p - 1)^2,$$

d'où :

$$\log_2 p_0 + \log_2 p_1 + \cdots + \log_2 p_k < \log_2 p + (\log_2 p - 1)^2 < (\log_2 p)^2.$$ ∎

Théorème 44 (Pratt).
Si p est premier, il existe un certificat le prouvant en temps polynomial.

Démonstration. On construit d'abord comme ci-dessus — même si c'est très long — une suite $p_0 > p_1 > \cdots > p_k$. D'après le corollaire ci-dessus, il existe pour chacun des p_i un entier a_i, $1 < a_i < p_i$, tel que $a_i^{(p_i-1)/2} \equiv -1 \pmod{p_i}$ et tel que, pour $p_j \mid p_i - 1$,

on ait $a_i^{(p_i-1)/p_j} \not\equiv 1 \pmod{p_i}$. Réciproquement, l'existence de ces a_i prouve que tous les p_i sont premiers, en particulier p. La preuve que p est premier à partir de ce certificat (acte de certification) consiste en le calcul de tous les $a_i^{(p_i-1)/2}$ et $a_i^{(p_i-1)/p_j}$ modulo p_i. D'après les estimations rappelées au paragraphe 5.1.3 de la page 41, chacun de ces calculs s'effectue en temps $\mathrm{O}\big((\log_2 p_i)^3\big)$, et, pour chaque p_i, le nombre de tels calculs est égal au nombre de facteurs premiers de $p_i - 1$, qui est strictement majoré par $\log_2 p_i$. Le coût total de la certification est donc $\mathrm{O}\big(\sum(\log_2 p_i)^4\big)$. Des majorations :

$$\sum(\log_2 p_i)^4 < (\sum \log_2 p_i)^4 < (\log_2 p)^8,$$

on conclut enfin que le coût de la certification est $\mathrm{O}(|p|^8)$. ∎

Exemple. Nous prenons $p := 947$, donc $p - 1 = 2 \times 11 \times 43$, puis :

$$11 - 1 = 2 \times 5, \quad 5 - 1 = 2 \times 2, \quad 43 - 1 = 2 \times 3 \times 7, \textit{etc.}$$

On obtient la liste : $p_0 = 947$, $p_1 = 43$, $p_2 = 11$, $p_3 = 7$, $p_4 = 5$ et $p_5 = 3$. Nous prendrons de plus (après tâtonnements) $a_5 := 2$, $a_4 := 2$, $a_3 := 3$, $a_2 := 2$, $a_1 := 3$, $a_0 := 2$. Une fois ces listes obtenues, la certification de la primalité de p (et simultanément de $p_1, \ldots, p_5$) comporte les calculs suivants (pour les deux premières lignes, $p_i - 1$ n'a pas de facteur premier impair) :

$$\text{certification de } p_5 : 2^{(3-1)/2} \equiv -1 \pmod 3,$$

$$\text{certification de } p_4 : 2^{(5-1)/2} \equiv -1 \pmod 5,$$

$$\text{certification de } p_3 : \begin{cases} 3^{(7-1)/2} \equiv -1 \pmod 7, \\ 3^{(7-1)/3} \equiv 2 \not\equiv 1 \pmod 7, \end{cases}$$

$$\text{certification de } p_2 : \begin{cases} 2^{(11-1)/2} \equiv -1 \pmod{11}, \\ 2^{(11-1)/5} \equiv 4 \not\equiv 1 \pmod{11}, \end{cases}$$

$$\text{certification de } p_1 : \begin{cases} 3^{(43-1)/2} \equiv -1 \pmod{43}, \\ 3^{(43-1)/3} \equiv 36 \not\equiv 1 \pmod{43}, \\ 3^{(43-1)/7} \equiv 41 \not\equiv 1 \pmod{43}, \end{cases}$$

$$\text{certification de } p_0 : \begin{cases} 2^{(947-1)/2} \equiv -1 \pmod{947}, \\ 2^{(947-1)/11} \equiv 215 \not\equiv 1 \pmod{947}, \\ 2^{(947-1)/43} \equiv 41 \not\equiv 1 \pmod{947}. \end{cases}$$

5.4.2 PRIMES est dans P

De 1983, date de son invention, à 2004 (voir ci-dessous), l'algorithme d'Adleman[11], Pomerance[12] et Rumely a été le meilleur test de primalité « généraliste », *i.e.* ne s'appliquant pas à une classe particulière de nombres. Son coût était estimé à $\mathrm{O}\big((\ln n)^{c \ln\ln\ln n}\big)$

[11] Il s'agit de Leonard Adleman, l'un des inventeurs de RSA.

[12] Il s'agit de Carl Pomerance, l'un des auteurs de la preuve qu'il y a une infinité de nombres de Carmichael.

pour un certain $c > 0$. Ce n'est donc pas un coût polynomial en $|n|$, puisque l'exposant $c \ln\ln\ln n$ de $\ln n = \Theta(|n|)$ n'est pas borné. Cependant cet exposant croît très lentement : par exemple, pour $n = 10^{500}$ on a $\ln\ln\ln n < 2$; et pour obtenir $\ln\ln\ln n = 3$, il faut prendre $n > 10^{220000000}$ (l'exposant se lit *deux cents vingt millions*). L'algorithme peut être considéré comme quasi-polynomial.
En 2004, les mathématiciens indiens Agrawal, Kayal et Saxena (un professeur et deux thésards) ont mis au point le premier algorithme vraiment polynomial. Leurs idées ont depuis été raffinées pour mettre au point de bons algorithmes. L'algorithme AKS est décrit en détail dans [Hindry] ; et un intéressant article (accessible en ligne) est consacré à cette découverte dans les *Notices of the American Mathematical Society* de mai 2003.

5.4.3 Fabrication de grands nombres premiers

Nous verrons en 5.5 des méthodes adaptées à des classes particulières de nombres premiers et permettant d'en découvrir de très grands. Malheureusement, la structure qui définit ces classes peut causer des problèmes de sécurité (informations indirectes) dans les applications informatiques souhaitées. Il est donc nécessaire de pouvoir faire appel à des tests généralistes.

Nous admettrons donc que nous disposons d'un test de primalité à coût polynomial (par exemple un raffinement de AKS) ; nous admettrons également, bien que nous ne l'ayons pas démontré, le théorème des nombres premiers : $\pi(n) \sim \dfrac{n}{\ln n}$ lorsque n tend vers $+\infty$. On en déduit que $\pi(10^N) \sim \dfrac{10^N}{N \ln 10}$ et que $\pi(10^{N-1}) \sim \dfrac{10^{N-1}}{(N-1)\ln 10} \sim \dfrac{10^{N-1}}{N \ln 10}$ donc que :

$$\pi(10^N) - \pi(10^{N-1}) \sim \frac{9.10^{N-1}}{N \ln 10}.$$

Il s'agit du nombre de nombres premiers qui s'écrivent avec N chiffres décimaux. En interprétant intuitivement la fréquence comme une probabilité, on voit que *la probabilité qu'un nombre de N chiffres décimaux « pris au hasard » soit premier est proche de* $\dfrac{1}{\alpha N}$, où $\alpha := \ln 10$. Si l'on suppose de nombreux « tirages indépendants » (nous utilisons toujours l'interprétation probabiliste dans un sens intuitif), l'espérance de premier succès est proche de αN (c'est une propriété de la loi de Bernoulli). Si l'on effectue CN tirages, la probabilité d'échec est de l'ordre de $\left(1 - \dfrac{1}{\alpha N}\right)^{CN} \approx e^{-C/\alpha}$, que l'on pourra rendre inférieure à tout $\varepsilon > 0$ donné en prenant $C > \alpha \ln(1/\varepsilon)$.

Par ailleurs, tester la primalité d'un nombre de N chiffres s'effectue en temps $\mathrm{O}(N^d)$ pour un certain $d > 0$, et, C ayant été choisi, le temps nécessaire pour trouver un nombre premier de N chiffres avec probabilité supérieure à $1 - \varepsilon$ est $\mathrm{O}(N^{d+1})$.

5.5 Deux classes spéciales de nombres premiers

Pour motiver l'étude qui va suivre, il est bon de résoudre l'exercice simple suivant.

Exercice 18.
Soient b, m des entiers naturels, $b, m \geqslant 2$. À quelle condition les nombres suivants sont-ils premiers : $b^m - 1$? $2^m + 1$?

Solution. Dans tous les cas, $b^m - 1$ est divisible par $b - 1$. Si $b \neq 2$, c'est un facteur non trivial ; si $m = kl$, $k, l \geqslant 2$, $b^m - 1$ est divisible par $b^k - 1$ qui en est un facteur non trivial. Pour que $b^m - 1$ soit premier, il est donc *nécessaire* que $b = 2$ et que m soit premier.
Si $m = kl$, $k \geqslant 2$, k impair, alors $2^m + 1 = (2^l)^k - (-1)^k$ est divisible par $2^l + 1$ qui en est un facteur non trivial. Pour que $2^m + 1$ soit premier, il est donc *nécessaire* que m n'admette aucun diviseur impair autre que 1, *i.e.* que ce soit une puissance de 2.

Comme nous le verrons, aucune des deux conditions trouvées n'est suffisante. Dans les deux numéros suivants, nous allons examiner de plus près les nombres qui les satisfont.

Remarque. Depuis 1925, le *Cunningham Project* vise à obtenir des factorisations d'entiers de la forme $b^m \pm 1$; voir leur site :

`http://homes.cerias.purdue.edu/~ssw/cun/index.html`

5.5.1 Nombres premiers de Fermat

Rappelons que nous notons $F_n := 2^{2^n} + 1$ les nombres dont Fermat se demandait s'ils étaient tous premiers (on prendra garde à ne pas confondre cette notation avec celle de la suite de Fibonacci). Il est clair que $F_0 = 3$, $F_1 = 5$, $F_2 = 17$ et $F_3 = 257$ sont premiers, mais qu'en est-il par exemple de $F_4 = 65537$ et de $F_5 = 4294967297$? Voici un critère commode pour aider à confirmer ou infirmer cette hypothèse dans certains cas simples.

Proposition 45 (Euler).
On suppose que $n \geqslant 2$. Tout diviseur de F_n est congru à 1 modulo 2^{n+2} :

$$\forall d \in \mathbb{N}\ ,\ d \mid F_n \Longrightarrow d \equiv 1 \pmod{2^{n+2}}.$$

Démonstration. Le produit de deux nombres congrus à 1 modulo 2^{n+2} l'est encore, il suffit donc de vérifier l'assertion lorsque p est premier, $p \mid F_n$. Cette dernière relation dit que $2^{2^n} \equiv -1 \pmod{p}$, ce qui entraîne que $2^{2^{n+1}} \equiv 1 \pmod{p}$, donc que l'ordre de $\overline{2} \in \mathbb{F}_p{}^*$ est une puissance de 2, donc (en vertu de la congruence de départ) qu'il est exactement égal à 2^{n+1}. Comme cet ordre divise $|\mathbb{F}_p{}^*| = p - 1$ (théorème de Lagrange), on a $p \equiv 1 \pmod{2^{n+1}}$.
Puisque $n \geqslant 2$, on a $p \equiv 1 \pmod 8$, d'où $p^2 \equiv 1 \pmod{16}$ (calcul immédiat), d'où $\left(\dfrac{2}{p}\right) = (-1)^{(p^2-1)/8} = 1$, de sorte que 2 est résidu quadratique modulo p, *i.e.* $2^{(p-1)/2} \equiv 1 \pmod{p}$. L'ordre de $\overline{2} \in \mathbb{F}_p{}^*$ divise donc $(p-1)/2$, or nous avons vu que c'est 2^{n+1}, et la conclusion s'ensuit. ∎

Si l'on cherche des diviseurs de F_n, $n \geqslant 2$, on pourra donc les chercher sous la forme $1 + k.2^{n+2}$, $k \in \mathbb{N}$.

Exemples. Tout diviseur de $F_4 = 2^{16} + 1 = 65537$ est de la forme $1 + 64k$. On cherche donc les nombres premiers de la forme $p = 1 + 64k$ et tels que $p \leqslant \lfloor\sqrt{F_4}\rfloor = 2^8 = 256$. Le seul est 193, qui ne divise pas 65537 qui est donc premier.
Tout diviseur de $F_5 = 2^{32} + 1 = 4294967297$ est de la forme $1 + 128k$. On cherche donc les nombres premiers de la forme $p = 1 + 128k$ et tels que $p \leqslant \lfloor\sqrt{F_5}\rfloor = 2^{16} = 65536$, ce qui entraîne que $k < 2^9$, *i.e.* $k \leqslant 511$. L'essai $k = 2$ donne $p = 257$, qui ne divise pas 4294967297, mais le nombre premier suivant correspond à $k = 5$, c'est $p = 641$ qui divise F_5 :

$$F_5 = 641 \times 6700417.$$

Donc F_5 n'est pas premier et le calcul qui a permis d'en *trouver* une certification était très simple. Si l'on veut factoriser complètement F_5, on cherche maintenant les nombres premiers $p = 1 + 128k$ avec $k \geqslant 5$ et $p \leqslant \lfloor\sqrt{6700417}\rfloor$, ce qui par calcul mental implique facilement $k \leqslant 25$. La suite est laissée au lecteur ...

Voici une critère plus systématique qui se déduit du théorème 41 de la page 46. En effet, ce dernier faisait dépendre la primalité de p de calculs portant sur les diviseurs de $p - 1$, qui sont particulièrement simples lorsque $p = F_n$!

Théorème 46 (Test de Pépin, 1877). On a une équivalence logique :

$$m^{(F_n-1)/2} \equiv -1 \pmod{F_n} \Longleftrightarrow \left(F_n \text{ premier et } \left(\frac{m}{F_n}\right) = -1\right).$$

Démonstration. L'implication directe découle immédiatement du théorème 41 puisque le seul facteur premier de $F_n - 1$ est 2.
L'implication réciproque est triviale d'après la caractérisation par Euler du symbole de Legendre (théorème 20 de la page 22). ■

Tant que l'on ne sait pas que F_n est premier (mais il est certainement impair, et même de la forme $4m + 1$), c'est le symbole de Jacobi qui apparaît dans ce critère. La loi de réciprocité quadratique pour le symbole de Jacobi (avec ici $(F_n - 1)/2$ pair) autorise alors le calcul :

$$\left(\frac{3}{F_n}\right) = \left(\frac{F_n}{3}\right) = \left(\frac{2}{3}\right) = -1,$$

l'avant-dernière égalité découlant des congruences :

$$2 \equiv -1 \pmod 3 \Longrightarrow 2^{2^n} \equiv 1 \pmod 3 \Longrightarrow F_n \equiv 2 \pmod 3.$$

Corollaire 47 (Test de Pépin, 1877). Pour que F_n soit premier, il faut, et il suffit, que l'on ait la congruence $3^{(F_n-1)/2} \equiv -1 \pmod{F_n}$.

Démonstration. La nécessité vient de ce qu'autrement 3 serait un témoin de Fermat. La suffisance découle du théorème et du calcul qui précèdent. ■

Exemples. Pour savoir si F_4 est premier, on peut calculer le reste de la division euclidienne de $3^{2^{15}}$ par F_4. Après quinze multiplications (exponentiation dichotomique) et une division euclidienne, on trouve le reste $65536 = F_5 - 1$, d'où la primalité par le test de Pépin.

On ignore actuellement s'il y a des F_n premiers au-delà de F_4 ; on ignore de même s'il y a une infinité de F_n qui sont premiers, et aussi s'il y en a une infinité qui sont composés.

5.5.2 Nombres premiers de Mersenne

Le nombre $M_p := 2^p - 1$ n'est premier que si p est premier, on l'appelle alors *nombre premier de Mersenne*. On vérifie immédiatement que $M_2 = 3$, $M_3 = 7$, $M_5 = 31$ et $M_7 = 127$ sont premiers, mais $M_{11} = 2047 = 23 \times 89$ ne l'est pas. On ignore actuellement s'il y a une infinité de nombres premiers de Mersenne ; cependant le plus grand nombre premier connu à une époque donnée a souvent été un nombre de Mersenne. Fin 2014, le plus grand nombre premier connu était $M_{57885161}$. Fin 2020, c'est $M_{82589933}$ (trouvé en 2018), qui a près de vingt-cinq millions de chiffres[13] On ignore également (début 2021) s'il y a une infinité de M_p qui sont composés.

C'est ici la décomposition de $n+1$ qui est simple, les critères portant sur la décomposition de $n-1$ ne s'appliquent pas facilement. Voici un outil simple dû au Maître.

Théorème 48 (Fermat, 1640). Soit p un nombre premier impair et soit q un facteur premier de M_p. Alors $q \equiv 1 \pmod{2p}$.

Démonstration. Puisque $2^p \equiv 1 \pmod{q}$, l'ordre de $\overline{2} \in \mathbb{F}_q^*$ divise p, donc est égal à p, qui divise donc $\left|\mathbb{F}_q^*\right| = q-1$. Par ailleurs, q est impair donc $2 \mid q-1$, donc $2p \mid q-1$. ■

Exemples. Fermat a ainsi démontré que M_{23} et M_{37} sont premiers ; et Euler (1738) l'a fait pour M_{29} (voir les exercices à la fin du chapitre).

Exercice 19.
Examiner le cas de M_{11}.
Solution. On a $M_{11} = 2047$ et ses facteurs premiers doivent être de la forme $q = 1 + 22k$. On constate immédiatement que $2047 = 23 \times 89$.

Voici maintenant une application arithmétique des nombres premiers de Mersenne. Le résultat qui suit remonte à Euclide pour le sens direct (description d'une famille de nombres parfaits pairs), à Euler pour la réciproque (tous sont dans la famille).

Définition 4. Un entier naturel est dit *parfait* s'il est la somme de ses diviseurs stricts.

Exemples. Les entiers $6 = 1 + 2 + 3$ et $28 = 1 + 2 + 4 + 7 + 14$ sont parfaits.

[13] Ces informations sont régulièrement actualisées sur le site

`https://en.wikipedia.org/wiki/Largest_known_prime_number`

Théorème 49. Les nombres parfaits pairs sont les $2^{p-1}M_p$, où p est premier et où M_p est un nombre premier de Mersenne.

La preuve (facile) est proposée en exercice à la fin du chapitre. Les exemples ci-dessus correspondent à $p = 2, 3$. On ignore actuellement s'il existe des nombres parfaits impairs.

6 Fractions continues

En vue d'applications aux fonctions spéciales dans le module III.4[14], nous allons étudier des objets étroitement liés à l'algorithme d'Euclide de calcul du pgcd. Le lecteur est donc invité à revoir ce qui concerne cet algorithme dans les modules d'arithmétique et d'algorithmique du livre [L1], mais également dans le module sur les polynômes ; comme la théorie des fractions continues est largement de nature algorithmique, des révisions sur le module « Structures discrètes et récursivité » du livre [L2] sont également recommandées.
Comme l'étude des deux cas (arithmétique dans $\mathbb{Z}$ et dans $K[X]$) est similaire, nous la mènerons en parallèle aussi loin que possible (sections 6.1, 6.2 et 6.3). Lorsque les formules diffèrent, nous parlerons du *cas entier* et du *cas polynomial*. Dans ce dernier cas, la lettre K désignera un corps commutatif quelconque.

6.1 Compléments à l'algorithme d'Euclide

Commençons par réécrire l'algorithme d'Euclide sous forme récursive. Nous supposerons que a et b sont deux entiers naturels tels que $b < a$; ou bien deux polynômes tels que $\deg b < \deg a$. Nous ferons appel à la fonction **`diveucl(a,b)`** (module d'algorithmique du livre [L1]) qui calcule et rend le couple **`(q,r)`** formé du quotient et du reste de la division euclidienne de **`a`** par **`b`**.

```
pgcd(a,b)
    si (b = 0) alors rendre(a)
    sinon ((q,r) := diveucl(a,b) ; rendre(pgcd(b,r))) ;;
```

Profitons-en pour alléger notre style de programmation fonctionnelle récursive, en nous rapprochant de celui de CAML. On écrit alors :

```
pgcd(a,b) =
    si (b = 0) alors a
    sinon soit (q,r) = diveucl(a,b) dans pgcd(b,r) ;;
```

On ne dit pas que la fonction **`pgcd`** *rend* **`0`** ou une autre valeur, mais que la valeur de **`pgcd(a,b)`** est *égale* à **`0`** ou à celle de **`pgcd(b,r)`**. On ne dit pas non plus que l'on *affecte* les variables **`q`** et **`r`**, mais que l'on *dénote* par ces lettres le quotient et le reste en vue de leur utilisation dans l'expression **`pgcd(b,r)`**. Le style qui en résulte est un peu plus proche de celui d'un texte mathématique. Voici la version CAML :

```
let rec pgcd(a,b) =
    if (b = 0) then a
    else let (q,r) = diveucl(a,b) in pgcd(b,r) ;;
```

[14] Voir aussi des applications à l'arithmétique dans le livre [MPA1].

En notation mathématique, on a une suite définie par récurrence :

$$x_0 := a, \quad x_1 := b \quad \text{et} \quad x_{i+2} := x_i - q_i x_{i+1}, \tag{1}$$

où l'on a noté q_i le quotient de la division euclidienne de x_i par x_{i+1} lorsqu'elle est possible (attention : par rapport au module d'algorithmique du livre [L1], nous avons décalé les indices des q_i pour les rendre conformes aux notations usuelles dans l'étude des fractions continues). La suite s'arrête en un rang $k+2$ tel que $x_k = q_k x_{k+1}$ et $x_{k+2} = 0$. Le résultat de l'algorithme est le pgcd $\delta := x_{k+1}$. Les couples (x_i, x_{i+1}) sont les arguments successifs des appels récursifs de la fonction **pgcd**. Ils vérifient les inégalités : $x_{i+1} < x_i$ (cas entier), respectivement $\deg x_{i+1} < \deg x_i$ (cas polynomial). Il en résulte que, dans le cas entier, $q_i \geqslant 1$ et $q_k \geqslant 2$; et, dans le cas polynomial, $\deg q_i \geqslant 1$. Rappelons (exercices du module de calcul matriciel du livre [L1]) que les égalités ci-dessus admettent une interprétation matricielle :

$$\forall i \in [\![0, k]\!] \ , \ \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix} = \begin{pmatrix} q_i & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_{i+1} \\ x_{i+2} \end{pmatrix}. \tag{2}$$

Il y a de bonnes raisons pour conserver les résultats intermédiaires des calculs. Par exemple, nous avons vu que la connaissance des q_i permet de calculer les coefficients de Bézout ; et nous avons vu dans le module « Méthodes numériques » du livre [L2] que, dans le cas polynomial, si b est le polynôme dérivé de a, les x_i sont (au signe près) les termes d'une *suite de Sturm*. Nous nous concentrerons ici sur les quotients successifs q_i. Ceux-ci expriment la « structure arithmétique » de la fraction $\dfrac{a}{b}$. Remarquons en effet que, si l'on remplace (a, b) par (da, db) (où d est un entier ou un polynôme non nul) les x_i sont tous multipliés par d, mais les q_i ne changent pas. La suite des q_i ne dépend donc que de la fraction $\dfrac{a}{b}$.

Le programme suivant calcule la liste $[q_0, \ldots, q_k]$ de ces quotients. Nous noterons **q :: l** la liste obtenue par ajout d'un élément **q** en tête d'une liste **l**.

```
quotients(a,b) =
    si b = 0 alors []                    (* liste vide *)
    sinon soit (q,r) = diveucl(a,b)
        dans q :: quotients(b,r) ;;
```

Par exemple,

- **quotients(10,6) = 1 :: quotients(6,4)**,
- **quotients(6,4) = 1 :: quotients(4,2)**,
- **quotients(4,2) = 2 :: quotients(2,0)**,
- **quotients(2,0) = []**,

de sorte que **quotients(10,6) = [1,1,2]**.

De même,

- **quotients(X^3+X-1,X^2) = X :: quotients(X^2,X-1)**,
- **quotients(X^2,X-1) = (X+1) :: quotients(X-1,1)**,
- **quotients(X-1,1) = (X-1) :: quotients(1,0)**,
- **quotients(1,0) = []**,

de sorte que **quotients(X^3+X-1,X^2) = [X,X+1,X-1]**.

On peut reconstituer le couple (a, b) à partir de le suite des quotients q_i et du pgcd δ par récurrence descendante : $x_{k+2} := 0$, $x_{k+1} := \delta$ et, pour i de k à 0, $x_i := x_{i+2} + q_i x_{i+1}$. En notation matricielle :

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} q_0 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} q_k & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \delta \\ 0 \end{pmatrix}. \tag{3}$$

Voici le programme correspondant. Il procède par induction structurelle sur la liste l des quotients, qui est soit vide soit de profil $q :: s$, où q est l'élément de tête et s la suite de la liste. Comme en style CAML, nous indiquons explicitement le filtrage.

```
couple(l,delta) = filtrer l par  (* en CAML: match l with *)
  | [] -> (delta,0)
  | q :: s -> soit (b,r) = couple(s,delta) dans (q*b+r,b) ;;
```

Ainsi, **`couple([1,1,2],2) = (10,6)`**.

Le pgcd δ étant donné, il n'y a pas une réciprocité exacte entre les algorithmes décrits. Le lecteur vérifiera en effet que **`couple([1,1,1,1],2) = (10,6)`**. Pourtant, c'est bien **`[1,1,2]`**, et non **`[1,1,1,1]`** qui est la « vraie » liste de quotients attachée à **`(10,6)`**. Ce point sera précisé par le théorème 51 de la page 57.

Puisque les quotients q_i ne dépendent que de la fraction $\alpha := \dfrac{a}{b}$, on doit pouvoir mener les calculs à partir de cette dernière. Dans le cas entier, nous savons que le quotient de la division euclidienne de a par b est la partie entière $\left\lfloor \dfrac{a}{b} \right\rfloor$. Dans le cas polynomial, c'est encore vrai, avec la définition de la partie entière donnée dans le module sur les polynômes du livre [L1] : la partie entière $\lfloor \alpha \rfloor$ est l'unique polynôme $q \in K[X]$ tel que la fraction rationnelle $\alpha - q$ soit de degré strictement négatif. Rappelons que le degré d'une fraction rationnelle β est défini par la formule $\deg \beta := \deg c - \deg d$, où (c, d) est n'importe quel couple tel que $\dfrac{c}{d} = \beta$. Le programme suivant suppose $\alpha > 0$ (cas entier), respectivement $\deg \alpha > 0$ (cas polynomial). On note **`floor`** la partie entière.

```
quotients(alpha) = soit q = floor(alpha) dans
  si (alpha = q) alors [q]
  sinon q :: quotients(1/(alpha-q)) ;;
```

Réciproquement, pour calculer la fraction $\dfrac{a}{b}$ (et non le couple (a, b)) à partir de la liste non vide $[q_0, \ldots, q_k]$, on devra filtrer la liste selon les profils **`[q]`** (un seul élément) ou **`q :: s`** (avec une suite non vide) :

```
fractions(l) = filtrer l par
  | [q] -> q
  | q :: s -> q + 1/(fractions(s)) ;;
```

6.2 Développement en fraction continue dans $\mathbb{Q}$ et dans $K(X)$

Théorème et définition 50. Soit $\alpha > 1$ un rationnel (cas entier), respectivement $\alpha \in K(X)$ une fraction rationnelle de degré $\deg \alpha > 0$ (cas polynomial). On pose $\alpha_0 := \alpha$, $q_0 := \lfloor \alpha_0 \rfloor$; puis, tant que $q_i \neq \alpha_i$: $\alpha_{i+1} := \dfrac{1}{\alpha_i - q_i}$ et $q_{i+1} := \lfloor \alpha_{i+1} \rfloor$. Il existe alors un entier k tel que $q_k = \alpha_k$. La liste $[q_0, \ldots, q_k]$ est appelée *développement en fraction continue* de α. Les q_i sont appelés *quotients incomplets* de ce développement. On a l'égalité :

$$\alpha = q_0 + \cfrac{1}{q_1 + \cfrac{1}{\ddots + \cfrac{1}{q_k}}}. \tag{4}$$

L'expression du membre droit est appelée *fraction continue*. On la trouve parfois sous la forme $q_0 + \dfrac{1|}{|q_1} + \cdots + \dfrac{1|}{|q_k}$. Nous l'abrégerons en $[q_0, \ldots, q_k]$: donc, *avec la même notation que la liste des quotients incomplets*.
Dans le cas entier, les quotients incomplets vérifient $q_i \geqslant 1$ $(0 \leqslant i \leqslant k-1)$ et $q_k \geqslant 2$.
Dans le cas polynomial, les quotients incomplets vérifient $\deg q_i \geqslant 1$ $(0 \leqslant i \leqslant k)$.

Démonstration. Écrivons $\alpha = \dfrac{a}{b}$ et introduisons la suite des x_i de la section précédente (dont les hypothèses sur a et b sont vérifiées). Si l'on pose $y_i := \dfrac{x_i}{x_{i+1}}$, on voit que les y_i sont définis par les mêmes relations de récurrence que les α_i. On a donc $y_i = \alpha_i$ pour $0 \leqslant i \leqslant k$, et en particulier $\alpha_k = y_k = \dfrac{x_k}{x_{k+1}} = q_k$: le processus s'arrête donc à l'indice k. On a de plus : $\alpha = \alpha_0 = q_0 + \dfrac{1}{\alpha_1}, \ldots, \alpha_{k-1} = q_{k-1} + \dfrac{1}{\alpha_k}$, et $\alpha_k = q_k$, d'où la formule (4).
La dernière assertion résulte des remarques qui suivent l'équation (1) de la page 54. ■

Exercice 20.
Développer en fraction continue $\dfrac{11}{8}$ et $\dfrac{X^3 + X - 1}{X^2}$.

Solution. On écrit $\dfrac{11}{8} = 1 + \dfrac{3}{8}$, $\dfrac{8}{3} = 2 + \dfrac{2}{3}$, $\dfrac{3}{2} = 1 + \dfrac{1}{2}$ et $\dfrac{2}{1} = 2$, d'où :

$$\frac{11}{8} = 1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2}}} = [1, 2, 1, 2] = 1 + \frac{1|}{|2} + \frac{1|}{|1} + \frac{1|}{|2}.$$

Les calculs qui suivent le programme **quotients** de la page 54 donnent directement :

$$\frac{X^3+X-1}{X^2} = X + \cfrac{1}{X+1+\cfrac{1}{X-1}} = [X, X+1, X-1] = X + \frac{1|}{|X+1} + \frac{1|}{|X-1}.$$

Remarque. Ces notions peuvent être étendues à tout élément α de $\mathbb{Q}$ ou de $K(X)$. On pose encore $\alpha_0 := \alpha$ et $q_0 := \lfloor \alpha_0 \rfloor$. Si $q_0 = \alpha_0$, le développement en fraction continue de α est q_0. Sinon, à partir de $\alpha_1 := \dfrac{1}{\alpha_0 - q_0}$, les hypothèses précédentes sont vérifiées. On a donc exactement les mêmes conclusions, sauf en ce qui concerne q_0, qui est alors un entier ou un polynôme arbitraire. Pour simplifier l'étude, nous nous placerons par la suite dans les conditions du théorème 50.

Quels que soient les $q_0, \ldots, q_k$ rationnels strictement positifs ou éléments de $K(X)$ de degrés strictement positifs, on peut définir la *fraction continue* :

$$[q_0, \ldots, q_k] := q_0 + \cfrac{1}{q_1 + \cfrac{1}{\ddots + \cfrac{1}{q_k}}} = q_0 + \frac{1|}{|q_1} + \cdots + \frac{1|}{|q_k}.$$

C'est en effet clair s'il n'y a qu'un terme ($k = 0$) et l'on voit par récurrence que, si $[q_{i+1}, \ldots, q_k]$ est strictement positif ou de degré strictement positif, alors il en est de même de :

$$[q_i, \ldots, q_k] = q_i + \frac{1}{[q_{i+1}, \ldots, q_k]}.$$

Les fractions continues de ce type sont parfois appelées *simples* parce que tous les numérateurs sont égaux à 1. Dans la section 6.3.2 apparaîtront des fractions continues un peu plus générales. Pour les fractions continues simples, on a les formules utiles suivantes :

$$[q_0, \ldots, q_{i-1}, q_i, q_{i+1}] = [q_0, \ldots, q_{i-1}, q_i + \frac{1}{q_{i+1}}].$$

Plus généralement :

$$[q_0, \ldots, q_{i-1}, q_i, \ldots, q_j] = [q_0, \ldots, q_{i-1}, [q_i, \ldots, q_j]].$$

La fraction continue calculée à partir du développement en fraction continue de α a pour valeur α. La réciproque est vraie, sous certaines conditions :

Théorème 51. On suppose que $k \geqslant 1$ et que les q_i $(0 \leqslant i \leqslant k)$ sont des entiers naturels non nuls, ou des polynômes non constants. Alors $\alpha := [q_0, \ldots, q_k]$ vérifie les conditions du théorème 50 de la page précédente. Son développement en fraction continue est $[q_0, \ldots, q_k]$, sauf, dans le cas entier, si $q_k = 1$. Dans ce dernier cas, son développement en fraction continue est $[q_0, \ldots, q_{k-1} + 1]$.

Démonstration. Il est clair que $\alpha = q_0 + \dfrac{1}{[q_1, \ldots, q_k]}$ est tel que $\alpha > 1$ (cas entier), respectivement $\deg \alpha > 0$ (cas polynomial). Si $k \geqslant 2$, puisque $[q_1, \ldots, q_k]$ vérifie à son

tour ces conditions, la partie entière de α est q_0 et son développement en fraction continue commence par q_0 et se poursuit par celui de $[q_1, \ldots, q_k]$. Si $k = 1$, sa partie entière est encore q_0, sauf si $q_k = q_1$ est l'entier 1 : dans ce cas, sa partie entière est $q_0 + 1$. ■

Nous dirons qu'une fraction continue simple $[q_0, \ldots, q_k]$ est *propre* si elle vérifie les conditions du théorème et si de plus, dans le cas entier, $q_k \geqslant 2$. Pour de telles fractions, il y a bien réciprocité entre les algorithmes de développement et d'évaluation.

6.3 Les réduites d'une fraction continue sur un corps arbitraire

6.3.1 Fractions continues simples

La valeur d'une fraction continue $[q_0, \ldots, q_k]$ a été définie par récurrence descendante, comme $q_0 + \dfrac{1}{[q_1, \ldots, q_k]}$ (et $[q_0] = q_0$). Cette définition garde un sens pour des éléments q_i d'un corps commutatif L quelconque (au lieu de $\mathbb{Q}$ ou $K(X)$), sous réserve qu'aucun dénominateur ne s'annule. Le calcul en est également possible par récurrence ascendante.
Nous appellerons *réduites* de cette fraction continue les $r_i := [q_0, \ldots, q_i]$, pour $i \in [\![0, k]\!]$.
Pour les étudier, nous utiliserons les notations et résultats sur les homographies du module « Groupes classiques » du livre [L2]. En particulier, l'expression $q + \dfrac{1}{x}$ peut se réécrire : $\dfrac{qx+1}{x} = \begin{bmatrix} q & 1 \\ 1 & 0 \end{bmatrix}(x)$. La formule $[q_0, \ldots, q_k] = q_0 + \dfrac{1}{[q_1, \ldots, q_k]}$ se traduit donc par : $[q_0, \ldots, q_k] = \begin{bmatrix} q_0 & 1 \\ 1 & 0 \end{bmatrix}\big([q_1, \ldots, q_k]\big)$. En itérant, on trouve :

$$[q_0, \ldots, q_k] = \begin{bmatrix} q_0 & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} q_{k-1} & 1 \\ 1 & 0 \end{bmatrix}(q_k) = \begin{bmatrix} q_0 & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} q_k & 1 \\ 1 & 0 \end{bmatrix}(\infty), \tag{5}$$

en vertu de l'égalité $\begin{bmatrix} q & 1 \\ 1 & 0 \end{bmatrix}(\infty) = q$. Il en découle en particulier que $[q_0, \ldots, q_k]$ *peut être défini sans aucune condition sur les* q_i *comme élément de la droite projective* $\mathbf{P}^1(L) = L \cup \{\infty\}$.
Pour $i \in [\![0, k]\!]$, nous noterons :

$$\begin{pmatrix} u_i & u'_i \\ v_i & v'_i \end{pmatrix} := \begin{pmatrix} q_0 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} q_i & 1 \\ 1 & 0 \end{pmatrix}.$$

De la relation $\begin{pmatrix} u_i & u'_i \\ v_i & v'_i \end{pmatrix} = \begin{pmatrix} u_{i-1} & u'_{i-1} \\ v_{i-1} & v'_{i-1} \end{pmatrix} \begin{pmatrix} q_i & 1 \\ 1 & 0 \end{pmatrix}$, on tire que, pour $i \geqslant 1$, $u'_i = u_{i-1}$ et $v'_i = v_{i-1}$.

Théorème 52. Définissons deux suites $(u_i)_{i\in\mathbb{N}}$ et $(v_i)_{i\in\mathbb{N}}$ par les relations :

$$u_0 := q_0, \quad u_1 := q_0q_1 + 1 \qquad \text{et} \qquad \forall i \in [\![2,k]\!]\,,\ u_i := q_iu_{i-1} + u_{i-2}$$

$$v_0 := 1, \quad v_1 := q_1 \qquad \text{et} \qquad \forall i \in [\![2,k]\!]\,,\ v_i := q_iv_{i-1} + v_{i-2}.$$

Alors les réduites r_i sont données par les formules :

$$\forall i \in [\![0,k]\!]\ ,\ r_i = \frac{u_i}{v_i}\cdot \tag{6}$$

Elles vérifient l'équation suivante :

$$\forall i \geqslant 1\ ,\ r_i - r_{i-1} = \frac{(-1)^{i-1}}{v_iv_{i-1}}\cdot \tag{7}$$

Démonstration. D'après le calcul précédent, les initialisations et relations de récurrence indiquées traduisent l'égalité : $\begin{pmatrix} u_i & u_{i-1} \\ v_i & v_{i-1} \end{pmatrix} := \begin{pmatrix} q_0 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} q_i & 1 \\ 1 & 0 \end{pmatrix}$. On déduit alors des règles de calcul sur les homographies (module « Groupes classiques ») :

$$r_i = \begin{bmatrix} q_0 & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} q_i & 1 \\ 1 & 0 \end{bmatrix} (\infty) = \begin{bmatrix} u_i & u_{i-1} \\ v_i & v_{i-1} \end{bmatrix} (\infty) = \frac{u_i}{v_i}\cdot$$

En comparant les déterminants, on trouve : $u_iv_{i-1} - v_iu_{i-1} = (-1)^{i-1}$.
L'égalité (7) s'en déduit en divisant cette relation par v_iv_{i-1}. ∎

Proposition et définition 53. Soit $[q_0, \ldots, q_k]$ le développement en fraction continue de α, qui vérifie les conditions du théorème 50 de la page 56. On appelle *quotients complets* de α les $\alpha_i := [q_i, \ldots, q_k]$ qui apparaissent lors du déroulement de l'algorithme. On a alors :

$$\alpha = \frac{u_i\alpha_{i+1} + u_{i-1}}{v_i\alpha_{i+1} + v_{i-1}}\cdot$$

Démonstration. On écrit

$$\alpha = [q_0, \ldots, q_i, [q_{i+1}, \ldots, q_k]] = [q_0, \ldots, q_i, \alpha_{i+1}] = \begin{bmatrix} u_i & u_{i-1} \\ v_i & v_{i-1} \end{bmatrix} (\alpha_{i+1}).$$ ∎

6.3.2 Fractions continues générales

Les fractions continues apparaissent parfois sous la forme plus générale :

$$q_0 + \cfrac{p_1}{q_1 + \cfrac{p_2}{\ddots + \cfrac{p_k}{q_k}}} = q_0 + \frac{p_1|}{|q_1} + \cdots + \frac{p_k|}{|q_k}\cdot \tag{8}$$

Si l'un des p_i était nul, toutes les fractions suivantes seraient inutiles : nous les supposerons donc non nuls. Nous allons voir comment de telles fractions continues se ramènent à la

forme simple. Par exemple :

$$q_0 + \frac{p_1|}{|q_1} + \frac{p_2|}{|q_2} = q_0 + \frac{1|}{|p_1^{-1}q_1} + \frac{p_1^{-1}p_2|}{|q_2} = q_0 + \frac{1|}{|p_1^{-1}q_1} + \frac{1|}{|(p_1^{-1}p_2)^{-1}q_2}\cdot$$

Ce calcul repose sur l'utilisation répétée des égalités suivantes :

$$\frac{1}{\alpha}\left(q + \frac{p}{x}\right) = \frac{q}{\alpha} + \frac{1}{(p/\alpha)^{-1}x}\cdot \tag{9}$$

Introduisons donc, pour $p, \alpha \neq 0$ des notations pour les homographies $\dfrac{x}{\alpha}$ et $q + \dfrac{p}{x}$:

$$D(\alpha) := \begin{bmatrix} 1 & 0 \\ 0 & \alpha \end{bmatrix} \quad \text{et} \quad R(q,p) := \begin{bmatrix} q & p \\ 1 & 0 \end{bmatrix}.$$

La relation (9) se traduit par l'égalité :

$$D(\alpha)R(q,p) = R\left(\frac{q}{\alpha}, 1\right) D\left(\frac{p}{\alpha}\right).$$

Quitte à poser arbitrairement $p_{k+1} := 1$, l'expression (8) prend alors la forme :

$$R(q_0, p_1) \cdots R(q_k, p_{k+1})(\infty). \tag{10}$$

Comme à la section précédente, nous pouvons introduire les réduites :

$$r_i := q_0 + \frac{p_1|}{|q_1} + \cdots + \frac{p_i|}{|q_i}\cdot$$

Proposition 54. Définissons deux suites $(u_i)_{i\in\mathbb{N}}$ et $(v_i)_{i\in\mathbb{N}}$ par les relations :

$$u_0 := q_0, \quad u_1 := q_0q_1 + p_1 \qquad \text{et} \qquad \forall i \in [\![2,k]\!]\,,\ u_i := q_iu_{i-1} + p_iu_{i-2}$$

$$v_0 := 1, \quad v_1 := q_1 \qquad \text{et} \qquad \forall i \in [\![2,k]\!]\,,\ v_i := q_iv_{i-1} + p_iv_{i-2}.$$

Alors les réduites r_i sont données par les formules :

$$\forall i \in [\![0,k]\!]\,,\ r_i = \frac{u_i}{v_i}\cdot \tag{11}$$

Démonstration. C'est la même que pour le théorème 52 de la page précédente, avec ici la formule :

$$\begin{bmatrix} u_i & p_{i+1}u_{i-1} \\ v_i & p_{i+1}v_{i-1} \end{bmatrix} = R(q_0,p_1)R(q_1,p_2)\cdots R(q_i,p_{i+1}).$$

Pour transformer l'expression (10), on multiplie l'homographie $R(q_0,p_1)\cdots R(q_k,p_{k+1})$ à gauche par $D(1)$, qui est l'identité :

$$\begin{aligned} R(q_0,p_1)R(q_1,p_2)\cdots R(q_k,p_{k+1}) &= D(1)R(q_0,p_1)R(q_1,p_2)\cdots R(q_k,p_{k+1}) \\ &= R(q_0,1)D(p_1)R(q_1,p_2)\cdots R(q_k,p_{k+1}) \\ &= R(q_0,1)R(p_1^{-1}q_1,1)D(p_1^{-1}p_2)\cdots R(q_k,p_{k+1}), \text{etc.} \end{aligned}$$

En itérant, on obtient le lemme suivant :

Lemme 55. On définit deux suites par les relations de récurrence suivantes :

$$a_0 := 1,\ b_0 := a_0 p_1 \quad \text{et, pour } i = 1, 2, \ldots, k, \quad a_i := \frac{1}{b_{i-1}} \quad \text{et} \quad b_i := a_i p_{i+1}.$$

Posons également $q'_i := q_i a_i$ pour $i = 0, \ldots, k$. Alors :

$$R(q_0, p_1) \cdots R(q_k, p_{k+1}) = R(q'_0, 1) \cdots R(q'_k, 1) D(b_k). \tag{12}$$

Théorème 56. Avec les notations du lemme, on a la formule :

$$q_0 + \cfrac{p_1}{q_1 + \cfrac{p_2}{\ddots + \cfrac{p_k}{q_k}}} = q'_0 + \cfrac{1}{q'_1 + \cfrac{1}{\ddots + \cfrac{1}{q'_k}}}.$$

Démonstration. Il suffit d'appliquer les deux membres de l'égalité (12) à ∞. ∎

Exemple. Considérons, dans $K(X)$, la fraction continue suivante :

$$\frac{1|}{|A_1X + B_1} - \frac{C_2|}{|A_2X + B_2} - \frac{C_3|}{|A_3X + B_3} - \cdots - \frac{C_k|}{|A_kX + B_k},$$

où tous les C_i sont non nuls. Cette fraction continue est égale à :

$$\frac{1|}{|A_1X + B_1} + \frac{1|}{|A'_2X + B'_2} + \frac{1|}{|A'_3X + B'_3} + \cdots + \frac{1|}{|A'_kX + B'_k},$$

où $A'_2 := -\dfrac{A_2}{C_2}$, $B'_2 := -\dfrac{B_2}{C_2}$, $A'_3 := \dfrac{C_2A_3}{C_3}$, $B'_3 := \dfrac{C_2B_3}{C_3}$, $A'_4 := -\dfrac{C_3A_4}{C_2C_4}$, $B'_4 := -\dfrac{C_3B_4}{C_2C_4}$, etc.

6.4 Développement en fraction continue d'un réel

Si l'on part d'un *réel irrationnel* $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, on peut encore poser $\alpha_0 := \alpha$, $q_0 := \lfloor \alpha_0 \rfloor$; puis, tant que $q_i \neq \alpha_i$: $\alpha_{i+1} := \dfrac{1}{\alpha_i - q_i}$ et $q_{i+1} := \lfloor \alpha_{i+1} \rfloor$. Il existe alors *aucun* entier k tel que $q_k = \alpha_k$, car cela entraînerait que $\alpha = [q_0, \ldots, q_k]$ et α serait rationnel. La liste *infinie* $[q_0, \ldots, q_k, \ldots]$ est encore appelée *développement en fraction continue* de α et les q_i sont appelés *quotients incomplets* et les α_i *quotients complets* de ce développement. Pour tout indice $i \geqslant 1$, on a $\alpha_i > 1$, et le développement en fraction continue de α est formé de q_0 suivi de celui de α_1. À un décalage près, on peut donc supposer $\alpha > 1$, ce que nous ferons désormais. les formules suivantes sont encore valables :

$$\alpha = [q_0, \ldots, q_i, \alpha_{i+1}] = \begin{bmatrix} u_i & u_{i-1} \\ v_i & v_{i-1} \end{bmatrix} (\alpha_{i+1}) = \frac{u_i\alpha_{i+1} + u_{i-1}}{v_i\alpha_{i+1} + v_{i-1}},$$

Point Méthode

Les programmes **`quotients`** et **`fraction`** de la page 54 sont encore valides, à condition d'autoriser les listes infinies et la non terminaison. *Ce n'est pas absurde :* des programmes comme le système d'exploitation d'un ordinateur ou les systèmes interactifs (interpréteur de CAML, gestion de messagerie ...) sont supposés pouvoir opérer indéfiniment et manipuler des données non bornées : les entrées-sorties. En théorie de la programmation, cette situation est modélisée par la notion de *flot*.

Réciproquement, si l'on part d'une suite $(q_i)_{i\in\mathbb{N}}$ (ou d'une liste infinie) d'éléments de $\mathbb{N}^*$, on peut encore, comme à la section 6.2, définir les suites $(u_i)_{i\in\mathbb{N}}$ et $(v_i)_{i\in\mathbb{N}}$ et les réduites $r_i := \dfrac{u_i}{v_i}\cdot$ Les r_i sont également nommés *convergents* à cause du résultat suivant.

Théorème 57. (i) Les réduites r_i ont une limite réelle $\alpha > 1$. Pour tout indice pair n, on a $r_n < \alpha < r_{n+1}$. Pour tout indice n, on a $|\alpha - r_n| < \dfrac{1}{v_n v_{n+1}}\cdot$
(ii) Le développement en fraction continue de α est $[q_0, \ldots, q_k, \ldots]$.

Démonstration. De l'initialisation $v_0 := 1$, $v_1 := q_1 \geqslant 1$ et de la formule de récurrence $v_{i+1} := q_i v_i + v_{i-1} \geqslant v_i + v_{i-1}$, on déduit par récurrence que $v_i \geqslant F_{i+1}$ (suite de Fibonacci). Il est clair par ailleurs que la suite $(v_i)_{i\in\mathbb{N}}$ est croissante. On introduit la série de terme général $w_i := r_i - r_{i-1}$ (pour $i \geqslant 1$). D'après le théorème 52 de la page 59, $w_i = \dfrac{(-1)^{i-1}}{v_i v_{i-1}}$ qui est de signe alterné, de module décroissant et majoré : $|w_i| \leqslant \dfrac{1}{F_i F_{i+1}}$. Comme $F_i \sim \dfrac{1}{\sqrt{5}}\left(\dfrac{1+\sqrt{5}}{2}\right)^i$, on peut appliquer les résultats généraux sur les séries alternées (ce qui justifie l'encadrement et la majoration de l'erreur). En particulier, $\alpha > r_0 = q_0 \geqslant 1$.

En vertu de l'égalité $[q_0, \ldots, q_k] = q_0 + \dfrac{1}{[q_1, \ldots, q_k]}$, le réel α_1 obtenu en appliquant ce même résultat à la liste $[q_1, \ldots, q_k, \ldots]$ est tel que $\alpha = q_0 + \dfrac{1}{\alpha_1}\cdot$ Comme $\alpha_1 > 1$, la partie entière de α_0 est q_0, qui est donc la tête de son développement en fraction continue. Ce dernier se poursuit par le développement en fraction continue de α_1, qui commence par q_1, etc. ∎

Corollaire 58. Pour tout $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, la suite des réduites r_i converge vers α.

Il y a donc, *pour les irrationnels*, une véritable réciprocité entre les deux algorithmes.

Exemple. Supposons tous les q_i égaux à q. Alors $\alpha = q + \dfrac{1}{\alpha}$, donc $\alpha = \dfrac{q + \sqrt{q^2+4}}{2}$ (l'autre racine étant négative). On obtient ainsi les développements en fraction continue : $1 + \sqrt{2} = [2, 2, 2, 2, \ldots]$, d'où l'on déduit : $\sqrt{2} = [1, 2, 2, 2, \ldots]$;

et $\dfrac{1+\sqrt{5}}{2} = [1,1,1,\ldots]$ (nombre d'or). De même, pour calculer $\alpha := [1,2,1,2,\ldots]$, on écrit $\alpha = [1,2,\alpha] = \dfrac{3\alpha+1}{2\alpha+1}$, d'où $\alpha = \dfrac{1+\sqrt{3}}{2}$.

Exercice 21.
On suppose le développement périodique : $\alpha = [q_0,\ldots,q_k,q_0,\ldots,q_k,\ldots]$.
Comment calculer α ?

Solution. On a $\alpha = [q_0,\ldots,q_k,\alpha] = \dfrac{u_k\alpha+u_{k-1}}{v_k\alpha+v_{k-1}}$, d'où une équation du second degré dont la seule racine positive est α. C'est donc un nombre *quadratique*, c'est-à-dire solution d'une équation du second degré à coefficients rationnels (voir une réciproque dans le livre [MPA1]).

Pour calculer le début du développement en fraction continue d'un irrationnel α, il suffit de connaître ce dernier avec une certaine précision. En effet, supposons donnés α' et α'' dont les développements admettent les mêmes premiers termes $[q_0,\ldots,q_k]$. On peut donc écrire : $\alpha' = [q_0,\ldots,q_k,\alpha'_{k+1}]$ et $\alpha'' = [q_0,\ldots,q_k,\alpha''_{k+1}]$. L'homographie $\begin{bmatrix} u_k & u_{k-1} \\ v_k & v_{k-1} \end{bmatrix}$ étant monotone continue sur $\mathbb{R}_+^*$, l'application $\beta \mapsto [q_0,\ldots,q_k,\beta]$ est une bijection de l'intervalle de bornes $\alpha'_{k+1},\alpha''_{k+1}$ sur l'intervalle de bornes α',α''. Pour tout réel α compris entre α' et α'', le développement en fraction continue commence donc par $[q_0,\ldots,q_k]$.

Exemple. On a $3,14159265358 < \pi < 3,14159265359$ et les commandes Maple :

```
convert(3.14159265358,confrac);
convert(3.14159265359,confrac);
```

nous donnent les développements en fraction continue

$$3{,}14159265358 = [3,7,15,1,292,1,1,1,1,\ldots]$$
$$3{,}14159265359 = [3,7,15,1,292,1,1,1,2,\ldots]$$

donc $\pi = [3,7,15,1,292,1,1,1,\ldots]$. Calculons les premières réduites associées.

- Les u_i sont $3,22,333,355,\ldots$
- Les v_i sont $1,7,106,113,\ldots$
- Les réduites sont $3,\dfrac{22}{7},\dfrac{333}{106},\dfrac{355}{113},\ldots$

Notons que $v_4 = 292v_3 + v_2 > 292v_3$. L'erreur commise en utilisant l'approximation $\pi \simeq \dfrac{355}{113}$ est donc inférieure à $\dfrac{1}{292.113^2}$, ce qui explique la bonne réputation de cette approximation (due au gros quotient incomplet $q_4 = 292$).

Le développement en fraction continue fournit la « meilleure approximation » d'un irrationnel par un rationnel (voir le livre [MPA1]).

6.5 Développement en fraction continue d'une série formelle

Nous avons vu au module « Compléments d'algèbre » du livre [L2] que l'on peut considérer le passage des polynômes et fractions rationnelles aux séries formelles et séries de Laurent comme une complétion, au même titre que le passage de $\mathbb{Q}$ à $\mathbb{R}$. De même que nous avons associé à tout réel un développement en fraction continue *infini* à coefficients entiers (section 6.4), nous allons maintenant associer à toute série de Laurent un développement en fraction continue infini à coefficients polynomiaux.

Nous utiliserons le plongement φ de $K(X)$ dans $K((\xi))$ défini par le développement en série de Laurent *à l'infini* : ainsi, $\varphi(X) = \xi^{-1}$ (module « Compléments d'algèbre »). Bien que l'on puisse sans ambiguïté poser $\xi = X^{-1}$ et donc identifier $K(X)$ à un sous-corps de $K((\xi))$, nous écrirons souvent $\varphi(R)$ pour la série en ξ associée à $R(X)$.

Pour tout $\alpha \in K((\xi))$, on appelle *partie entière de* α et l'on note $\lfloor \alpha \rfloor$ l'unique polynôme $P \in K[X]$ tel que $\omega\big(\alpha - \varphi(P)\big) \geqslant 1$. Ainsi, si $\alpha = \sum\limits_{n \geqslant n_0} a_n \xi^n$, on a $\lfloor \alpha \rfloor = \sum\limits_{n_0 \leqslant n \leqslant 0} a_n X^{-n}$ (c'est bien un polynôme en X).

Pour l'algorithme de développement en fraction continue, nous supposons que $\alpha \in K((\xi))$ n'est pas rationnelle, autrement dit, $\alpha \notin \operatorname{Im} \varphi$. Nous posons $\alpha_0 := \alpha$, puis, pour tout $i \geqslant 0$, $q_i := \lfloor \alpha_i \rfloor$ et $\alpha_{i+1} := \dfrac{1}{\alpha_i - \varphi(q_i)}$ (tous ces termes sont bien définis, car sinon α serait rationnelle). Puisque $\omega\big(\alpha_i - \varphi(q_i)\big) \geqslant 1$, on voit que $\omega(\alpha_i) \leqslant -1$ pour $i \geqslant 1$, donc $\deg q_i \geqslant 1$ pour $i \geqslant 1$. Les q_i forment le *développement en fraction continue de* α.

Donnons-nous réciproquement une famille de polynômes $(q_i)_{i \geqslant 0}$ tels que $\deg q_i \geqslant 1$ pour $i \geqslant 1$. On a introduit les les réduites (ou convergents) $r_i := \dfrac{u_i}{v_i}$ à la section 6.3.

Théorème 59. (i) Les réduites r_i ont une limite $\alpha \in K((\xi))$. On a :

$$\omega(r_i - \alpha) = \deg(v_i\, v_{i+1}) = 2 \deg q_1 + \cdots + 2 \deg q_i + \deg q_{i+1} \geqslant 2i + 1.$$

(ii) Le développement en fraction continue de α est $[q_0, \ldots, q_k, \ldots]$.

Démonstration. De l'initialisation $v_0 := 1$, $v_1 := q_1 \geqslant 1$ et de la formule de récurrence $v_{i+1} := q_i v_i + v_{i-1} \geqslant v_i + v_{i-1}$ (section 6.3), on déduit par récurrence que $\deg u_i = \deg q_0 + \cdots + \deg q_i$ et $\deg v_i = \deg q_1 + \cdots + \deg q_i$. On en déduit :

$$\omega\big(\varphi(r_{i+1}) - \varphi(r_i)\big) = -\deg(r_{i+1} - r_i) = \deg(v_i\, v_{i+1}).$$

L'assertion (i) en découle. La preuve de (ii) est alors entièrement similaire à celle donnée pour le théorème 57 de la page 62. ∎

Corollaire 60. Pour tout $\alpha \in K((\xi)) \setminus \operatorname{Im} \varphi$, la suite des réduites associées au développement en fraction continue de α converge vers α.

Exercice 22.
Soit $\alpha := \sum_{n\geqslant 0} n!\,\xi^{n+1}$ la *série d'Euler*. Calculer r_2 et $\omega\big(\alpha - \varphi(r_2)\big)$.

Solution. On a $q_0 = 0$, puis, successivement :

$$\alpha_1 = \frac{1}{\xi + \xi^2 + 2\xi^3 + 6\xi^4 + \cdots} = \xi^{-1} - 1 - \xi - 3\xi^2 + \cdots \qquad \text{et} \quad q_1 = X - 1,$$

$$\alpha_2 = \frac{1}{-\xi - 3\xi^2 + \cdots} = -\xi^{-1} + 3 + \cdots \qquad \text{et} \quad q_2 = -X + 3.$$

On obtient ainsi la réduite $r_2 = \dfrac{1|}{X-1} - \dfrac{1|}{X-3} = \dfrac{X-3}{X^2 - 4X + 2}$, d'où :

$$\varphi(r_2) = \frac{\xi^{-1} - 3}{\xi^{-2} - 4\xi^{-1} + 2} = \xi\,\frac{1 - 3\xi}{1 - 4\xi + 2\xi^2} = \xi + \xi^2 + 2\xi^3 + 6\xi^4 + 20\xi^5 \cdots .$$

On a donc $\omega\big(\alpha - \varphi(r_2)\big) = 5$, ce qui est conforme au théorème.

On va voir que la réduite r_n est, à plusieurs titres, la *meilleure approximation de α par une fraction rationnelle à degrés donnés*.

Théorème 61. Soit $r := \dfrac{u}{v} \in K(X)$ une fraction rationnelle qui n'est pas une réduite de α. Soit n l'unique entier naturel tel que $\deg v_n \leqslant \deg v < \deg v_{n+1}$. Alors :

$$\omega\,(\alpha - r) \leqslant \deg(v\,v_n) < \omega(\alpha - r_n) \quad \text{et} \quad \omega(\alpha v - u) \leqslant \deg v_n < \omega(\alpha v_n - u_n).$$

Démonstration. Puisque la suite des entiers $\deg v_i$ est strictement croissante de 0 à $+\infty$, un tel n existe bien. Le système $\begin{pmatrix} u_n & u_{n+1} \\ v_n & v_{n+1} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix}$ est de déterminant $(-1)^{n+1}$: il admet donc une solution telle que $a, b \in K[X]$, et l'on a : $b = (-1)^n(uv_n - vu_n)$, d'où $r - r_n = \dfrac{(-1)^n b}{vv_n}\cdot$ Comme r n'est pas une réduite, $b \neq 0$, et l'on a :

$$\omega(r - r_n) = \deg(v\,v_n) - \deg b \leqslant \deg(v\,v_n) < \deg(v_{n+1}\,v_n) = \omega(\alpha - r_n).$$

On en déduit d'abord que $\omega(r - \alpha) = \omega(r - r_n) \leqslant \deg(v\,v_n)$, puis que :

$$\omega(\alpha v - u) \leqslant \deg v_n < \deg v_{n+1} = \omega(\alpha v_n - u_n).$$

Corollaire 62.
Si $r := \dfrac{u}{v} \in K(X)$ est telle que $\omega(\alpha - r) > 2\deg v$, alors r est une réduite.

Exemple. Reprenons l'exemple de la série d'Euler $\alpha := \sum_{n\geqslant 0} n!\,\xi^{n+1}$ (exercice 22).

Notons $f_n := \sum_{k=0}^{n} k!\,\xi^{k+1}$. Les commandes Maple suivantes :

```
f3 := sum(k! x^(-k-1),k=0..4): f4 := sum(k! x^(-k-1),k=0..4):
convert(f3,confrac,x); convert(f4,confrac,x);
```

donnent, par exemple, les développements en fraction continue (généralisée) de f_3 et f_4 :

$$\frac{1|}{|x-1} - \frac{1|}{|x-3} + \frac{20|}{|x+\frac{23}{5}} - \frac{81|}{|25(x-\frac{3}{5})}$$

et :

$$\frac{1|}{|x-1} - \frac{1|}{|x-3} - \frac{4|}{|x+25} + \frac{531|}{|x-\frac{1203}{59}} - \frac{1536|}{|3481(x-\frac{36}{59})}.$$

Du théorème, on déduit alors que les deux premiers quotients partiels sont corrects :

$$\alpha = \frac{1|}{|x-1} - \frac{1|}{|x-3} + \cdots .$$

En expérimentant ainsi avec Maple jusqu'à f_{11} et f_{12}, on voit apparaître un schéma :

$$\alpha = \frac{1|}{|x-1} - \frac{1|}{|x-3} - \frac{4|}{|x-5} - \frac{9|}{|x-7} - \frac{16|}{|x-9} - \frac{25|}{|x-11} + \cdots .$$

On peut donc conjecturer la formule suivante :

$$\alpha = \frac{1|}{|A_1X+B_1} - \frac{C_2|}{|A_2X+B_2} - \frac{C_3|}{|A_3X+B_3} - \cdots ,$$

où $C_i = (i-1)^2$ pour $i \geqslant 2$, et $A_i = 1$ et $B_i = -(2i-1)$ pour $i \geqslant 1$.

Exercice 23.

En admettant la formule précédente, expliciter la relation de récurrence des dénominateurs v_n des réduites.

Solution. D'après la proposition 54 de la page 60, on a $v_0 := 1$, $v_1 := X - 1$ et, pour $k \geqslant 2$: $v_k = \big(X - (2k-1)\big)v_{k-1} - (k-1)^2 v_{k-2}$. Les $(-1)^k v_k$ sont les *polynômes de Laguerre*, étudiés dans le module III.4, où l'on verra également que la formule conjecturée est correcte.

Exercice type corrigé

I.1.0 On se propose de démontrer que l'*équation diophantienne* $x^3 - y^2 = 2$ admet pour seules solutions entières $(x,y) = (+3,+5)$ et $(x,y) = (+3,-5)$. La démonstration repose sur l'arithmétique de l'anneau $A := \mathbb{Z}[\sqrt{-2}]$, sous-anneau de $\mathbb{C}$ engendré par $\sqrt{-2}$, où l'on convient que $\sqrt{-2} := \mathrm{i}\sqrt{2}$. (Toutes les questions ne sont pas nécessaires à la résolution de l'équation !)

(i) Vérifier que $A = \{a + b\sqrt{-2} \mid a, b \in \mathbb{Z}\}$ et que A est isomorphe à $\mathbb{Z}[X]/ < X^2 + 2 >$.

(ii) Vérifier que l'application $z \mapsto N(z) := z\overline{z}$ envoie A dans $\mathbb{N}$, que $N(zz') = N(z)N(z')$ et que N définit un stathme euclidien sur A.

(iii) Montrer que $A^* = \{+1, -1\}$.

(iv) Montrer que $\pi := \sqrt{-2}$ est premier dans A. Déterminer l'idéal premier $<\pi> \cap \mathbb{Z}$ de $\mathbb{Z}$. Comment se factorise 2 dans A ?

(v) Soit p un nombre premier impair. Montrer qu'il est inerte dans A si, et seulement si, $\left(\dfrac{-2}{p}\right) = -1$. Déterminer ces nombres premiers.

(vi) Déterminer les nombres premiers impairs qui se décomposent dans A et montrer qu'ils sont représentables sous la forme $p = x^2 + 2y^2$.
(vii) On suppose maintenant que $x^3 - y^2 = 2$, $x, y \in \mathbb{Z}$. Vérifier que y est impair et en déduire que $y \notin <\pi>$.
(viii) Soit $z := y + \sqrt{-2}$. Que vaut $N(z)$? Montrer que z et $\overline{z}$ sont premiers entre eux et en déduire que ce sont des cubes dans A.
(ix) En écrivant $y + \sqrt{-2} = (a + b\sqrt{-2})^3$, achever de résoudre l'équation avec de l'arithmétique élémentaire.
Solution. (i) Comme $(\sqrt{-2})^2 \in \mathbb{Z} + \mathbb{Z}\sqrt{-2}$, ce groupe, qui contient 1, est stable pour la multiplication ; c'est donc un sous-anneau de $\mathbb{C}$. Comme c'est le plus petit sous-groupe de $\mathbb{C}$ contenant 1 et $\sqrt{-2}$, c'est bien le sous-anneau de $\mathbb{C}$ engendré par $\sqrt{-2}$.
Le morphisme $P \mapsto P(\sqrt{-2})$ de $\mathbb{Z}[X]$ dans $\mathbb{C}$ a pour image A et pour noyau un idéal I contenant $X^2 + 2$. Pour tout $P \in I$, on peut effectuer la division euclidienne $P = (X^2 + 2)Q + R$ dans $\mathbb{Z}[X]$ car $X^2 + 2$ est un polynôme unitaire. Puisque P est dans le noyau, on a $P(\sqrt{-2}) = 0$, donc $R(\sqrt{-2}) = 0$. Comme $R \in \mathbb{Z}[X]$ est de degré < 2, cela entraîne $R = 0$, donc $P \in <X^2 + 2>$ et le noyau de notre morphisme est $I = <X^2 + 2>$. Le premier théorème d'isomorphisme fournit l'isomorphie $A \simeq \mathbb{Z}[X]/<X^2 + 2>$.
(ii) On a $N(a + b\sqrt{-2}) = a^2 + 2b^2 \in \mathbb{N}$.
L'égalité $N(zz') = N(z)N(z')$ est évidemment valable pour tous $z, z' \in \mathbb{C}$.
Soient $u, v \in A$, $v \neq 0$. Alors $w := u/v = (u\overline{v})/N(v)$ est de la forme $x + y\sqrt{-2}$, $x, y \in \mathbb{Q}$ (puisque $u\overline{v} \in A$ et $N(v) \in \mathbb{N}^*$). Soient $a, b \in \mathbb{Z}$ tels que $|x - a|, |y - b| \leqslant 1/2$ et soit $q := a + b\sqrt{-2}$. Alors $N(w - q) \leqslant 1/4 + 2/4 < 1$, donc $N(u - qv) < N(v)$. Posant $r := u - qv$, on a bien une division euclidienne $u = qv + r$.
(iii) On établit d'abord que $u \in A^* \Leftrightarrow N(u) = 1$. En effet, si $u\overline{u} = N(u) = 1$, alors $\overline{u}$ est l'inverse de u dans A ; et si $uv = 1$, $u, v \in A$, alors $1 = N(1) = N(u)N(v)$, d'où $N(u) = 1$ puisque $N(u), N(v) \in \mathbb{N}$.
Ensuite on écrit $u = a + b\sqrt{-2}$, et l'on voit que $N(u) = 1$, *i.e.* $a^2 + 2b^2 = 1$ a pour seules solutions entières $(a, b) = (+1, 0)$ et $(a, b) = (-1, 0)$, donc $u = \pm 1$.
(iv) Puisque A est euclidien, il est principal et factoriel ; donc « premier » et « irréductible » sont des propriétés équivalentes. Si $\pi = uv$, alors $N(u)N(v) = N(\pi) = 2$, donc $N(u) = 1$ ou $N(v) = 1$. Donc π est irréductible et premier.
Il est clair que $2 = \pi\overline{\pi} \in <\pi> \cap \mathbb{Z}$, d'où l'inclusion $2\mathbb{Z} \subset <\pi> \cap \mathbb{Z}$. Réciproquement, si $a \in <\pi> \cap \mathbb{Z}$, on peut écrire $a = \pi u$, d'où $a^2 = N(a) = N(\pi)N(u)$, donc $2 = N(\pi)$ divise a^2, donc a. On a donc $<\pi> \cap \mathbb{Z} = 2\mathbb{Z}$.
Puisque $2 = \pi\overline{\pi} = \pi(-\pi)$, 2 est ramifié : $2 = -\pi^2$.
(v) Puisque $A \simeq \mathbb{Z}[X]/<X^2 + 2>$, on a $A/pA \simeq \mathbb{Z}[X]/< p, X^2 + 2> \simeq \mathbb{F}_p[X]/<X^2 + 2>$ qui est intègre si, et seulement si, -2 n'est pas un carré dans $\mathbb{F}_p$, c'est-à-dire si, et seulement si, $\left(\dfrac{-2}{p}\right) = -1$.
Or $\left(\dfrac{-2}{p}\right) = \left(\dfrac{-1}{p}\right)\left(\dfrac{2}{p}\right) = (-1)^{(p-1)/2}(-1)^{(p^2-1)/8}$ dépend seulement de la classe de p modulo 8. les classes qui conviennent sont $p \equiv 5 \pmod 8$ et $p \equiv 7 \pmod 8$.
(vi) Si au contraire $\left(\dfrac{-2}{p}\right) = +1$, l'anneau $A/pA \simeq \mathbb{F}_p[X]/<X^2 + 2>$ n'est pas intègre, p n'est pas premier dans A, donc il est réductible. L'égalité $p = uv$, $u, v \notin A^*$, en-

traîne $p^2 = N(p) = N(u)N(v)$ donc $N(u) = N(v) = p$. Donc d'une part p se décompose, d'autre part p est représentable sous la forme $x^2 + 2y^2$. Les premiers impairs de ce type sont ceux tels que $p \equiv 1 \pmod 8$ et $p \equiv 3 \pmod 8$.

À titre d'exemple, les premiers sont $3 = 1^2+2.1^2$, $11 = 3^2+2.1^2$, $17 = 3^2+2.2^2$, $19 = 1^2+2.3^2$, $41 = 3^2 + 2.4^2$, $43 = 5^2 + 2.3^2$.

(vii) Si y était pair, x^3 serait congru à 2 modulo 4 ; mais 2 n'est pas résidu cubique modulo 4. Donc $y \notin 2\mathbb{Z}$.

Comme $y \in \mathbb{Z}$, d'après la question (iv), $y \notin < \pi >$.

(viii) Si un élément premier u de A divisait z et $\overline{z}$, il diviserait leur différence $2\sqrt{-2} = -\pi^3$, donc il serait équivalent à π. Mais π ne divise pas y, donc pas non plus $z = y + \pi$.

On a $z\overline{z} = x^3$. En tout premier u de A, on a $v_u(z) = 0$ ou $v_u(\overline{z}) = 0$ (puisqu'ils sont premiers entre eux) et $v_u(z) + v_u(\overline{z}) = 3v_u(x)$. Rappelons que v_u désigne la valuation associée à l'élément premier u (paragraphe 1.2.2). Dans tous les cas, $v_u(z)$ est multiple de 3. Donc z est équivalent à un cube (on reconnaît là un cas particulier de l'exercice 4 de la page 10). Mais les unités de A sont $+1$ et -1 qui sont eux mêmes des cubes. Donc z est un cube dans A.

(ix) On peut donc écrire $y + \sqrt{-2} = (a + b\sqrt{-2})^3$, $a, b \in \mathbb{Z}$, ce qui équivaut à $y = a^3 - 6ab^2$ et $1 = 3a^2b - 2b^3$. De cette dernière égalité on déduit que b divise 1, donc $b = \pm 1$. L'égalité $b = -1$ donnerait $1 = -3a^2 + 2$, qui est impossible. On a donc $b = 1$, donc $1 = 3a^2 - 2$, donc $a = \pm 1$ et $y = \pm 5$. Chacune de ces valeurs donne $x = 3$.

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 Effectuer dans $\mathbb{Z}[\mathrm{i}]$, la division euclidienne de 5 par $1 + \mathrm{i}$.

I.1.2 (i) Démontrer que N est un stathme euclidien sur l'anneau $\mathbb{Z}[j]$.

(ii) Sur l'anneau $\mathbb{Z}[\sqrt{2}]$, on pose $g(a + b\sqrt{2}) = \left|a^2 - 2b^2\right|$. Vérifier que $g(xx') = g(x)g(x')$, puis que g est un stathme euclidien.

I.1.3 Soit g un stathme euclidien sur A. Alors $g(x)$ est minimum si, et seulement si, x est inversible. De plus, si $x, y \neq 0$, $x \mid y$ et $g(x) = g(y)$, alors x et y sont associés.

I.1.4 $**$ (i) Pour toute matrice rectangulaire M sur un anneau A, notons $F_k(M)$ l'idéal engendré par les mineurs d'ordre k. Démontrer que, si M, N et P sont quelconques (de formats compatibles), $F_k(MN)$ et $F_k(PM)$ sont inclus dans $F_k(M)$.

(ii) En déduire que, dans la proposition 3 de la page 6, les d_i qui apparaissent dans la matrice (membre gauche de l'égalité) sont uniques à facteur inversible près.

I.1.5 À l'aide de l'algorithme de la proposition 3 de la page 6, discuter la résolubilité dans $\mathbb{Z}$ du système : $\begin{cases} 6x + 8y + 4z + 20t = a, \\ 12x + 12y + 18z + 30t = b, \\ 18x + 4y + 4z + 10t = c. \end{cases}$

I.1.6 $*$ Compléter l'algorithme d'Euclide (proposition 2 de la page 5) pour qu'il fournisse des coefficients de Bézout.

I.1.7 Pour que l'anneau $A[X]$ soit principal, il faut, et il suffit, que A soit un corps.
À quelle condition $A[X]$ est-il un corps ?

I.1.8 Pour que les idéaux $<a>$ et $<b>$ soient étrangers, il faut, et il suffit, qu'il existe $u, v \in A$ tels que $ua + vb = 1$. Cela implique que a et b sont étrangers. La réciproque est fausse. Que dire du cas d'un anneau principal ?

I.1.9 $*$ Soit $A = \mathcal{C}(I, \mathbb{R})$ l'anneau des fonctions continues de l'intervalle $I \subset \mathbb{R}$ dans $\mathbb{R}$. Démontrer que $f, g \in A$ ont un diviseur commun non trivial si, et seulement si, elles ont un zéro commun. Démontrer que, si f et g n'ont pas de diviseur commun, $<f, g> = A$. L'anneau A est-il principal ?

I.1.10 Montrer que toute suite croissante d'idéaux d'un anneau principal est stationnaire.

I.1.11 $**$ Soit A un anneau principal. Démontrer que tout idéal de $A[X]$ est de type fini, *i.e.* engendré par un nombre fini d'éléments.

I.1.12 Trouver deux éléments x, y d'un anneau commutatif non intègre A tels que $x \mid y$ et $y \mid x$, mais il n'existe pas de $u \in A^*$ tel que $y = ux$.

I.1.13 $*$ Dans un anneau factoriel, les idéaux premiers non nuls minimaux sont principaux.

I.1.14 $*$ (i) Pour tout nombre premier p, calculer $v_p(n!)$.
(ii) En déduire une nouvelle démonstration du fait que $\dfrac{n!}{m!(n-m)!}$ est un entier pour $0 \leqslant m \leqslant n$.

I.1.15 $**$ Démontrer qu'un anneau factoriel A est *intégralement clos*, autrement dit : si $x \in K$ (corps des fractions de A) est tel que $x^n + a_1 x^{n-1} + \cdots + a_n = 0$, $a_1, \ldots, a_n \in A$, alors $x \in A$.

I.1.16 $*$
(i) Soient A un anneau factoriel et $p \in A$ un irréductible. Soit $P := a_0 X^n + \cdots + a_n \in A[X]$ tel que p divise $a_1, \ldots a_n$ mais ne divise pas a_0 et que p^2 ne divise pas a_n.
Démontrer que P est irréductible dans $\mathbb{K}[X]$, où $\mathbb{K}$ désigne le corps des fractions de A (critère d'Eisenstein).
(ii) Soit $p \in \mathbb{N}$ un nombre premier.
Démontrer que le polynôme $P := X^{p-1} + \cdots + 1$ est irréductible dans $\mathbb{Z}[X]$.

I.1.17 Soient A un anneau factoriel de corps des fractions K et $F, G \in K[X]$ unitaires tels que $FG \in A[X]$. Démontrer que $F, G \in A[X]$.

I.1.18 $*$ (i) Décrire tous les irréductibles de $\mathbb{Z}[X]$.
(ii) Soit K un corps. Décrire tous les irréductibles de $K[X, Y]$.

I.1.19 $**$ (i) On reprend les notations de l'exercice type corrigé du module « Compléments d'algèbre » du premier tome [L1]de ce cours : $\mathcal{P}$ est un idéal premier de $\mathbb{Z}[X]$ et l'on suppose que $\mathcal{P} \cap \mathbb{Z} = \{0\}$. Montrer que $\mathcal{P}\mathbb{Q}[X] = \left\{ \frac{1}{d}P \mid d \in \mathbb{N}^* \quad \text{et} \quad P \in \mathcal{P} \right\}$ et que c'est un idéal premier de $\mathbb{Q}[X]$. Montrer que $\mathcal{P}\mathbb{Q}[X] \cap \mathbb{Z}[X] = \mathcal{P}$.
(ii) Montrer que $\mathcal{P}$ est principal, et qu'il est soit nul soit engendré par un polynôme primitif.

I.1.20 $**$ On suppose le corps K algébriquement clos.
Déterminer tous les idéaux premiers de $K[X, Y]$.

Exercices sur la section 2

I.1.21 Vérifier que, m étant fixé, l'équation $\varphi(n) = m$ n'a qu'un nombre fini de solutions.

I.1.22 (i) À quelle condition le groupe produit $G_1 \times G_2$ est-il cyclique ?
(ii) À quelle condition le groupe $(\mathbb{Z}/n\mathbb{Z})^*$ est-il cyclique ?

I.1.23 $*$
(i) Montrer que si $f : \mathbb{N}^* \to \mathbb{C}$ est multiplicative, l'application $Tf : n \mapsto \sum_{d|n} f(d)$ l'est aussi.
(ii) Calculer $T\varphi$ et en déduire une nouvelle démonstration de la proposition 13 de la page 14.

I.1.24 Donner un exemple de sous-groupe fini de K^*, où K est un corps non commutatif, qui ne soit pas cyclique.

I.1.25 Soient $a, b \in G$ (groupe commutatif) d'ordres respectifs d, e premiers entre eux.
Montrer que ab est d'ordre de.

I.1.26 Combien l'équation $x^2 = 1$ a-t-elle de solutions dans $(\mathbb{Z}/n\mathbb{Z})^*$?

I.1.27 (i) Préciser la structure de $(\mathbb{Z}/2^r\mathbb{Z})^*$ pour $r = 1, 2, 3$.
(ii) Vérifier que $(\mathbb{Z}/2^r\mathbb{Z})^*$ n'est pas cyclique pour $r \geqslant 3$.

I.1.28 Soient p premier impair et $r \geqslant 2$. Quel est le nombre de générateurs du groupe cyclique $(\mathbb{Z}/p^r\mathbb{Z})^*$?

I.1.29 (i) On appelle *exposant* d'un groupe G le plus petit $m > 0$ tel que $x^m = 1$ pour tout $x \in G$, *i.e.* le ppcm des ordres des éléments de G. Quel est l'exposant d'un groupe cyclique ? D'un groupe produit ?
(ii) Calculer l'exposant de $(\mathbb{Z}/n\mathbb{Z})^*$ pour n impair.

I.1.30 $**$ Soient p premier impair et $a \in \mathbb{Z}$ tel que la classe de a soit un générateur de $(\mathbb{Z}/p^2\mathbb{Z})^*$. Alors cela reste vrai dans tout $(\mathbb{Z}/p^r\mathbb{Z})^*$.

I.1.31 Soient p, q deux nombres premiers impairs distincts, $n = pq$, et $f \in \mathbb{N}$ tel que $f \equiv 1 \pmod{\varphi(n)}$. Démontrer que pour tout $M \in \mathbb{Z}$ on a $M^f \equiv M \pmod{n}$.
Ce résultat justifie donc le cryptage RSA d'un message secret $M \in \{0, \ldots, n-1\}$ *arbitraire* (non nécessairement premier à n) tel que nous l'avons décrit au 2.4.2, pourvu que la clé n soit de la forme $n = pq$.

Exercices sur la section 3

I.1.32 $*$ Soit K un corps commutatif fini de caractéristique $p > 0$ (voir le livre [L1]).
(i) Vérifier que l'application $m \mapsto m.1_K$ de $\mathbb{Z}$ dans K induit un isomorphisme du corps $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z}$ sur le plus petit-sous-corps de K ; on identifiera $\mathbb{F}_p$ à ce sous-corps.
(ii) Montrer que la loi de groupe de K et la loi $\mathbb{F}_p \times K \to K$ induite par la multiplication $K \times K \to K$ font de K un $\mathbb{F}_p$-espace vectoriel. En déduire que $q := \operatorname{card} K$ est de la forme p^r.
(iii) Montrer que l'application $F : x \mapsto x^p$ est un automorphisme du corps K (*automorphisme de Frobenius*) et que $F(x) = x$ équivaut à $x \in \mathbb{F}_p$.
(iv) Montrer que l'ordre de F en tant qu'élément du groupe des automorphismes du corps K est r.

I.1.33 $*$ Soit K un corps commutatif fini. Pour tout $d \in \mathbb{N}$, calculer $\sum\limits_{x \in K} x^d$.

I.1.34 $*$ Retrouver le corollaire 22 de la page 23 en invoquant le fait que $\mathbb{F}_p^*$ est cyclique.

I.1.35 $*$ Donner une condition nécessaire et suffisante pour que le nombre premier p reste premier dans $\mathbb{Z}[j]$, où $j := e^{2\mathrm{i}\pi/3} = \dfrac{-1+\mathrm{i}\sqrt{3}}{2}$.

I.1.36 $*$ Soient p un nombre premier impair, $r \geqslant 2$ un entier et $a \in \mathbb{Z}$. Pour que a soit résidu quadratique modulo p^r, il faut, et il suffit, que $\left(\dfrac{a}{p}\right) = 1$.

I.1.37 Soit $p = 2k+1$, k impair, un nombre premier. Montrer, que pour tout $a \in \mathbb{Z} \setminus p\mathbb{Z}$, soit a soit $-a$ est un résidu quadratique modulo p et calculer b tel que $b^2 \equiv \pm a \pmod{p}$.

I.1.38 $**$ Soit $d := \varepsilon p_1 \cdots p_k$, où $\varepsilon = \pm 1$ et les p_i sont des premiers impairs distincts. On va établir une condition *nécessaire* pour qu'un nombre premier impair p soit représentable sous la forme $x^2 + dy^2$, $x, y \in \mathbb{Z}$.
(i) Vérifier qu'alors p ne divise pas y et que $\left(\dfrac{-d}{p}\right) = 1$.
(ii) En déduire que $\prod\limits_{i=1}^{k} \left(\dfrac{p}{p_i}\right) (-1)^{\frac{p-1}{2} \sum \frac{p_i - 1}{2}} = (-\varepsilon)^{\frac{p-1}{2}}$.
(iii) Appliquer à $d = -15$.

Exercices sur la section 4

I.1.39 $*$ Déterminer les corps résiduels $\mathbb{Z}[\mathrm{i}]/\pi$ pour tous les irréductibles π de $\mathbb{Z}[\mathrm{i}]$.

I.1.40 $**$ (i) Soit $j := e^{2\mathrm{i}\pi/3} = \dfrac{-1+\mathrm{i}\sqrt{3}}{2}$.
Déterminer les unités et les éléments irréductibles de l'anneau $\mathbb{Z}[j]$.
(ii) En déduire une condition nécessaire et suffisante pour qu'un nombre premier p puisse être représenté sous la forme $p = x^2 + xy + y^2$, $x, y \in \mathbb{Z}$.

I.1.41 $**$ Pour $n \in \mathbb{N}^*$, on note $r_2(n)$ le nombre de couples $(a,b) \in \mathbb{Z}^2$ tels que $a^2+b^2=n$.
(i) Démontrer que $f(n) := r_2(n)/4$ est une fonction multiplicative, *i.e.* :

$$n \wedge n' = 1 \Longrightarrow f(nn') = f(n)f(n').$$

(ii) Calculer $r_2(n)$ lorsque n est primaire (puissance d'un nombre premier) et en déduire une formule générale.

I.1.42 $**$ Démontrer que $r_2(n)$ (voir exercice I.1.41) est égal à $4\sum\left(\frac{-1}{m}\right)$, où m parcourt l'ensemble des diviseurs impairs de n.

I.1.43 Appliquer l'algorithme de Cornacchia à 89.

I.1.44 Décomposer 79 en somme de quatre carrés par la méthode du texte.

Exercices sur la section 5

I.1.45 (i) On note $F_m := 2^{2^m}+1$ les nombres de Fermat. Démontrer l'égalité :

$$F_0 \cdots F_{m-1} = F_m - 2,$$

et en déduire que les nombres de Fermat sont premiers entre eux deux à deux.
(ii) Déduire de la question précédente une nouvelle preuve qu'il y a une infinité de nombres premiers.

I.1.46 Soit p un facteur premier de $n!+1$. Montrer que $p > n$ et en déduire une nouvelle preuve qu'il y a une infinité de nombres premiers.

I.1.47 $*$ En améliorant la démonstration du théorème 38 de la page 38, obtenir la majoration dans l'encadrement de Tchebychef.

I.1.48 $*$ Du théorème des nombres premiers, déduire que $p_n \sim n \ln n$.

I.1.49 $*$ Calculer la valeur des expressions suivantes :

$$-1+\sum_{j=1}^{n}\left\lfloor \cos^2 \pi \frac{(j-1)!+1}{j}\right\rfloor \qquad \text{et} \qquad \sum_{j=2}^{n}\left\lfloor \frac{(j-1)!+1}{j} - \left\lfloor \frac{(j-1)!}{j}\right\rfloor\right\rfloor.$$

I.1.50 Estimer la taille des nombres de Fermat et de Mersenne.

I.1.51 Comment peut-on améliorer le crible d'Ératosthène en utilisant un tableau d'entiers ?

I.1.52 $*$ Rendre plus rigoureux le raisonnement probabiliste justifiant le test de Solovay-Strassen.

I.1.53 $*$ Proposer un algorithme de calcul rapide du symbole de Jacobi.

I.1.54 $*$ Le lemme suivant fait partie des bases de l'algorithme AKS.
Montrer que n est premier si, et seulement si, $X^n - 1 \equiv (X-1)^n \pmod{n\mathbb{Z}[X]}$.

I.1.55 Certifier la primalité de 659.

I.1.56 Vérifier que les entiers 561, 1105, 1729 et 41041 sont des nombres de Carmichael.

I.1.57 $*$ Trouver une variante du test de Pépin avec 5 au lieu de 3.

Exercices sur la section 6

I.1.58 $*$ Calculer $[q_0, \ldots, q_k]$ lorsque tous les q_i sont égaux à $2\mathrm{i}X$.

I.1.59 (i) Soit $a \in \mathbb{C}$. Calculer les réduites et discuter la convergence de $[a, a, a, \ldots]$.
(ii) Même question pour $\dfrac{a|}{|1} + \dfrac{a|}{|1} + \cdots$.

I.1.60 $*$ Démontrer que si tous les q_i sont réels et minorés par $\varepsilon > 0$, alors les réduites de $[q_0, q_1, \ldots]$ convergent.

I.1.61 Démontrer la propriété suivante des réduites : $r_{i+1} - r_{i-1} = \dfrac{(-1)^{i+1}\, q_{i+1}}{v_{i+1}\, v_{i-1}}$.

I.1.62 Calculer $r_i - r_{i-1}$ dans le cas des réduites généralisées de la proposition 54 de la page 60.

I.1.63 $*$ On définit les *polynômes continuants* en les indéterminées X_i à coefficients dans $\mathbb{Q}$ (cas entier), respectivement dans K (cas polynomial) en posant $U_0 := X_0$, $U_1 := X_0X_1 + 1$, $V_0 := 1$, $V_1 := X_1$ et, pour $i \geqslant 2$, $U_i := X_iU_{i-1} + U_{i-2}$ et $V_i := X_iV_{i-1} + V_{i-2}$. Calculer R_i et $R_i - R_{i-1}$.

I.1.64 Montrer que, parmi deux réduites consécutives du développement en fraction continue d'un irrationnel α, l'une des deux au moins vérifie $|\alpha - r_n| \leqslant \dfrac{1}{2v_n^2}$.

I.1.65 Utiliser la monotonie des homographies pour donner une autre preuve de l'encadrement de α dans le théorème 57 de la page 62.

I.1.66 Dans l'exemple de la page 62, calculer les réduites.

I.1.67 (i) Programmer le développement en fraction continue d'un irrationnel et la fonction réciproque.
(ii) Programmer la fonction **`quotients`** avec un argument qui indique le nombre de quotients incomplets à produire et la fonction **`fraction`** avec un argument qui indique la précision souhaitée.

I.1.68 Calculer, à l'aide de Maple, le début du développement en fraction continue de la série exponentielle et tracer le graphe pour apprécier la qualité de l'approximation.

Compléments sur les formes bilinéaires alternées et sur les formes quadratiques

I.2

Le but de ce module est de compléter le module *Algèbre bilinéaire* du livre [L2]. Nous nous intéresserons d'abord à la *classification des formes bilinéaires alternées* sur un corps K. C'est un problème important, même s'il n'est pas très populaire dans les enseignements français au niveau de la licence... C'est le premier pas vers l'étude des *espaces symplectiques*. Ceux-ci sont aujourd'hui un outil indispensable en mathématiques, mécanique et physique. Le cas bilinéaire alterné est plus simple que le cas quadratique. De plus, son étude est une bonne préparation à celle de la théorie de Witt dans le cas quadratique.
Nous reviendrons ensuite sur le problème de la classification des formes quadratiques sur un corps K. Dans le livre [L2], nous l'avons résolu pour $K = \mathbb{C}$ et $K = \mathbb{R}$. Classifier revient à trouver « suffisamment d'invariants ». Nous connaissons déjà *le rang* et, dans le cas réel, *la signature*. Nous en définirons de nouveaux : *le discriminant* et *l'invariant de Witt*. Le discriminant nous permettra de classifier les formes quadratiques sur un corps fini $K = F_q$, en utilisant les résultats du module I.1 sur les carrés de F_q. L'invariant de Witt apparaît dans la théorie de Witt. Celle-ci est, *grosso modo*, ce que l'on peut dire des formes quadratiques sans prendre en compte les propriétés particulières du corps de base K.

Des méthodes voisines de celles que nous allons présenter permettent de traiter le cas des *formes sesquilinéaires*[1] sur un corps K, muni d'un automorphisme, qui généralisent le cas des formes sesquilinéaires sur $\mathbb{C}$, muni de l'automorphisme de conjugaison, que nous avons étudié dans le module I.IV du livre [L2]. Nous ne le ferons pas. C'est un peu plus technique et, fondamentalement, c'est une variation, *mutatis mutandis*, à partir des cas symplectique, orthogonaux et hermitiens.

Si $K = \mathbb{Q}$, le problème de classification des formes quadratiques est *beaucoup plus difficile*, ce qui peut surprendre le mathématicien débutant... Nous dégagerons les points essentiels en admettant des résultats difficiles. Le lecteur intéressé pourra se reporter au livre [Serre] pour les preuves manquantes. L'idée de la classification est de se ramener aux classifications sur $\mathbb{R}$ et sur les corps p-adiques $\mathbb{Q}_p$, définis dans le module III.3. Pour le cas p-adique, il faut introduire de nouveaux invariants, les *invariants de Hasse*, qui remplacent la signature du cas réel.

[1] Le lecteur intéressé pourra consulter le livre [Perrin] ou la monographie, plus difficile, [D3].

Les techniques introduites dans ce module permettent de prolonger l'étude des *groupes classiques* commencée dans le module I.6 du livre [L2]. Nous ne donnerons que quelques résultats de base, certains sous forme d'exercice. Le lecteur intéressé par ce sujet riche et important (les groupes classiques jouent un grand rôle en algèbre, théorie des nombres, analyse, théorie des équations différentielles, systèmes dynamiques, mécanique, physique théorique et géométrie, dans l'esprit du *Programme d'Erlangen* de F. Klein) pourra consulter le livre [Perrin] et, pour aller plus loin, la monographie (plus difficile) de Jean Dieudonné [D3].

Rappelons que, selon une terminologie introduite par Herman Weyl[2], les groupes classiques sont le groupe général linéaire à n variables $GL_n(K)$ et ses sous-groupes laissant invariante une forme quadratique (groupe orthogonal), une forme sesquilinéaire (groupe unitaire) ou une forme alternée (groupe symplectique). Les cas où $K = \mathbb{R}$, $K = \mathbb{C}$ ou $K = F_q$ sont particulièrement intéressants. Notons que le groupe $GL_n(F_q)$ est *fini*, donc tous les groupes classiques sur un corps fini sont des *groupes finis*. En particulier des familles de groupes finis *simples* apparaissent ainsi.

Dans tout le module on désigne par K un corps commutatif. Contrairement à ce qui est fait dans la très grande majorité des présentations élémentaires, nous ne « mettrons pas sous le tapis » le cas de la caractéristique 2 (nous expliquerons ce choix plus loin). On note en général E un K-espace vectoriel. On le supposera toujours *de dimension finie*. Pour les notations, définitions et résultats que nous utiliserons le lecteur se reportera au module I.3 du livre [L2], que le présent module complète. Nous ferons quelques rappels pour faciliter la lecture. Ils ne remplacent pas la (re)lecture du module I.3 du livre [L2].

1 Compléments sur les formes bilinéaires alternées

Dans tout ce paragraphe K est un corps commutatif et E est un K-espace vectoriel de dimension finie. Le plus souvent, cette dimension est *paire* et elle est en général notée $2n$. Contrairement à ce que nous avons fait dans [L1] I.3, 2, nous n'excluons pas la caractéristique 2. Nous verrons dans la partie 2 que, en caractéristique 2, les formes bilinéaires alternées sont un outil fondamental pour l'étude des formes quadratiques. En caractéristique 2, il ne faut pas confondre[3] *alternée* et *antisymétrique*. C'est essentiellement la seule « difficulté » nouvelle quand on n'écarte pas la caractéristique 2.

Nous rappelons qu'une forme bilinéaire φ sur un K-espace vectoriel E est dite *alternée* si, pour tout $x \in E$, $\varphi(x,x) = 0$. Elle dite *antisymétrique* si, pour tous $x, y \in E$, $\varphi(y,x) = -\varphi(x,y)$. En caractéristique différente de 2 les deux notions sont équivalentes. En caractéristique 2 une forme alternée est antisymétrique mais une forme antisymétrique n'est pas nécessairement alternée.

[2] The classical groups 1939.

[3] Le lecteur prendra garde au fait que, pour cette raison, certains résultats de [L1] I.3, 2 ne sont pas vrais dans ce cas. Par exemple la proposition 27 de 2.4.1.

Soit φ une forme alternée sur E. Un sous-espace $F \subset E$ tel que, pour tous $x, y \in F$, $\varphi(x, y) = 0$ est dit[4] *isotrope*. Pour tout $x \in E$, on a $\varphi(x, x) = 0$, puisque φ est *alternée*. Ainsi, toutes les droites de E sont isotropes. On dit que $x, y \in E$ sont *φ-orthogonaux* (ou plus simplement orthogonaux) si $\varphi(x, y) = 0$. On note alors $x \perp y$. Si $F \subset E$ est un sous-espace, son *orthogonal* est le sous-espace $F^\perp := \{x \in E \mid x \perp F\}$ ($x \perp F$ voulant dire $x \perp y$ pour tout $y \in F$). Par définition le *noyau* $\operatorname{Ker}\varphi$ de φ est le sous-espace $E^\perp$. Une forme bilinéaire alternée φ est dite *non dégénérée* si $\operatorname{Ker}\varphi = E^\perp = (0)$. Une *forme bilinéaire alternée non dégénérée* φ sur E est dite aussi *symplectique* et l'on dit que (E, φ) est un *espace symplectique*.

Le *noyau* de φ est, par définition, le sous-espace $\operatorname{Ker}\varphi := E^\perp \subset E$. Le rang de φ est la dimension de $\widetilde{E} := E/\operatorname{Ker}\varphi$, c'est-à-dire $\dim E - \dim \operatorname{Ker}\Phi$; φ est non dégénérée si, et seulement si, $\operatorname{Ker}\varphi = (0)$, c'est-à-dire si, et seulement si, son rang est égal à $\dim E$.

La forme bilinéaire alternée φ induit une forme bilinéaire alternée *non dégénérée* $\widetilde{\varphi}$ sur $\widetilde{E}$. L'étude de φ est équivalente à celle de $\widetilde{\varphi}$. Si F est un sous-espace supplémentaire de $\operatorname{Ker}\varphi$ dans E, alors c'est un *supplémentaire orthogonal* : $F \perp \operatorname{Ker}\varphi$. Le passage au quotient $E/\operatorname{Ker}\varphi$ induit un isomorphime $\pi : F \to E/\operatorname{Ker}\varphi$. L'étude de φ se ramène ainsi à celle la forme bilinéaire alternée non-dégénérée $\varphi_{|_F}$. On a, pour $x, y \in F$, $\widetilde{\varphi}(\pi(x), \pi(y)) = \varphi(x, y)$, c'est-à-dire, avec la notation de [L2] (I.3, 2.4.2, page 122) : $\widetilde{\varphi} \star \pi = \varphi_{|_F}$. Avec une définition introduite un peu plus loin, on dira que les formes *symplectiques* $\widetilde{\varphi}$ et $\varphi_{|_F}$ sont *équivalentes*.

Définition 1. Soit (E, Φ) un espace symplectique. Un sous-espace $F \subset E$ est dit *régulier* si la restriction de Φ à F n'est pas dégénérée ou si $F = (0)$.

Un sous-espace est régulier si, et seulement si, $F \cap F^\perp = (0)$. Sinon il est dit singulier[5]

On dit que les deux espaces symplectiques (E, φ) et (E', φ') sont *isomorphes*, ou que les deux formes bilinéaires φ et φ' sont *équivalentes*, s'il existe un isomorphisme de K-espaces vectoriels $u : E' \to E$ tel que, pour tous $x, y \in E$, $\varphi'(x, y) = \varphi(u(x), u(y))$, c'est-à-dire $\varphi' = \varphi \star u$. On dit alors que u est une *équivalence*. Deux formes symplectiques équivalentes ont « les mêmes propriétés ». On note $\varphi \sim \varphi'$ (resp. $(E, \varphi) \sim (E', \varphi')$) l'équivalence de deux formes symplectiques (resp. l'existence d'un isomorphisme entre deux espaces symplectiques).

On dit qu'une matrice $A = (a_{ij}) \in M_m(K)$ est dite *alternée* si pour tous $i \in [\![1, m]\!]$, $a_{j,i} = -a_{i,j}$ *et*, pour tout $i \in [\![1, m]\!]$, $a_{i,i} = 0$.

On rappelle (*cf.* [L2], I.3, 2.3.2) que si φ est une forme bilinéaire définie sur E, on définit la matrice de φ dans une base $\mathcal{B} = (e_1, \ldots, e_m)$ de E par $Mat(\varphi; \mathcal{B}) := (\varphi(e_i, e_j))$. Si φ est alternée (resp. antisymétrique), $Mat(\varphi; \mathcal{B})$ est alternée (resp.antisymétrique-. Les matrices de φ dans deux bases $\mathcal{B}$ et $\mathcal{B}'$ sont *congruentes* : $Mat(\varphi; \mathcal{B}') = {}^t P Mat(\varphi; \mathcal{B}) P$, où $P := Pass(\mathcal{B}, \mathcal{B}')$ est la *matrice de passage*.

[4] C'est une terminologie traditionnelle, mais elle n'est pas utilisée par tout le monde... Elle est différente de celle que nous utiliserons plus loin pour les formes quadratiques.

[5] Attention ! Certain auteurs appellent sous-espace-isotrope ce que nous appelons sous-espace singulier.

Nous verrons plus loin que si $M \in M_m(K)$ est une matrice alternée *inversible*, alors m est *pair*. Si $m = 2n$, on note $A_n(K) \subset M_{2n}(K)$ le sous-espace des matrices *alternées*. Si $M \in A_n(K)$, la forme bilinéaire définie par $(x, y) \mapsto {}^tXMY$ est alternée.

En caractéristique impaire une matrice est alternée si, et seulement si, elle est antisymétrique. Ce n'est pas le cas en caractéristique 2. Dans ce cas, par exemple, la matrice identité est antisymétrique et elle n'est pas alternée. Nous utiliserons seulement les matrices alternées, ce qui évite de s'inquiéter de la caractéristique.

Soit $n \in \mathbb{N}^*$. On note :

$$J_1 := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad J_n := \left(\begin{array}{c|c} 0_n & I_n \\ \hline -I_n & 0_n \end{array}\right), \quad J'_n := \mathrm{Diag}(\underbrace{J_1, \ldots, J_1}_{n}).$$

Ce sont des matrices carrées *alternées* inversibles de déterminant 1 et d'ordre $2n$.
Les formes bilinéaires alternées φ^{can} et ψ^{can} définies sur K^{2n} par les matrices J_n et J'_n sont particulièrement importantes. Elles sont dites *canoniques* :

$$\varphi^{can}(x, y) = {}^tXJ_nY, \quad \psi^{can}(x, y) = {}^tXJ'_nY,$$

$$\varphi^{can}\left((x_1, \ldots, x_n), (x_1, \ldots, x_n)\right) = \sum_{j=1}^{n} (x_j y_{n+j} - x_{n+j} y_j)$$

$$\psi^{can}\left((x_1, \ldots, x_n), (x_1, \ldots, x_n)\right) = \sum_{j=1}^{n} (x_{2j-1} y_{2j} - x_{2j} y_{2j-1}).$$

Si $(e_1, \ldots, e_{2n})$ est la base canonique de K^{2n}, on définit une nouvelle base $(\varepsilon_1, \ldots, \varepsilon_{2n})$ par *permutation* :

$$\varepsilon_1 := e_1,\ \varepsilon_2 := e_3,\ \ldots,\ \varepsilon_n := e_{2n-1}, \varepsilon_{n+1} := e_2,\ \varepsilon_{n+2} := e_4,\ \ldots, \varepsilon_{2n} := e_{2n}.$$

La permutation correspondante des indices est :

$$\sigma : (1, 2, \ldots, 2n) \mapsto (1, 3, \ldots, 2n-1, 2, 4, \ldots, 2n).$$

On vérifie que la matrice de ψ^{can} dans cette nouvelle base est φ^{can}. Par suite : $J_n = {}^tP_nJ'_nP_n$, où P_n est *une matrice de permutation* (*cf.* [L2], I.3, 3.2.2, définition 21).
Les matrices J_n et J'_n sont congruentes.
On définit $u \in \mathcal{GL}(E)$ par $u : x \mapsto u(x) = x'$, avec $X' = P_nX$. Soient $x, y \in E$. On a $u(y) = y'$, avec $Y' = P_nY$, et :

$$\varphi^{can}(x, y) = {}^tXJ_nY = {}^tX{}^tP_nJ'_nP_nY = {}^tX'J'_nY' = \psi^{can}(x', y') = \psi^{can}\left(u(x), u(y)\right). \tag{1}$$

Ainsi les espaces symplectiques (K^{2n}, φ^{can}) et (K^{2n}, ψ^{can}) sont *isomorphes*.

Nous avons remarqué dans le livre [L2] (*cf.* la remarque de la page 122) que les résultats sur l'orthogonalité que nous avons établi pour le cas symétrique dans le paragraphe *propriétés d'orthogonalité et d'isotropie* (*cf.* la page 129) restent vrais, *mutatis mutandis*, dans le cas antisymétrique (c'est encore vrai pour le cas alterné). Nous utiliserons ici, de façon essentielle, cette remarque. On a, en particulier, le résultat suivant.

Proposition 1.
Soient (E, φ) un *espace symplectique* et $F \subset E$ un sous-espace. Alors F est régulier si, et seulement si, $E = F \oplus F^{\perp}$. Si F est régulier, alors $F^{\perp}$ est régulier.

C'est la variante alternée de la proposition 39 de [L2], I.3, 3.1.1 (rappelée plus loin, *cf.* la proposition 20 de la page 92).

Soient φ une forme bilinéaire alternée sur E et $F, F' \subset E$ deux sous-espaces. Si $F \cap F' = (0)$ et $F \perp F'$, on dit que la somme directe $G := F \oplus F' \subset E$ est *orthogonale* et l'on écrit $G = F \overset{\perp}{\oplus} F'$. Si $F \subset E$ est un sous-espace régulier, alors $E = F \overset{\perp}{\oplus} F^{\perp}$.

1.1 Sous-espaces isotropes, involutifs, Lagrangiens

Soit (E, φ) un *espace symplectique*. Soient $F, G \subset E$ des sous-espaces. On rappelle les résultats suivants :

- $(F^{\perp})^{\perp} = F$;
- $\dim F + \dim F^{\perp} = \dim E$;
- si $F \subset G$, alors $G^{\perp} \subset F^{\perp}$;
- $(F \cap G)^{\perp} = F^{\perp} + G^{\perp}$, $(F + G)^{\perp} = F^{\perp} \cap G^{\perp}$.

Définition 2. Soit (E, φ) un *espace symplectique*.
Un sous-espace $F \subset E$ tel que $F^{\perp} \subset F$ est dit *involutif* (ou *coisotrope*). Un sous-espace tel que $F^{\perp} = F$ est dit *Lagrangien*.

Par ailleurs, un sous-espace $F \subset E$ est isotrope si, et seulement si, $F \subset F^{\perp}$. Un sous-espace est Lagrangien si, et seulement s'il est isotrope et involutif.
La terminologie *Lagrangien* vient de la mécanique (où l'on utilise des formalismes dits *Lagrangiens* et *Hamiltoniens*, en relation avec la géométrie symplectique, *cf.* [Arnold]).

Proposition 2. Soient (E, φ) un *espace symplectique* de dimension $2n$ et $F \subset E$ un sous-espace

(i) Si F est isotrope, $dim F \leqslant n$.

(ii) Si F est involutif, $dim F \geqslant n$.

(iii) Si F est Lagrangien, $\dim F = n$.

(iv) Le sous-espace F est isotrope si et seulement si $F^{\perp}$ est involutif.

(v) Les conditions suivantes sont équivalentes :

  a) F est Lagrangien ;

  b) F est isotrope maximal ;

  c) F est involutif minimal.

(vi) Si F est isotrope, il est contenu dans un sous-espace Lagrangien.

(vii) Si F est involutif, il contient un sous-espace Lagrangien.

Démonstration. Montrons (i). Si F est isotrope, $F \subset F^\perp$, donc :

$$\dim F \leqslant \dim F\perp = 2n - \dim F, \qquad \text{puis} \qquad 2\dim F \leqslant 2n \quad \text{et} \quad \dim F \leqslant n.$$

On déduit (ii) de (i) en utilisant $\perp$; (iii) résulte trivialement de (i) et (ii) ; (iv) est évident en utilisant $\perp$.

Montrons (v). Supposons a) et montrons b). La dimension de F est n, donc maximale pour la dimension d'un sous-espace involutif. Ainsi F est involutif maximal.

Supposons b) et montrons a). On raisonne par l'absurde. Si F n'est pas Lagrangien, alors $\dim F < n$ et l'inclusion $F \subset F^\perp$ est *stricte*. Il existe $e \in F^\perp \setminus F$. On considère $G := F + Ke$, on a $\dim G = \dim F + 1$. Montrons que G est *isotrope*. Soient $x, x' \in F$ et $\lambda, \lambda' \in K$. On a : $\varphi(x+\lambda e, x'+\lambda' e) = \varphi(x, x') + \lambda\varphi(e, x') + \lambda'\varphi(x, e) + \lambda\lambda'\varphi(e, e) = 0$. On a donc abouti à une contradiction. On montre l'équivalence de b) et de c) en utilisant $\perp$.

Les assertions (vi) et (vii) résultent trivialement des précédentes. ■

1.2 Classification des formes bilinéaires alternées

Le résultat suivant dit que, à équivalence près, *il n'y a qu'un seul espace symplectique de dimension* 2.

Proposition 3. Soit (E, φ) un *espace symplectique* de dimension 2. Il existe une base de E dans laquelle la matrice de φ est J_1. De plus on peut *choisir arbitrairement* le premier vecteur de la base.

Un espace symplectique de dimension 2 est apparlé *plan symplectique*.

Démonstration. Soit $\mathcal{B} = (e_1, e_2)$ une base de E. On a $\varphi(e_1, e_1) = \varphi(e_2, e_2) = 0$ et $\alpha := \varphi(e_1, e_2) \neq 0$. Sinon l'on aurait $e_1 \perp E$ et φ serait dégénérée. Soit $\lambda \in K^*$. On a $\varphi(e_1, \lambda e_2) = \lambda\alpha$. Par suite, si $\varepsilon := \alpha^{-1} e_2$, alors (e_1, ε) est une base de E et $\varphi(e_1, \varepsilon) = 1$, donc la matrice de φ dans la base (e_1, ε) est J_1. ■

Le résultat suivant est important. Il dit qu'à équivalence près, il n'y a qu'un seul espace symplectique de rang donné et que ce rang est nécessairement pair.

Théorème 4. Soit $m \in \mathbb{N}^*$. Soit (E, φ) un *espace symplectique* de dimension m. Alors m est *pair*, $m = 2n$, et E peut s'écrire $E = E_1 \overset{\perp}{\oplus} E_2 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} E_n$, où les $(E_i, \varphi_i = \varphi_{|_{E_i}})$ $(i \in [\![1, n]\!])$ sont des *plans symplectiques*.

Démonstration. On montre le résultat par récurrence sur $n = \dim E/2$.

Si $n = 1$ le résultat est évidemment vrai. Soit $n \geqslant 2$. On suppose que le résultat est vrai pour tout espace symplectique de dimension $2(n-1)$. Soit (E, φ) un espace symplectique de dimension $2n$. Soit $e_1 \in E$ un vecteur non nul. On va compléter la droite Ke_1 en un plan symplectique. Puisque φ est non dégénérée, il existe $v \in E$ tel que $\alpha := \varphi(e_1, v) \neq 0$. Alors $v \neq 0$ et $v \neq e_1$, donc $E_1 := Ke_1 + Kv$ est un plan. La matrice de la restriction $\varphi_1 := \varphi_{|_{E_1}}$ de φ à ce plan est αJ_1. Son déterminant est $\alpha^2 \neq 0$, donc φ_1 n'est pas dégénérée. Autrement dit $E_1 \subset E$ est un sous-espace *régulier*. Alors, d'après la proposition 1, $E = E_1 \overset{\perp}{\oplus} E_1^\perp$ et $E^\perp \subset E$ est un sous-espace régulier. On a $\dim E_1^\perp = 2n - 2 = 2(n-1)$.

La restriction de φ à $E_1^\perp$ est non dégénérée, donc, d'après l'hypothèse de récurrence, $E_1^\perp$ est une somme de $(n-1)$ plans symplectiques. Ainsi E est une somme directe orthogonale de n plans symplectiques. ∎

Pour tout $n \in \mathbb{N}^*$, *tous les espaces symplectiques de dimension* $2n$ *sont isomorphes*. Le seul invariant est la dimension. Si φ est dégénérée sur E de dimension k, on a $E = F \oplus \operatorname{Ker}\varphi$ et la restriction de φ à F est non dégénérée. Le rang de φ est $r = \dim F = 2n$. Les invariants sont donc k et r.
En utilisant la proposition 3 et les formules (1) de la page 78, on déduit du théorème 4 ci-dessus le résultat suivant.

Corollaire 5. Soit (E, φ) un *espace symplectique* de dimension $2n$.

(i) Il existe une base de E dans laquelle la matrice de φ est J'_n.

(ii) Il existe une base de E dans laquelle la matrice de φ est J_n.

Une base satisfaisant la condition (ii) est dite *symplectiques* pour φ.
Nous rappelons que deux matrices $M, M' \in M_m(K)$ sont dites *congruentes* s'il existe $P \in GL_n(K)$ telle que $M' = {}^tPMP$.

Corollaire 6. Soit $n \in \mathbb{N}^*$.

(i) Les matrices J_n et J'_n sont congruentes.

(ii) Toute matrice alternée $M \in A_n(K)$ inversible est congruente à la matrice J_n.

(iii) Soit $M \in A_n(K)$ non inversible et non nulle. Il existe $1 \leqslant m < n$ unique tel que M soit congruente à la matrice : $\widetilde{J}_{m,n} := \left(\begin{array}{c|c} J_m & 0_{m,2n-m} \\ \hline 0_{2n-m,m} & 0_{2n-m} \end{array}\right)$.

(iv) Si $M \in A_n(K)$, alors $\det M$ est *un carré de* K.

Démonstration. Les assertions (i) et (ii) sont évidentes. Montrons (iv). Soit $M \in A_n(K)$. Si $\det M = 0$, le résultat est évident.
Si $\det M \neq 0$, alors il existe $P \in GL_n(K)$ tel que $M = {}^tPJ_nP$ et $\det M = (\det P)^2$.
Pour montrer (iv), on commence par monter que M est congruente à une matrice :
$\tilde{M} = \left(\begin{array}{c|c} M' & 0_{m,2n-m} \\ \hline 0_{2n-m,m} & 0_{2n-m} \end{array}\right)$, avec $M' \in A_m(K)$ inversible. On applique ensuite (ii) à la matrice M'. ∎

Exercice 1.
Soit (E, φ) un *espace symplectique*. Montrer qu'il existe, dans E, deux sous-espaces Lagrangiens supplémentaires.
Solution. On utilise une décomposition de E en somme directe orthogonale de plans symplectiques : $E = E_1 \overset{\perp}{\oplus} E_2 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} E_n$. Pour tout $i \in [\![1, n]\!]$, on choisit une base (e_i, e'_i) de E_i. On pose $L = Ke_1 + Ke_2 + \cdots + Ke_n$ et $L' = Ke'_1 + Ke'_2 + \cdots + Ke'_n$. Les sous-espaces L et L' sont involutifs et de dimension n, donc Lagrangiens. On a $L \cap L' = (0)$,

car $(e_1, e_2, \ldots, e_n, e'_1, e'_2, \ldots, e_n)$ est une base de E, et $\dim L + \dim L' = 2n = \dim E$, donc $E = L \oplus L'$.

Théorème 7. Soit (E, φ) un *espace symplectique* de dimension $2n$. Soit $L \subset E$ un sous-espace *Lagrangien*. Alors L admet un supplémentaire Lagrangien L' dans E. Si l'on choisit une base $(e_1, \ldots, e_n)$ de L, on peut trouver une base $(e'_1, \ldots, e'_n)$ de L' telle que, dans la base $(e_1, \ldots, e_n, e'_1, \ldots, e'_n)$ de $L \oplus L' = E$, la matrice de φ ait pour matrice J_n.

Démonstration. Soit $F \subset E$ isotrope et tel que $L \cap F = (0)$. Il existe un tel sous-espace, puisque $F = (0)$ vérifie ces conditions. On va « gonfler » ce sous-espace jusqu'à en faire un espace Lagrangien L'.
Si $F = F^\perp$, c'est terminé. Sinon $\dim F^\perp > n$ et $\dim F^\perp + \dim L > 2n = \dim E$, donc $F^\perp \cap L \neq (0)$. Puisque $L \cap F = (0)$, $F^\perp \cap L$ n'est pas contenu dans F. on en déduit, en passant aux orthogonaux que $F^\perp$ n'est pas contenu dans $F + L^\perp = F + L$. Soit $e \in F^\perp \setminus (F + L)$. On pose $F_1 := F \oplus Ke$. On a $L \cap F_1 = (0)$, car $L \cap F = (0)$ et $e \notin F + L$. Le sous-espace $F_1 \subset E$ est de dimension $\dim F + 1$ et il est isotrope, car $e \in F^\perp$. Si $\dim F_1 = n$, F_1 est Lagrangien. Sinon on reprend le même argument en remplaçant F par F_1.On termine la preuve de l'existence de L' en un nombre fini d'étapes.
Il reste à prouver le résultat sur les bases. On utilise *l'isomorphisme* d_φ entre E et E^* associé à la forme bilinéaire non dégénérée φ (*cf.* [L2], I.3, 2.3.1). On choisit une base $(\varepsilon_1, \ldots, \varepsilon_n)$ de L'. alors $(e_1, \ldots, e_n, \varepsilon_1, \ldots, \varepsilon_n)$ est une base de E^*. On considère la *base duale* : $(e_1^*, \ldots, e_n^*, \varepsilon_1^*, \ldots, \varepsilon_n^*)$ de E^*. On pose $e'_i := d_\varphi^{-1}(e_i^*)$ $(i \in [\![1, n]\!])$. Alors $(e'_1, \ldots, e'_n)$ est une base de L'. En effet $e'_i \in L'^\perp$, car, pâr définition de d_φ :

$$\varphi(e'_i, \varepsilon_j) = -\varphi(\varepsilon_j, e'_i) = -d_\varphi(e'_i)(eps_j) = -e_i^*(\varepsilon_j) = 0.$$

donc $e'_i \in L' = (L')^\perp$, car L' est Lagrangien; et, par ailleurs, la famille $(e'_1, \ldots, e'_n)$ est l'image par un isomorphisme d'une famille libre, elle est donc libre. On a, pour tous $i, j \in [\![1, n]\!]$: $\varphi(e'_i, e_j) = -e_i^*(e_j) = -\delta_{i,j}$, $\varphi(e_i, e_j) = 0$ et $\varphi(e'_i, e'_j) = 0$, donc la matrice de φ dans la base $(e_1, \ldots, e_n, e'_1, \ldots, e'_n)$ est J_n. ∎

Remarque. Le théorème 7 ci-dessus redonne une preuve du théorème 4. Le lecteur le vérifiera facilement.

1.3 Pfaffien

Nous avons vu (*cf.* le corollaire 6) que le déterminant d'une matrice alternée $M \in A_n(K)$ est *le carré d'un élément de* K. On va voir que c'est, plus précisément, le carré d'une fonction polynomiale (à coefficients entiers) en les coefficients de M, appelée le *Pfaffien*.
Le Pfaffien a de nombreuses applications en mathématiques (théorie des classes caractéristiques, graphes planaires) et en physique (verres de spin). L'appellation Pfaffien est due au mathématicien anglais A. Cayley qui a introduit cet objet en relation avec l'étude des systèmes différentiels dits de Pfaff (vers 1850).

1.3.1 Le pfaffien universel

Soit $n \in \mathbb{N}^*$. On considère l'anneau de polynômes $B := \mathbb{Z}[X_{i,j}]_{1 \leqslant i < j \leqslant 2n}$ et son corps des fractions $L := \mathbb{Q}(X_{i,j})_{1 \leqslant i < j \leqslant 2n}$ (il y a $n(2n-1)$ variables).

On définit une matrice alternée « universelle » à coefficients dans B : $\widetilde{M} := (\tilde{m}_{i,j})$, où $\tilde{m}_{i,i} := 0$, pour $i \in [\![1, 2n]\!]$ et $\tilde{m}_{i,j} := -m_{i,j} := X_{i,j}$, pour $1 \leqslant i < j \leqslant 2n$. Le déterminant D de $\widetilde{M}$ appartient à B et c'est un carré dans L : $D = R^2$, $R \in L$. L'anneau B est *factoriel*. On écrit $R = F/G$, avec $F, G \in B$, sans facteur commun (aux inversibles près). On a $F^2/G^2 = D$, donc $F^2 = DG^2$. Les facteurs irréductibles des deux membres sont égaux, aux inversibles près. Si G avait un facteur irréductible non trivial, ce serait un facteur irréductible de F, mais c'est, par hypothèse, impossible, donc G est inversible et $R = F/G \in B$. Il existe donc un polynôme, unique au signe près, dont le carré est $\det \widetilde{M}$.

Soit $M = (m_{i,j}) \in A_n(K)$. Soit $f : B \to K$ le morphisme de substitution des $a_{i,j}$ aux $X_{i,j}$ (pour $1 \leqslant i < j \leqslant 2n$). On a $\det M = f(\det \widetilde{M}) = f(R)^2$, soit $\det M = (R(a_{i,j}))^2$.

Si $K := \mathbb{Q}$ et $M := J'_n$, on a $\det J'_n = 1$, donc la valeur de R en les coefficients de J'_n strictement au-dessus de la diagonale principale est ± 1. On choisir R de telle sorte que cette valeur soit 1. On l'appelle alors *Pfaffien universel* d'ordre $2n$ et on le note Pf. Pour toute matrice alternée $M \in A_n(K)$, on a $\det M = (\mathrm{Pf}\, M)^2$, en notant $\mathrm{Pf}\, M$ la valeur du polynôme Pf en les coefficients $m_{i,j}$ de M tels que $1 \leqslant i < j \leqslant 2n$. On dit que $\mathrm{Pf}\, M \in K$ est *le Pfaffien de* M. Le Pfaffien universel est évidemment le Pfaffien de la matrice $\widetilde{M} \in A_n(L)$.

1.3.2 Formule explicite

Nous allons donner une formule explicite pour le Pfaffien. Pour ce faire, nous définirons, par une formule combinatoire, un élément de K, noté $\mathrm{pf}\, M$ attaché à une matrice alternée $M \in A_n(K)$. Les propriétés de pf, établies par des arguments purement combinatoires, nous permettrons ensuite de montrer que $\mathrm{Pf}\, M = \mathrm{pf}\, M$. Nous supposons dans ce paragraphe que K est *de caractéristique nulle* (on pourrait s'en passer en « raffinant » la combinatoire, comme on peut le deviner sur les exemples ci-dessous pour $n = 1, 2$).

La formule pour pf est « parachutée ». La situation est similaire à celle du déterminant que certains auteurs définissent en « parachutant » la formule classique[6]. Pour le déterminant, il est plus élégant d'utiliser les formes multilinéaires et alternées. Il en est de même pour le Pfaffien. On peut faire apparaître assez simplement la formule[7] en utilisant l'algèbre extérieure (*cf.* [MPA1]). Les propriétés sont aussi plus faciles à établir dans ce cadre, mais nous ne disposons pas ici de l'outillage nécessaire.

Si $M \in A_n(K)$, on note $M = (m_{i,j})_{i,j \in [\![1,2n]\!]}$. On note $\mathfrak{S}_{2n}$ le groupe des permutations de $2n$ éléments. On note $\varepsilon(\sigma) \in \{\pm 1\}$ la signature d'une permutation $\sigma \in \mathfrak{S}_{2n}$.

Définition 3. On définit une application : $\mathrm{pf} : A_{2n}(K) \to K$ par :

$$\mathrm{pf}\, M := \frac{1}{2^n n!} \sum_{\sigma \in \mathfrak{S}_{2n}} \varepsilon(\sigma) \prod_{i=1}^{n} m_{\sigma(2i-1),\sigma(2i)}.$$

[6] Historiquement la première définition du déterminant a été donnée de cette façon.

[7] Dans le cadre de l'algèbre extérieure, on considère la puissance extérieure $n^{ème}$ d'un bi-vecteur.

Exemples.

$$\text{pf}\begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix} = a$$

$$\text{pf}\begin{pmatrix} 0 & a_{1,2} & a_{1,3} & a_{1,4} \\ -a_{1,2} & 0 & a_{2,3} & a_{2,4} \\ -a_{1,3} & -a_{2,3} & 0 & a_{3,4} \\ -a_{1,4} & -a_{2,4} & -a_{3,4} & 0 \end{pmatrix} = a_{1,2}a_{3,4} - a_{1,3}a_{2,4} + a_{1,4}a_{2,3}.$$

Dans les deux cas on peut vérifier la formule $\det M = (\text{pf } M)^2$ (laborieusement dans le deuxième cas...). On notera que ces formules gardent un sens en caractéristique arbitraire.

On vérifie facilement : pf $(\lambda M) = \lambda^n \text{pf } M$ $(\lambda \in K)$, pf ${}^tM = (-1)^n \text{pf } M$,
En utilisant les propriétés de la signature d'une permutation, on vérifie le résultat suivant.

Proposition 8. Si $M = \text{Diag}(M_1, M_2)$ (par blocs), on a pf $M = (\text{pf } M_1)(\text{pf } M_2)$.

Proposition 9.
Pour tous $M \in A_{2n}(K)$ et tout $P \in M_{2n}(K)$, on a

$$\text{pf } (PM^tP) = (\text{pf } M) \det P. \tag{2}$$

Démonstration. Nous allons faire une démonstration purement combinatoire.
On note $C := PM^tP$, $P = (p_{i,j})_{i,j\in[\![1,2n]\!]}$, $C = (c_{i,j})_{i,j\in[\![1,2n]\!]}$.
On a $c_{i,j} = \sum\limits_{k,\ell\in[\![1,2n]\!]} p_{i,k}m_{k,\ell}p_{j,\ell}$. Par suite :

$$\text{pf } (PM^tP) = \frac{1}{2^n n!}\sum_{\sigma\in\mathfrak{S}_{2n}} \varepsilon(\sigma)\prod_{i=1}^{n}\sum_{k_i,\ell_i\in[\![1,2n]\!]} p_{\sigma(2i-1),k_i}m_{k_i,\ell_i}p_{\sigma(2i),\ell_i}$$

On a, en réarrangeant les termes :

$$\sum_{\sigma\in\mathfrak{S}_{2n}} \varepsilon(\sigma)\prod_{i=1}^{n}\sum_{k_i,\ell_i\in[\![1,2n]\!]} p_{\sigma(2i-1),k_i}m_{k_i,\ell_i}p_{\sigma(2i),\ell_i} =$$

$$\sum_{k_1,\ell_1,k_2,\ell_2\ldots,k_n,\ell_n}\left(\sum_{\sigma\in\mathfrak{S}_{2n}} \varepsilon(\sigma)\prod_{i=1}^{n} p_{\sigma(2i-1),k_i}p_{\sigma(2i),\ell_i}\right) m_{k_1,\ell_1}m_{k_2,\ell_2}\ldots m_{k_n,\ell_n}.$$

La première somme porte sur les $k_i, \ell_j \in [\![1, 2n]\!]$. On rappelle la formule du déterminant :

$$\det P = \sum_{\sigma\in\mathfrak{S}_{2n}} \varepsilon(\sigma)p_{\sigma(1),1}p_{\sigma(2),2}\cdots p_{\sigma(2n-1),2n-1}p_{\sigma(2n),2n}.$$

On va utiliser le fait que le déterminant est une fonction multilinéaire et alternée des vecteurs colonnes. Si l'on remplace dans l'expression ci-dessus les indices $1, 2, \ldots, 2n-1, 2n$ par $k_1, \ell_1, \ldots, k_n, \ell_n$, on obtient le développement du déterminant d'une matrice fabriquée avec des colonnes de P, avec des répétitions éventuelles. Il y a deux cas possibles :

1. ou bien il y a au moins une répétition de colonne et alors l'expression est nulle ;

2. ou bien $k_1, \ell_1, \ldots, k_n, \ell_n$ est une permutation τ de $1, 2, \ldots, 2n-1, 2n$, c'est-à-dire : $\tau(1) = k_1, \tau(2) = \ell_1, \ldots, \tau(2n-1) = k_n, \tau(2n) = \ell_n$, et alors :

$$\varepsilon(\tau) \det P = \sum_{\sigma \in \mathfrak{S}_{2n}} \varepsilon(\sigma) \prod_{i=1}^{n} p_{\sigma(2i-1),\tau(2i-1)} p_{\sigma(2i),\tau(2i)}.$$

Par suite :

$$\text{pf}\,(PM^tP) = \frac{1}{2^n n!} \det P \sum_{\tau \in \mathfrak{S}_{2n}} \varepsilon(\tau) \prod_{i=1}^{n} m_{\tau(2i-1),\tau(2i)} = (\text{pf}\, M)\, \det P. \qquad \blacksquare$$

Remarque. Dans la définition 3 on peut remplacer le corps K par un anneau commutatif (de caractéritique nulle). On vérifie facilement que les propositions 8 et 9 restent vraies dans ce cadre.

Exercice 2.

Calculer pf J'_n et pf J_n.

Solution. D'après la proposition 8, on a pf $J'_n = (\text{pf}\, J_1)^n = 1^n = 1$.

On a $J_n = PJ'_n{}^tP$ où P est une matrice de permutation. Le déterminant de P est la signature de la permutation $\sigma : (1, 2, \ldots, 2n) \mapsto (1, 3, \ldots, 2n-1, 2, 4, \ldots, 2n)$. Le *nombre d'inversions* $\ell(\sigma)$ de la permutation σ est $n-1+n-2+\cdots+2+1 = n(n-1)/2$. Sa signature est donc $\varepsilon(\sigma) = (-1)^{\ell(\sigma)} = (-1)^{n(n-1)/2}$ (*cf.* la proposition 27 de II.2 2.3 du livre [L1]). Par suite :

$$\text{pf}\, J_n = (\det P)\, \text{pf}\, J'_n = \det P = (-1)^{n(n-1)/2}. \tag{3}$$

Exercice 3.

Soit $M := \left(\begin{array}{c|c} 0_n & A \\ \hline -{}^tA & 0_n \end{array}\right)$, avec $A \in M_n(\mathbb{K})$. Montrer que pf $M = (-1)^{n(n-1)/2} \det A$.

Solution. On vérifie : $M = \left(\begin{array}{c|c} 0_n & I_n \\ \hline {}^tA & 0_n \end{array}\right) \left(\begin{array}{c|c} 0_n & -I_n \\ \hline I_n & 0_n \end{array}\right) \left(\begin{array}{c|c} 0_n & A \\ \hline I_n & 0_n \end{array}\right) = P(-J_n)\,{}^tP$, avec $P := \left(\begin{array}{c|c} 0_n & I_n \\ \hline {}^tA & 0_n \end{array}\right)$.

On a $\det P = (-1)^n \det A$, donc pf $M = (-1)^n (\det A)\text{pf}(-J_n)$. Par suite :

$$\text{pf}\, M = (\det A) \det J_n = (-1)^{n(n-1)/2} \det A.$$

Proposition 10. Soit $M \in A_n(K)$. Si $\det M = 0$, alors pf $M = 0$.

Démonstration. Si $\det M = 0$, d'après le corollaire 6, il existe $1 \leqslant m < n$ et une matrice $P \in M_{2n}(K)$ telle que $M = P\widetilde{J}\,{}^tP$, avec $\widetilde{J} := \left(\begin{array}{c|c} J_m & 0_{m,n-m} \\ \hline 0_{n-m,m} & 0_{n-m} \end{array}\right)$. D'après la proposition 8, on a pf $\widetilde{J} = 0$. D'après la proposition 9, on a :

$$\text{pf}\, M = \text{pf}\,(P\widetilde{J}\,{}^tP) = (\det P)\, \text{pf}\, \widetilde{J} = 0. \qquad \blacksquare$$

Proposition 11. Pour tous $M \in A_{2n}(K)$, on a $\det M = (\text{pf } M)^2$.

Démonstration. Si $\det M = 0$, on a, d'après la proposition ci-dessus, $\text{pf } M = 0$ et le résultat est vrai.
On peut donc supposer que $\det M \neq 0$. Pour toute matrice $P \in M_{2n}(K)$, on a,

$$\left(\text{pf } (PM^tP)\right)^2 = (\text{pf } M)^2 (\det P)^2 .$$

On a $\det(PM^tP) = (\det M)(\det P)^2$. Supposons $\det P \neq 0$, alors :

$$\frac{(\text{pf } (M))^2}{\det(M)} = \frac{\left(\text{pf } (PM^tP)\right)^2}{\det(PM^tP)}.$$

On calcule le second membre en utilisant un « bon choix » de la matrice P. Il existe $P = P'$ (non singulière) telle que ${}^tP'MP' = J'_n$. On a $\text{pf} J'_n = 1$ et $\det J'_n = 1$, donc :

$$\frac{\left(\text{pf } (P'M^tP')\right)^2}{\det(P'M^tP')} = \frac{(\text{pf} J'_n)^2}{\det J'_n} = 1.$$

On a donc $\det M = (\text{pf } M)^2$. ∎

Corollaire 12. Soit $\widetilde{M} \in L = \mathbb{Q}(X_{i,j}; 1 \leqslant i < j \leqslant 2n)$ la matrice alternée universelle. On a $\text{Pf} = \text{pf } \widetilde{M}$.

Notons que Pf est un polynôme homogène de degré n.

1.4 Introduction au groupe symplectique

Le *groupe symplectique* d'une forme bilinéaire alternée φ non dégénérée sur E est le groupe des automorphismes de E conservant φ. On le note $Sp(\varphi)$. C'est un sous-groupe de $\mathcal{GL}(\mathcal{E})$. C'est un analogue du groupe orthogonal d'une forme quadratique. A isomorphisme près (en utilisant le corollaire 5), il suffit de considérer le cas de la forme canonique φ_n^{can} (ou ψ_n^{can}) sur K^{2n}. Nous ne donnerons que quelques propriétés de base. Le lecteur intéressé pourra se reporter à [MPA1].

Proposition 13. Soient (E, φ) un espace symplectique et J la matrice de φ dans ue base $\mathcal{B}$ de E. Pour que $u \in Sp(\varphi)$, il faut et il suffit que sa matrice M dans la base $\mathcal{B}$ vérifie l'égalité ${}^tMJM = J$.

Démonstration. Une matrice $M \in M_{2n}(K)$ est la matrice, dans la base $\mathcal{B}$, d'une transformation linéaire u conservant φ^{can} si, et seulement si, :

$$\forall\, x, y \in K^{2n}, \quad \varphi(x, y) = \varphi\left(u(x), u(y)\right).$$

Ceci équivaut à ${}^tXJY = {}^tX{}^tMJMY$, pour tous $x, y \in K^{2n}$, c'est-à-dire à ${}^tMJM = J$. ∎

On en déduit l'énoncé suivant.

Proposition et définition 14. Une matrice $M \in M_{2n}(K)$ est dite *symplectique* si ${}^tMJ_nM = J_n$.

(i) Les matrices symplectiques forment un sous-groupe $Sp_{2n}(K)$ de $GL_{2n}(K)$, appelé groupe symplectique à $2n$ variables sur K.

(ii) Une matrice M est symplectique si, et seulement si, elle est la matrice dans la base canonique de K^{2n} d'une transformation de $Sp(\varphi^{can})$.

Une matrice symplectique M est inversible et l'on a : $M^{-1} = J_n^{-1}\,{}^tMJ_n = -J_n\,{}^tMJ_n$ (en utilisant $J_n^{-1} = -J_n$).

Théorème 15. Toute transformation symplectique est de déterminant 1.

Ainsi $Sp(\varphi)$ est un sous-groupe de $\mathcal{SL}(E)$ et $Sp_{2n}(K)$ est un sous-groupe de $SL_{2n}(K)$.

Démonstration. On se ramène au cas symplectique canonique sur K^{2n}. La matrice M d'une transformation symplectique vérifie ${}^tMJ_nM = J_n$. On a alors :

$$(\det M)^2 \det J_n = \det J_n,$$

donc $\det M = \pm 1$ ($\det J_n \neq 0$). Il est plus difficile de montrer que $\det M = 1$. Nous allons utiliser le Pfaffien, en nous limitant au cas où K est de caractéristique nulle. Pour le cas général, il faut utiliser une autre preuve : le groupe symplectique est engendré par des transvections symplectiques (*cf.* le théorème 19 ci-dessous) et le déterminant d'une transvection est 1.

Supposons K de caractéristique nulle. On a Pf $(tMJ_nM) =$ Pf J_n. D'après la proposition 9, on a Pf $(tMJ_nM) = (\det M)$Pf J_n, donc $(\det M)$Pf $J_n =$ Pf J_n et $\det M = 1$ (Pf $J_n \neq 0$). ∎

En utilisant la multiplication des matrices 2×2 par blocs $n \times n$, on vérifie facilement le résultat suivant.

Proposition 16. Soit $M := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$ une matrice 2×2 par blocs, avec $A, B, C, D \in M_n(K)$. Pour que M soit symplectique, il faut et il suffit que A, B, C, D vérifient les conditions suivantes :

$$ {}^tAD - {}^tCB = I_n, \quad {}^tCA = {}^tAC \quad \text{et} \quad {}^tDB = {}^tBD. \tag{4}$$

Pour $n = 1$, on obtient : $Sp_2(K) = SL_2(K)$. En effet les conditions (4) sont alors équivalentes à $\det M = 1$.

Proposition et définition 17. Les transvections de E appartenant à $Sp(\varphi)$ sont les transformations $t_{a,\lambda} : x \mapsto x + \lambda\varphi(a, x)a$, où $a \in E$ et $\lambda \in K$ sont fixés. Ces transvections sont appelées *transvections symplectiques*.

On admet ce résultat. Une preuve fait l'objet de l'exercice I.2.2 de la page 118.

Corollaire 18. Le centre du groupe $Sp(\varphi)$ est $\{Id, -Id\}$.

Démonstration. Les transformations Id et $-Id$ appartiennent évidemment au centre. Inversement, soit g appartenant au centre de $Sp(\varphi)$. Pour tout $a \in E$ non nul, g commute avec la transvection $t_{a,1}$: $g \circ t_{a,1} \circ g^{-1} = t_{a,1}$. En écrivant la commutation, on en déduit que, pour tout $x \in E$, l'on a $\varphi(a, g(x))\, a = \varphi(a, x) g(a)$. En choisissant x tel que $\varphi(a, x) \neq 0$, on en déduit que $g(a) = \lambda_a a$ avec $\lambda_a \in K$.
On en déduit que g laisse stable la droite Ka. Puisque a est arbitraire, toutes les droites de E sont stables par g et g est donc une *homothétie* de rapport λ (*cf.* l'exercice I.2.19).
On a $\varphi(x, y) = \varphi(\lambda x, \lambda y) = \lambda^2 \varphi(x, y)$, donc $\lambda = \pm 1$. ■

On admet le résultat suivant. Une preuve fait l'objet de l'exercice I.2.3 de la page 118.

Théorème 19. Soit (E, φ) un espace symplectique. Le groupe symplectique $Sp(\varphi)$ est engendré par les transvections symplectiques.

2 Compléments sur les formes quadratiques

Dans le module du livre [L2] nous avons, pour les formes quadratiques, exclu le cas des corps de caractéristique 2. Nous souhaitons ici nous affranchir de cette restriction. Pour cette raison, nous utiliserons une définition *différente* des formes quadratiques[8].

Définition 4. Soit K un corps commutatif de caractéristique arbitraire. On dit que l'application $\Phi : E \to K$ est une forme quadratique si les deux conditions suivantes sont satisfaites :

a) pour tous $\lambda \in K$ et $x \in E$, $\Phi(\lambda x) = \lambda^2 \Phi(x)$;

b) pour tous $x, y \in E$, l'application $B_\Phi : (x, y) \in E^2 \mapsto \Phi(x + y) - \Phi(x) - \Phi(y)$ est une forme bilinéaire.

On dit que B_Φ est *l'application bilinéaire associée à* Φ. Si la caractéristique de K est différente de 2, alors la forme bilinéaire $\varphi := \frac{1}{2} B_\Phi$ est dite *la forme polaire* de Φ.

Les formes B_Φ et φ (quand elle est définie) sont évidemment *symétriques*. On a :

$$B_\Phi(x, x) = 4\, \Phi(x) - 2\, \Phi(x) = 2\, \Phi(x).$$

En caractéristique différente de 2, on en déduit $\Phi(x) = \varphi(x, x)$. En caractéristique 2, on en déduit $B_\Phi(x, x) = 2\Phi(x) = 0$: la forme B_Φ est *alternée*. On vérifie, en utilisant une identité de polarisation que, en caractéristique impaire, la définition ci-dessus est *équivalente* à celle donnée dans le livre [L2] (*cf.* I.3, 3.1.1, théorème et définition 34). Dans ce cas, tous les résultats de [L2] restent donc valides dans notre nouveau cadre et le lecteur pourra ne rien changer à ses habitudes.
Notons qu'avec la définition 4 ci-desus, en caractéristique 2, un polynôme homogène de degré deux arbitraire $P \in K[X_1, \ldots, X_n]$ définit une forme quadratique Ce n'est pas toujours le cas avec l'ancienne : considérer $(x, y) \mapsto xy$ sur K^2.

[8] C'est la définition de N. Bourbaki, Algèbre, Ch. 9, §3, 4, Définition 2, page 54.

Remarques.

1. Nous nous limitons aux formes quadratiques sur un *corps*, mais on peut étendre de façon évidente la définition 4 en remplaçant le corps K par un *anneau commutatif* A et le K-espace vectoriel E par un A-module. Nous n'étudierons pas cette question. Voici toutefois quelques indications dans le cas très important (en particulier sur le plan historique) où $A := \mathbb{Z}$. Si $A := \mathbb{Z}$, on obtient la notion de *forme quadratique entière*. Pour les forme quadratiques entières sur $\mathbb{Z}^n$, la relation d'équivalence naturelle est $\Phi \sim \Psi$ si $\Psi = \Phi \circ u$, la matrice de u dans la base canonique appartenant à $GL_n(\mathbb{Z})$. Une famille importante de formes quadratiques entières est celle des formes quadratiques entières *binaires* $ax^2 + bxy + cy^2$, avec $a, b, c \in \mathbb{Z}$. Elle a été initialement étudiée par Fermat, Lagrange Legendre, puis par Gauß (dans *Disquisitiones Arithmeticae*, 1801), et plus tard par Minkowski. C'est encore un sujet très actif (arithmétique, algorithmes de factorisation, cryptographie ...). Si deux formes entières binaires sont équivalentes sous l'action de $SL_2(\mathbb{Z})$, elles représentent les mêmes entiers et on peut utiliser l'équivalence pour résoudre des problèmes arithmétiques. On peut considérer une forme entière comme une forme réelle et lui appliquer les techniques de ce module. Nous proposerons quelques exercices dans cette direction : I.2.7, I.2.8 , I.2.15.

   Le vocabulaire utilisé pour les formes binaires est traditionnel. Il est un peu différent[9] de celui que nous utilisons pour les formes quadratiques générales. Le *discriminant* D de la forme binaire $\Phi(x, y) = ax^2 + bxy + cxy$ est, par définition, $D := b^2 - 4ac$ (c'est celui du lycée pour $at^2 + bt + c$). Il est clairement invariant sous l'action de $SL_2(\mathbb{Z})$. Si $D > 0$, on dit que la forme quadratique binaire est *réelle*. Si $D < 0$ on dit que la forme quadratique binaire est *imaginaire* et l'on suppose $a > 0$ (c'est-à-dire Φ *définie positive*). Il est clair que D est un carré modulo 4, on a donc $D \equiv 0$ ou $1 \mod 4$. On note $Q(D)$ l'ensemble des formes binaires de discriminant D et $H(D)$ l'ensemble des orbites de $Q(D)$ sous l'action de $SL_2(\mathbb{Z})$. On montre que $H(D)$ est *fini*. (Pour le cas *imaginaire*, une preuve fait l'objet de l'exercice I.2.8 de la page 120). Les éléments de $H(D)$ sont les *classes* de D[10]. Le nombre d'éléments de $H(D)$ s'appelle le *nombre de classes* de D. Gauß a défini une structure de groupe abélien sur les classes. Le calcul (des points de vue théoriques et algorithmiques) de la structure et du cardinal de ces groupes est encore un sujet de recherche actif.

2. La théorie des formes quadratiques en caractéristique 2 est intéressante[11]. En particulier, elle a été beaucoup utilisée en topologie depuis les années 50.[12]. Il est clair que, dans plusieurs questions, par exemple pour l'étude des coniques sur un corps fini, il n'est pas naturel de restreindre la caractéristique. Nous donnerons quelques résultats « de base » sur les formes quadratiques en caractéristique 2. Au delà, la plupart des techniques nécessaires sont assez différentes des techniques classiques en caractéristique impaire et elles sortent pour l'essentiel du cadre de nos livres. Toutefois les outils dont nous disposons dans ce module sur les espaces symplectiques permettent d'obtenir plusieurs résultats « significatifs ». On peut, par exemple, classifier les formes quadratiques non dégénérées sur les corps finis de caractéristique 2 en introduisant un invariant dit *de Arf*. Cest l'objet des exercices I.2.4 et I.2.9.

---

[9] On verra plus loin que, pour nous, le discriminant δ est défini à partir de $ac - b^2/4 = -D/4$.

[10] Ou du corps quadratique $\mathbb{Q}(\sqrt{D}) = \mathbb{Q}[X]/(X^2 - D)/Q[X]$.

[11] Son étude systématique a été commencée par le mathématicien turc Cahit Arf (un élève de Helmut Hasse) en 1941. Auparavant, en 1901 le mathématicien américain L. E. Dickson avait classifié les formes quadratiques sur F_2. Cette étude a été prolongée, ente autres, par Witt et Dieudonné et elle se poursuit activement.

[12] Invariants d'un nœud, invariants des variétés de dimension $4m + 2$ (Kervaire), homotopie des sphères, sphères exotiques....

Dans cette partie, sauf mention *explicite* du contraire, le corps K est supposé de caractéristique différente de 2. Nous utiliserons alors φ au lieu de B_Φ. Nous signalerons tout au long du texte les particularités de la caractéristique 2. Dans les parties 2.4 et 2.5 la caractéristique de K est *toujours* supposée impaire. On pourra bien sûr, en première lecture, supposer que la caractéristique est toujours impaire dans tout le texte, en « sautant » tout ce qui concerne la caractéristique 2.

Pour les notations, définitions (à l'exception de la définition 4) et résultats que nous utiliserons le lecteur se reportera au module I.3 du livre [L2] (et plus précisément à I.3, 3), que le présent module complète. Nous ferons quelques rappels pour faciliter la lecture et préciser quelques points. Ils ne remplacent pas la (re)lecture de [L2], I.3, 3.

Conseil de lecture. Les propriétés d'orthogonalité jouent un rôle essentiel dans cette partie, comme dans la précédente. Le lecteur uniquement habitué au cas *euclidien* aura vraisemblablement au début quelques difficultés avec les propriétés de *l'orthogonalité* en présence de phénomènes *d'isotropie*. Nous lui conseillons de *relire soigneusement* le paragraphe *Propriétés d'orthogonalité et d'isotropie* de [L2], I.3, 3.1.1 (page 129). Pour le cas quadratique, la terminologie concernant l'isotropie n'est pas complètement fixée, la notre peut différer de celle de certains cours ou documents, nous conseillons au lecteur d'y prendre garde.

2.1 Rappels et compléments

Nous rappelons quelques définitions et résultats de [L2] I.3.3. Les compléments portent sur la caractéristique 2.

La donnée d'une forme quadratique Φ sur un K-espace vectoriel E est appelée *espace quadratique*. Cet espace est noté (E, Φ). On emploiera les mêmes qualificatifs pour la forme et l'espace : non-dégénéré(e), isotrope, ... En caractéristique impaire, la forme polaire de Φ (resp. Φ', Ψ ...) sera notée φ (resp. φ', ψ ...). Si $\dim E = 2$, on dit que (E, Φ) est un *plan quadratique*.

On dit qu'une forme quadratique Φ sur le K-espace vectoriel E *représente* $a \in K$, s'il existe $x \in E^*$ tel que $\Phi(x) = a$. On dit aussi que a est *représenté par* Φ.

A deux espaces quadratiques (E_1, Φ_1) et (E_2, Φ_2) on associe leur somme directe orthogonale (E, Φ). Elle est définie par $E = E_1 \times E_2$ et $\Phi(x_1, x_2) = \Phi(x_1) + \Phi(x_2)$, pour tout $x = (x_1, x_2) \in E$. On note $(E, \Phi) = (E_1, \Phi_1) \overset{\perp}{\oplus} (E_2, \Phi_2)$ et parfois, par abus de langage, $\Phi = \Phi_1 \overset{\perp}{\oplus} \Phi_2$. On étend de façon évidente la notion de somme directe orthogonale au cas d'un nombre fini d'espaces quadratiques : $(E, \Phi) := (E_1, \Phi_1) \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} (E_p, \Phi_p)$.

On dit que les deux espaces quadratiques (E, Φ) et (E', Φ') sont *isomorphes*, ou que les deux formes quadratiques Φ et Φ' sont *équivalentes*, s'il existe un isomorphisme de K-espaces vectoriels $u : E' \to E$ tel que $\Phi' = \Phi \circ u$. On dit alors que u est une *isométrie*, ou une *équivalence*. L'équivalence est évidemment une relation d'équivalence. Deux formes quadratiques équivalentes ont « les mêmes propriétés », en particulier *elles représentent les mêmes éléments de* K. On note $\Phi \sim \Phi'$ (resp. $(E, \Phi) \sim (E', \Phi')$) l'équivalence de deux formes quadratiques (resp. l'existence d'un isomorphisme entre deux espaces quadratiques).

Suivant une méthode initiée par Gauß, pour étudier un problème sur une forme quadratique, on se ramène souvent, par équivalence, à un cas où cette étude est plus simple.

Soient (E, Φ) un espace quadratique et $B_\Phi : E^2 \to K$ la forme bilinéaire associée (resp., en caractéristique différente de 2, $\varphi = E^2 \to K$ la forme polaire de Φ). On dit que $x, y \in E$ sont Φ-*orthogonaux* (ou plus simplement orthogonaux) si $B_\Phi(x, y) = 0$ (resp. $\varphi(x, y) = 0$). On note alors $x \perp y$. Si $F \subset E$ est un sous-espace, son *orthogonal* est le sous-espace $F^\perp := \{x \in E \mid x \perp F\}$ ($x \perp F$ voulant dire $x \perp y$ pour tout $y \in F$).
Les propriétés de l'orthogonalité sont donc, par définition, celles d'une forme bilinéaire. Elles sont indépendantes de la caractéristique. Si la caractéristique est différente de 2, on utilise φ et les résultats du livre [L2]. Sinon, on utilise B_Φ, qui est alors *alternée* et on peut appliquer les résultats sur les formes bilinéaires alternées de la partie 1.

Soit (E, Φ) un espace quadratique. On dit qu'il est *non dégénéré* si la forme bilinéaire associée à Φ est non dégénérée, c'est-à-dire si $E^\perp = (0)$. On dit alors aussi que la forme quadratique Φ est non dégénérée.
Un vecteur $x \in E$ et dit *isotrope*[13], pour Φ, si $\Phi(x) = 0$ (*en caractéristique impaire*, c'est équivalent à $x \perp x$). Un sous-espace $F \subset E$ est dit *isotrope*[14], pour Φ, s'il contient *au moins un* vecteur isotrope non nul. Il est dit *totalement isotrope* si *tous* ses éléments sont isotropes. Un sous-espace totalement isotrope est dit *maximal* s'il n'est pas strictement contenu dans un autre sous-espace totalement isotrope.
Un sous-espace est dit *défini* (ou *anisotrope*) s'il n'est pas isotrope, c'est-à-dire si le seul vecteur isotrope est 0. On emploie la même terminologie pour une forme quadratique.
Attention, en caractéristique 2, une forme quadratique peut-être dégénérée *et* définie. Par exemple la forme quadratique $\Phi : x \mapsto x^2$ sur K est dégénérée et définie.
Le noyau d'une forme quadratique Φ est, par définition :

$$\operatorname{Ker} \Phi := \{x \in \operatorname{Ker} B_\Phi \mid \Phi(x) = 0\}.$$

On vérifie facilement que c'est un sous-espace.
En caractéristique impaire, $\operatorname{Ker} \Phi = \operatorname{Ker} \varphi = E^\perp$, comme dans [L2] I.3.3. En caractéristique 2, l'inclusion $\operatorname{Ker} \Phi \subset \operatorname{Ker} B_\Phi$ peut être *stricte*.
Le rang r de Φ est, par définition $r = \dim E - \dim \operatorname{Ker} \Phi$. Il est supérieur ou égal au rang de B_Φ. En caractéristique impaire il y a égalité et le rang peut être calculé en utilisant la matrice de Φ dans une base.
Supposons la caractéristique différente de 2. Alors $\operatorname{Ker} \Phi = \operatorname{Ker} \varphi$ est totalement isotrope et Φ est non dégénérée si, et seulement si, $\operatorname{Ker} \Phi = (0)$, c'est-à-dire si, et seulement si, son rang est égal à $\dim E$. La forme quadratique Φ induit une forme quadratique *non dégénérée* $\widetilde{\Phi}$ sur $\widetilde{E} := E / \operatorname{Ker} \Phi$,. Si F est un supplémentaire de $\operatorname{Ker} \Phi$ dans E, alors c'est un *supplémentaire orthogonal* : $E = F \overset{\perp}{\oplus} \operatorname{Ker} \Phi$ et le passage au quotient induit un isomorphisme d'espaces quadratiques entre $(F, \Phi_{|F})$ et $(\widetilde{E}, \widetilde{\Phi})$. L'étude de Φ se ramène donc à l'étude de la forme quadratique non dégénérée associée $\widetilde{\Phi}$, c'est-à-dire à celle de $\Phi_{|F}$.

[13] *Attention !* En caractéristique 2, tous les vecteurs sont isotropes pour B_Φ mais pas nécessairement pour Φ.
[14] C'est la terminologie la plus courante actuellement. Attention, des ouvrages plus anciens utilisent une terminologie différente et certains auteurs appellent « espace isotrope » ce que nous appelons « espace singulier ».

En caractéristique 2 la restriction de Φ à $\operatorname{Ker} B_\Phi$ peut être non triviale (elle est diagonalisable) et la question est plus difficile.

Soit (E, Φ) un espace quadratique *non-dégénéré*. Soient $F, G \subset E$ des sous-espaces. On rappelle les résultats suivants :

- $(F^\perp)^\perp = F$;
- $\dim F + \dim F^\perp = \dim E$;
- si $F \subset G$, alors $G^\perp \subset F^\perp$;
- $(F \cap G)^\perp = F^\perp + G^\perp$, $(F + G)^\perp = F^\perp \cap G^\perp$.

Ils sont similaires à ceux rappelés plus haut pour les espaces symplectiques.

Définition 5. Soit (E, Φ) un espace quadratique. Un sous-espace $F \subset E$ est dit *régulier* si la restriction de Φ à F est une forme quadratique non dégénérée ou si $F = (0)$.

Un sous-espace qui n'est pas régulier est dit *singulier*. Attention, certains auteurs appellent « non isotrope » un sous-espace régulier et « isotrope » un sous-espace singulier. Pour nous un sous-espace régulier peut être isotrope.

On dit que deux sous-espaces $F, F' \subset E$ sont *en somme directe orthogonale* si $F \cap F' = (0)$ et $F \perp F'$. On note alors $F \overset{\perp}{\oplus} F'$. On dit que ce sont des supplémentaires orthogonaux si $E = F \overset{\perp}{\oplus} F'$. Si Φ est non-dégénérée, on a alors : $F' = \mathbb{F}^\perp$ et $F = (F')^\perp$.

Les définitions s'étendent de façon évidente à une famille finie de sous-espaces.

Comme plus haut dans le cas symplectique, nous utiliserons souvent le résultat suivant (*cf.* [L2], I.3, 3.1.1, proposition 39). Le lecteur vérifiera qu'il reste vrai en caractéristique 2.

Proposition 20. Soient (E, Φ) un espace quadratique et $F \subset E$ un sous-espace.

(i) Le sous-espace F est *régulier* si, et seulement si, $F \cap F^\perp = (0)$.

(ii) Le sous-espace F est *régulier* si, et seulement si, $E = F \oplus F^\perp$.

(iii) Supposons que Φ est *non dégénérée*. Alors F est régulier si, et seulement si, $F^\perp$ est régulier.

Si $F \subset E$ est régulier, alors $E = F \overset{\perp}{\oplus} F^\perp$. Un sous-espace F est *singulier* si, et seulement si, $F \cap F^\perp \neq (0)$.

On rappelle que deux matrices $M, M' \in S_n(K)$ sont dites *congruentes* s'il existe une matrice $P \in M_n(K)$ inversible telle que $M' =^t PMP$. La congruence est une relation d'équivalence.

Soit (E, Φ) un espace quadratique (en caractéristique différente de 2). La matrice de Φ dans une base $(e_1, \ldots, e_n)$ de E est, par définition, $(a_{i,j} := \varphi(e_i, e_j))$. Elle est symétrique. Les matrices de Φ dans deux bases de E sont *congruentes*.

Soit $\mathcal{B} = (e_1, \ldots, e_m)$ une base donnée de E. En caractéristique impaire, la connaissance de Φ est équivalente à celle de la matrice de B_Φ dans cette base.

Ce n'est plus du tout le cas en caractéristique 2 : ce qui « manque » est la donnée des $\Phi(e_i)$ ($i \in [\![1, m]\!]$).

En caractéristique différente de 2, il existe toujours une *base Φ-orthonormée* (e_i) (*cf.* [L2], I.3, 3.1.2, théorème 44). La matrice de Φ dans cette base (qui est par définition celle de φ) est *diagonale* : $\mathrm{Diag}(a_1, \dots a_n)$, avec $a_i = \Phi(e_i)$. La forme quadratique Φ est *équivalente*[15] à la forme quadratique $(x_1, \dots, x_n) \mapsto a_1x_1^2 + \cdots + a_nx_n^2$ sur K^n. Si Φ n'est pas identiquement nulle, on peut toujours choisir la base orthonormée de telle sorte que $a_1 = 1$. La forme Φ est non dégénérée si, et seulement si, $a_1 \dots a_n \neq 0$. Le rang de Φ est le nombre de a_i non nuls. Les matrices de Φ dans deux bases orthonormées sont des matrices diagonales congruentes. Si l'on permute les éléments diagonaux de $\mathrm{Diag}(a_1, \dots a_n)$, on obtient une matrice diagonale congruente. Si $b_1, \dots, b_n$ sont des *carrés* non nuls de K, alors les matrices diagonales $\mathrm{Diag}(a_1, \dots, a_n)$ et $\mathrm{Diag}(b_1a_1, \dots, b_na_n)$ sont congruentes. En caractéristique impaire, classifier les formes quadratiques sur K est équivalent à classifier les matrices diagonales à coefficients dans K, modulo congruence.
Notation. Comme beaucoup d'auteurs, on notera souvent, avec un abus de notation :
$\Phi := (a_1, \dots, a_n)$ la forme quadratique Φ sur K^n dont la matrice dans la base canonique est $\mathrm{Diag}(a_1, \dots, a_n)$, c'est-à-dire $\Phi : (x_1, \dots x_n) \mapsto a_1x_1^2 + \cdots + a_nx_n^2$.

En caractéristique 2, l'existence d'une base orthogonale équivaut à $B_\Phi = (0)$. Si c'est le cas, toute base est orthogonale. S'il existe une base dans laquelle Φ s'écrit $x \mapsto \sum a_ix_i^2$, cette base est orthogonale. En pratique, dans le cas non dégénéré, on remplace l'utilisation des bases orthogonales par celle des *bases symplectiques* (celles dont la matrice est J_n). On écrit ainsi par exemple toute forme quadratique non dégénérée de rang 2 sous la forme $(x, y) \mapsto bx^2 + xy + cy^2$. La matrice associée est J_1.

2.2 Discriminant d'une forme quadratique

L'ensemble des carrés de K^* (c'est-à-dire $\{x^2 \mid x \in K^*\}$) est l'image de l'homomorphisme de groupes multiplicatifs $x \in \mathbb{K}^* \mapsto x^2 \in \mathbb{K}^*$. Il forme un *sous-groupe* du groupe multiplicatif K^*. On le note $(K^*)^2$ et l'on considère le *groupe quotient* : $K^*/(K^*)^2$. La classification des formes quadratiques sur un K-espace vectoriel de dimension 1 se ramène à l'étude de $K^*/(K^*)^2$.

Exemples.

1. Dans $\mathbb{C}$ tout élément est un carré et si $x^2 = 0$, alors $x = 0$, donc $\mathbb{C}^*/(\mathbb{C}^*)^2 = (1)$. Plus généralement, si K est un corps algébriquement clos, alors $K^*/(K^*)^2 = \{1\}$.
2. Dans $\mathbb{R}$ les carrés non nuls sont les éléments strictement positifs et un élément non nul qui n'est pas un carré est strictement négatif, donc $\mathbb{R}^*/(\mathbb{R}^*)^2 = (\pm 1)$: l'image de $x \in \mathbb{R}^*$ dans $\mathbb{R}^*/(\mathbb{R}^*)^2$ est interprétée comme *le signe* de x.
3. Si K est un *corps fini* (de caractéristique impaire), alors $K/(K^*)^2$ est aussi un groupe à deux éléments, noté $\{\pm 1\}$: 1 pour les images des carrés et -1 pour celles des éléments qui ne sont pas des carrés. Voir I.1, 3.

[15] En caractéristique 2 une forme quadratique n'est pas en général diagonalisable.

4. Soit $K := \mathbb{Q}$. Montrons que $K/(K^*)^2$ a *une infinité d'éléments*. Soient p et p' deux nombres premiers distincts. Leurs images dans $\mathbb{Q}/(\mathbb{Q}^*)^2$ sont distinctes. Supposons le contraire. Alors p/p' est un carré de $\mathbb{Q}^*$: $p/p' = (a/b)^2$, avec $a, b \in \mathbb{N}^*$ et premiers entre eux. On a $b^2p = a^2p'$. Puisque p divise a^2p' et qu'il est premier avec p', il divise a^2, donc a : $a = pa'$. On a $b^2 = (a')^2pp'$ et p divise donc b^2, donc b. Ainsi p est un facteur commun de a et b et l'on a abouti à une contradiction. Pour plus de détails sur $\mathbb{Q}/(\mathbb{Q}^*)^2$, *cf.* l'exercice I.2.12 de la page 122.
5. Si K est un corps p-adique (*cf.* le module III.3) : $K = \mathbb{Q}_p$, on montre que $K/(K^*)^2$ a un nombre fini d'éléments : 4 si le nombre premier p est impair et 8 si $p = 2$ (*cf.* les exercices III.3.28 de la page 526 et III.3.29 de la page 527).
6. Si K est un corps fini de caractéristique 2, l'application $x \in K \mapsto x^2 \in K$ est un *isomorphisme*, donc $K/(K^*)^2$ a *un seul* élément.

En caractéristique impaire, le groupe $K^*/(K^*)^2$ est utile pour étudier l'équivalence des formes quadratiques, grâce à la notion de *discriminant* d'une forme quadratique que nous allons définir. Notons que $K^*/(K^*)^2$ représente les classes de congruence de matrices carrées inversibles $(1,1)$ à coefficient dans K.

Proposition et définition 21.
Soient K un corps de caractéristique différente de 2 et $n \geqslant 1$.

(i) Soit $A \in S_n(K)$ une matrice symétrique inversible. On appelle *discriminant* de A et l'on note $\delta(A)$ l'image de $\det A$ dans $K^*/(K^*)^2$. Le discriminant ne dépend que de la classe de congruence de A.

(ii) Soient E un K-espace vectoriel de dimension n et $\Phi \in Q(E)$ une forme quadratique non dégénérée. Si $\mathcal{B}$ est une base de E, le discriminant de la matrice de Φ dans la base $\mathcal{B}$ ne dépend pas du choix de $\mathcal{B}$, on l'appelle *discriminant* de Φ et l'on note $\delta(\Phi) := \delta\left(\mathrm{Mat}(\Phi, \mathcal{B})\right) \in K^*/(K^*)^2$.

Démonstration.

(i) Soient $A, B \in S_n(K)$. Si A et B sont congruentes, on a $B = PA^tP$, avec $P \in GL_n(K)$). Alors :

$$\det B = \det A(\det P)^2 \quad \text{et} \quad (\det P)^2 \in (K^*)^2,$$

donc, si A et B sont inversibles $\delta(A) = \delta(B)$.

(ii) Soient $\Phi \in Q(E)$ non dégénérée et $\mathcal{B}$, $\mathcal{C}$ deux bases de E. Alors les matrices respectives de Φ dans ces bases sont inversibles et congruentes, donc $\delta(A) = \delta(B)$. ■

En utilisant la proposition 43 de [L2] I.3, 3.1.2, on obtient le critère *nécessaire* suivant pour l'équivalence de deux formes quadratiques. Il n'est pas, en général, suffisant.

Proposition 22. En caractéristique différente de 2, deux formes quadratiques équivalentes ont des discriminants égaux.

Le discriminant est un *invariant* pour la relation d'équivalence entre formes quadratiques.

Exercice 4.

Soient $K := \mathbb{Q}$ et $r \in \mathbb{N}^*$. Montrer qu'il existe *une infinité* de formes quadratiques non dégénérées de rang r qui ne sont pas deux à deux équivalentes.

Solution. A tout nombre premier p on associe la forme quadratique $\Phi_p := (1, \ldots, 1, p)$ sur $\mathbb{Q}^r$. Son discriminant est l'image $\tilde{p}$ de p dans $\mathbb{Q}/(\mathbb{Q}^*)^2$. Nous avons vu que si p et p' sont deux nombres premiers distincts, alors $\tilde{p} \neq \tilde{p}'$ (*cf.* les exemples de la page 93).
Donc Φ_p et $\Phi_{p'}$ ne sont pas équivalentes.

Exemples.

1. Soient $a, b, c \in K$ et Φ définie par $\Phi : (x, y) \mapsto ax^2 + bxy + cy^2$. Sa matrice dans la base canonique est $A := \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix}$. Le discriminant est l'image de $\det A = ac - b^2/4$ dans $K^*/(K^*)^2$. Cet exemple est à l'origine de la terminologie[16]. Si $(x, y) = (t, 1)$, $\Phi(x, y) = at^2 + bt + c$. Supposons $ac \neq 0$. Si l'équation du second degré $at^2 + bt + c = 0$ a une racine $t_0 \in K$, c'est-à-dire si, et seulement si, $b^2 - 4ac$ est un carré de K, alors Φ a un vecteur isotrope non nul : $(1, t_0)$. Inversement, si (x, y) est un vecteur isotrope non nul, alors $x \neq 0$ ou $y \neq 0$. Supposons $y \neq 0$, alors $(t_1 := xy^{-1}, 1)$ est isotrope, donc $at_1^2 + bt_1 + c = 0$ et $D := b^2 - 4ac$ est un carré de K. On peut raisonner de façon analogue si $x \neq 0$.
   Si $D = b^2 - 4ac$ est un carré non nul, l'image de $-D/4 = ac - b^2/4$ dans $K^*/(K^*)^2$, c'est-à-dire $\delta(\Phi)$ est l'image de $-1 \in K$.
2. Soient $n \in \mathbb{N}^*$ et $A := \mathrm{Diag}(a_1, \ldots, a_n) \in S_n(K)$ (avec $a_1, \ldots, a_n \neq 0$). Alors $\delta(A)$ est l'image du produit $a_1, \ldots, a_n$ dans $K^*/(K^*)^2$. Si A est congruente à la matrice identité I, alors $a_1, \ldots, a_n$ est un carré de K^*. Mais cette condition nécessaire n'est pas suffisante : si $K := \mathbb{R}$, la matrice $A := -I_2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ n'est pas congruente à la matrice identité I (la signature de A est $(0, 2)$) bien que l'on ait $\delta(A) = (-1)^2 = 1$.
3. Si $K = \mathbb{R}$, les matrices $B := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ et I_2 ne sont pas congruentes. En effet $\det B = -1$, $\det I = 1$ et $-1 \neq 1$ dans $\mathbb{R}^*/(\mathbb{R}^*)^2$.

Remarque.

Supposons que $K := \mathbb{R}$ et revenons sur l'exemple 1 ci-dessus. On considère l'ensemble $\mathcal{Q}$ des formes quadratiques réelles sur $\mathbb{R}^2$, dites *binaires* : $[a, b, c] : (x, y) \mapsto ax^2 + bxy + cy^2$.

Une matrice $P = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in GL_2(\mathbb{R})$ transforme $[a, b, c]$ en une forme équivalente : $[a, b, c]^P = [a', b', c']$ définie par $a'x^2 + b'xy + c'^2 = ax'^2 + bx'y' + cy'^2$, où $x' = \alpha x + \beta y$

[16] Le discriminant (au sens élémentaire) de l'équation du second degré $at^2 + bt + c = 0$ est $b^2 - 4ac$.

et $y' = \gamma x + \delta y$. En général, le discriminant « au sens du lycée » $D = b^2 - 4ac$ n'est pas conservé. Il est multiplié par $(\det P)^2$ Pour le conserver, on va se limiter à l'action du sous-groupe $SL_2(\mathbb{R})$. De plus on se limite au cas des formes *définies positives*, c'est-à-dire telles que $D < 0$. On note $\mathcal{Q}^+$ l'ensemble de ces formes. Il est invariant par équivalence.

La connaissance d'une forme quadratique $[a, b, c] \in \mathcal{Q}^+$ est équivalente à celle des racines dans $\mathbb{C}$ de $at^2 + bt + c$. Elles sont conjuguées et il suffit de connaître celle dont la partie imaginaire est > 0. On obtient ainsi une application $\xi : \mathcal{Q}^+ \to \mathcal{H}$, où $\mathcal{H} := \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$ est le *demi-plan de Poincaré*, définie par $\xi([a, b, c]) := -\frac{b}{2a} + i\frac{\sqrt{-D}}{2a}$. Cette application est *bijective* : si $[a, b, c] := [1, -2x, x^2 + y^2]$, on a $\xi([a, b, c]) = x + iy$. Par ailleurs, nous savons que l'on a une action de $SL_2(\mathbb{R})$ sur $\mathcal{H}$ qui permet d'interpréter $SL_2(\mathbb{R})$ comme un *groupe de déplacements en géométrie hyperbolique* (*cf.* [L2] I.6, 3.2.2). On peut montrer que ξ est compatible avec les actions de $SL_2(\mathbb{R})$ à la source (à droite) et au but (à gauche), on dit qu'elle est *équivariante* (*cf.* l'exercice I.2.15 de la page 122). Ceci permet d'étudier les classes d'équivalence de formes de $\mathcal{Q}^+$ (en particulier de formes *entières*) modulo $SL_2(\mathbb{Z})$ en utilisant le demi-plan de Poincaré.

Exercice 5.

Soient (E, Φ) un plan quadratique non dégénéré et $d \in K^*$ un représentant du discriminant $\delta(\Phi) \in K^*/(K^*)^2$.

1. Soit a représenté par Φ. Montrer que la matrice de Φ (dans une base arbitrairement choisie) est congruente à $\operatorname{Diag}(a, ad)$.
2. On suppose que $-1 \in K^*$ est un représentant du discriminant $\delta(\Phi) \in K^*/(K^*)^2$. Montrer que Φ est équivalente à la forme quadratique $(1, -1)$ (autrement dit que (Φ, E) est un plan hyperbolique, *cf.* la définition 6 ci-dessous).
3. Montrer qu'une forme quadratique Φ_1 sur E est équivalente à Φ si, et seulement si, $\delta(\Phi_1) = \delta(\Phi)$ et si Φ et Φ_1 représentent un même élément de K^*.

Solution.

1. Soit $v \in E$, non nul, tel que $\Phi(v) = a$. On peut prendre v comme premier vecteur d'une base Φ-orthogonale, il existe donc $b \in K$ tel que la matrice de Φ est congruente à $\operatorname{Diag}(a, b)$. Alors ab et d sont deux représentants de $\delta(\Phi)$: $ab = d \mod (K^*)^2$. Il existe donc $c \in K^*$ tel que $d = abc^2$. On a :
$$\operatorname{Diag}(a, b) = \operatorname{Diag}\left(a, \frac{d}{ac^2}\right) = \operatorname{Diag}\left(a, ad\left(\frac{1}{ac}\right)^2\right)$$
et la dernière matrice est congruente à $\operatorname{Diag}(a, ad)$.
2. Supposons que $d = -1$ soit un représentant $\delta(\Phi)$. En utilisant la question 1, on montre que la matrice de Φ est congruente à la matrice $\operatorname{Diag}(a, -a)$. On termine en montrant que les matrices $\operatorname{Diag}(a, -a)$ et $\operatorname{Diag}(1, -1)$ sont congruentes, ce qui n'est pas si évident... (*cf.* [L2], I.3, 3.1.2, exercice 18, ou utiliser les propriétés des plans hyperboliques).
3. Les deux conditions sont évidemment nécessaires. Inversement, notons $d \in \mathbb{K}^*$ un représentant de $\delta(\Phi_1) = \delta(\Phi)$. D'après la question 1 les matrices de Φ et Φ_1 dans une

même base sont toutes deux congruentes à $\mathrm{Diag}(a, ad)$. Elles sont donc congruentes et $\Phi \sim \Phi_1$.

Remarque. Comparons les plans quadratiques non-dégénérés en caractéristique impaire et en caractéristique 2. Soit (E, Φ) un tel plan.

- En caractéristique impaire, Φ est diagonalisable, elle s'écrit dans une base orthogonale $(x, y) \mapsto bx^2 + cy^2$, avec $bc \neq 0$. L'image de bc dans $K^*/(K^*)^2$ est le *discriminant*. C'est un invariant pour l'équivalence : si (b, c) et (b', c') sont équivalentes, alors bc et $b'c'$ ont la même image dans $K^*/(K^*)^2$.
- En caractéristique 2, Φ s'écrit dans une base symplectique :

$$\Phi_{b,c} : (x, y) \mapsto bx^2 + xy + cy^2, \qquad \text{avec} \qquad b, c \in K.$$

  On remplace l'homomorphisme de groupes *multiplicatifs* $x \in K^* \mapsto x^2 \in K^*$ par l'homomorphisme de groupes *additifs* $\sigma : t \in K \mapsto t^2 + t \in K$, appelé *application d'Artin-Schreier*. On remplace le quotient $K^*/(K^*)^2$ par le quotient $K/\sigma(K)$. On montre que si les formes quadratiques $\Phi_{b,c}$ et $\Phi_{b',c'}$ sont équivalentes, alors bc et $b'c'$ ont la même image dans $K/\sigma(K)$ (*cf.* l'exercice I.2.9 de la page 120). On appelle cette image *l'invariant de Arf*. Il « remplace » le discriminant.

2.3 Classification des formes quadratiques sur un corps fini F_q

Soient p un nombre premier, $m \in \mathbb{N}^*$ et $q := p^m$. Soit $K = F_q$ le corps fini à q éléments (on admet qu'il existe un tel corps, unique à isomorphisme près, *cf.* par exemple [Perrin], [Serre] ou [MPA1]).
On rappelle que, si p est *impair*, le goupe quotient $F_q^*/(F_q^*)^2$ a deux éléments (cf les exemples de la page 93). On le note $F_q^*/(F_q^*)^2 = \{\pm 1\}$. Dans tout ce paragraphe (sauf dans le lemme 23) le nombre premier p est supposé impair.

Il faut manipuler la notation $F_q^*/(F_q^*)^2 = \{\pm 1\}$ avec précaution. Pour certaines valeurs de q, l'image de -1 dans F_q peut être un carré et alors l'image de -1 dans $F_q^*/(F_q^*)^2 = \{\pm 1\}$ est 1. Par exemple, pour $q = p = 5$, on a $-\bar{1} = \bar{4} = (\bar{2})^2$.

Lemme 23. Soit $K = F_q$ un corps fini (sans restriction sur la caractéristique). Soient $a, b, c \in F_q^*$. Alors, il existe $x, y \in F_q$ tels que $ax^2 + by^2 = c$.

Démonstration. On définit deux sous-ensembles $A, B \subset F_q$ par :

$$A := \{ax^2 \mid x \in F_q\} \quad \text{et} \quad B := \{c - by^2 \mid y \in F_q\}$$

Le cardinal de l'ensemble $x^2 \mid x \in F_q\}$ est $1 + \frac{q-1}{2} = \frac{q+1}{2}$, donc $\mathrm{card}\ A = \mathrm{card}\ B = \frac{q-1}{2}$.
Par suite $\mathrm{card}\ A + \mathrm{card}\ B = q + 1 > q = \mathrm{card}\ F_q$, donc $A \cap B \neq \varnothing$.
Il existe donc $x, y \in F_q$ tels que $ax^2 = c - by^2$, c'est-à-dire $ax^2 + by^2 = c$. ∎

Proposition 24. Soit $q = p^m$, où p est un nombre premier impair. Soit (E, Φ) une forme quadratique *non dégénérée* sur $K = F_q$, avec $\dim E = 2$. Elle représente tout élément de F_q^*.

Démonstration.
En utilisant une base de E on se ramène au cas de $\Phi : (x, y) \to ax^2 + bx^2$, avec $a, b \in F_q^*$. On utilise ensuite le lemme 23 ci-dessus. ■

Corollaire 25. Soit $q = p^m$, où p est un nombre premier impair. Soit (E, Φ) une forme quadratique *non dégénérée* sur $K = F_q$, avec $\dim E \geqslant 2$. Elle représente tout élément de F_q^*.

Démonstration. En utilisant une base orthonormée de E, on se ramène au cas où $E = K^n$ et $\Phi = (a_1, \ldots, a_n)$. Soit $c \in F_q^*$. D'après le lemme 23, il existe $x_1, x_2 \in F_q$ tels que $a_1x_1^2 + a_2x_2^2 = c$. On pose $x_3 = \ldots = x_n = 0$. On a alors $\Phi(x_1, x_2, \ldots, x_n) = c$. ■

Exercice 6.
Soient $a \in F_q^*$ (en caractéristique impaire) et $\Phi = (a, a)$ montrer que $\Phi \sim (1, 1)$.
Solution. D'après la proposition 24 ci-dessus, Φ représente $1 \in F_q^*$. Soit $u \in E = \mathbb{F}_q^2$ tel que $\Phi(u) = 1$. En utilisant $u^\perp$, on montre qu'il existe $d \in F_q^*$ tel que $\Phi \sim (1, d)$. Le discriminant de Φ est la classe de a^2, il est donc égal à 1 et d est un carré.
Donc $\Phi \sim (1, 1)$.

Théorème 26. Soit $q = p^m$, où p est un nombre premier *impair*. Soit (E, Φ) une forme quadratique *non dégénérée* sur $K = F_q$. Soit $d \in F_q^*$ dont la classe dans $K^*/(K^*)^2$ est le discriminant $\delta(\Phi)$ de Φ. Alors $\Phi \sim (1, \ldots, 1, d)$.

Démonstration. On procède par récurrence sur la dimension $n := \dim E$.
Le résultat est évidemment vrai pour $n = 1$. Supposons $n \geqslant 2$ et que le résultat est vrai pour une forme quadratique non dégénérée sur un espace de dimension $n - 1$. D'après le corollaire 25, Φ représente $1 \in F_q^*$. Soit $x \in E$ tel que $\Phi(x) = 1$. Notons $E' = x^\perp$ et Φ' la restriction de Φ à E'. Elle n'est pas dégénérée (Kx est régulier), donc $E' = x^\perp$ est régulier. On a $E = Kx \overset{\perp}{\oplus} E'$, $\dim E' = n - 1$ et $\delta(\Phi) = \delta(\Phi')$. D'après l'hypothèse de récurrence on a $\Phi' \sim (1, \ldots, 1, d)$, donc $\Phi \sim (1, 1, \ldots, 1, d)$. ■

Il y a donc exactement *deux classes d'équivalence* de formes quadratiques non dégénérées de rang n sur F_p : la classe de $(1, \ldots, 1)$ et celle de $(1, \ldots, 1, a)$, où a n'est pas un carré dans F_q. Si $n = 2$, l'une de ces classes est formée de formes quadratiques définies, l'autre de formes quadratiques isotropes : $(1, 1)$ est définie si, et seulement si, (-1) n'est pas un carré de F_q (*cf.* l'exercice I.2.13 de la page 122).

Corollaire 27. Soit $q = p^m$, $p \neq 2$. Pour que deux formes quadratiques non dégénérées sur F_q soient équivalentes, il faut et il suffit qu'elles aient même rang et même discriminant.

Exercice 7.
Soit E un F_q-espace vectoriel de dimension finie n ($q = p^m$, $p \neq 2$). Montrer que si $n \geqslant 3$, alors toute forme quadratique Φ sur E possède un vecteur isotrope non nul.
Solution. Il suffit évidemment de prouver le résultat pour $n = 3$ et l'on peut supposer que Φ est non dégénérée (sinon le résultat est évident). D'après le théorème 26 ci-dessus, Φ est équivalente à $\Psi := (1, 1, c)$, avec $c \in K^*$. D'après la proposition 24, il existe $x, y \in K$ tels que $x^2 + y^2 = -c$. Alors $\Psi(x, y, 1) = x^2 + y^2 + c = 0$: $(x, y, 1)$ est un vecteur isotrope et il n'est pas nul.

La version géométrique du résultat de l'exercice ci-dessus est : toute conique projective sur un corps fini a un point rationnel.

Remarques.

1. Le théorème 26 ne s'applique que si le corps *fini* K est de caractéristique impaire. En caractéristique 2, on connaît aussi la classification des formes quadratiques non dégénérées sur un corps fini. Le résultat est assez similaire. On remplace le discriminant par l'invariant de Arf (*cf.* les remarques page 89 et page 97). Celui-ci appartient au groupe quotient $K/\sigma(K)$, où $\sigma : (K, +) \to (K, +)$ est l'homomorphisme d'Artin-Schreier ($\sigma : t \mapsto t^2 + t$). On montre facilement que le cardinal du quotient $K/\sigma(K)$ est *deux*. Il y a deux classes d'équivalence correspondantes : celles de $(x, y) \mapsto xy$ et de $(x, y) \mapsto x^2 + xy + ay^2$, avec $a \notin K/\sigma(K)$. Voir les exercices I.2.4 et I.2.9.
2. Le résultat de l'exercice 7 ci-dessus reste vrai en caractéristique 2. Une preuve est proposée dans l'exercice I.2.10 de la page 121. On peut aussi le déduire du théorème de Chevalley (*cf.* [Serre, I.2.1, 2.2]).

2.4 Espaces hyperboliques Théorème de décomposition de Witt

Dans toute cette partie et les suivantes, le corps K est *toujours* supposé de caractéristique différente de 2.
La théorie de Witt[17] *ne dépend pas des propriétés particulières du corps de base* K (à la restriction sur la caractéristique près). Elle est utile pour étudier les propriétés du groupe des isométries d'une forme quadratique. Elle permet aussi de ramener la classification des formes quadratiques à celle des formes quadratiques *définies*. Mais cette dernière, par contre, *dépend des propriétés du corps* K et « le plus dur reste à faire » ...

2.4.1 Plan hyperbolique. Espaces hyperboliques.

Rappelons la définition d'un plan hyperbolique.

Définition 6. Un plan hyperbolique est un espace quadratique *non dégénéré* (E, Φ) de dimension deux et *isotrope*.

[17] Le mathématicien allemand Ernst Witt a été l'un des premiers à étudier les formes quadratiques sur un corps *arbitraire*, en 1937.

Autrement dit, un plan quadratique non dégénéré (E, Φ) est hyperbolique si, et seulement si, Φ représente 0.
Nous avons montré dans le livre [L2] que *tous les plans quadratiques sont équivalents* (*cf.* I.3, 3.1.2, exemple de la page 134).

Proposition 28. Soit $(E; \Phi)$ un espace quadratique de dimension deux, non dégénéré. C'est un plan hyperbolique si, et seulement s'il est équivalent à un plan hyperbolique.

Exercice 8.
Soit $a \in K^*$. Montrer que la forme quadratique $(1, -a)$ est hyperbolique si, et seulement si, a est un carré de K^*.
Solution.
Si $a = b^2$, on a : $1.b^2 - a.1 = 0$, donc $(b, 1)$ est un vecteur isotrope non nul. Inversement, si $(1, -a)$ est hyperbolique, il existe un vecteur isotrope non nul (x_0, y_0) : $x_0^2 - ay_0^2 = 0$. On ne peut pas avoir $y_0 = 0$ (sinon l'on aurait $x_0 = 0$), donc $a = (x_0/y_0)^2 \in (K^*)^2$.

Rappelons que toute forme quadratique non dégénérée de rang 2 est équivalente à une forme quadratique $(1, b)$, avec $b \neq 0$. Elle est donc hyperbolique si, et seulement si, $-b$ est un carré de K^*. Il y a de nombreuses caractérisations des plans hyperboliques.

Proposition 29. Soit (E, Φ) un espace quadratique de dimension 2, non dégénéré. Les conditions suivantes sont équivalentes :

(i) (E, Φ) est un plan hyperbolique ;

(ii) Φ est équivalente à la forme quadratique $\Phi : (x, y) \mapsto xy$ sur $E = \mathbb{K}^2$;

(iii) pour $b \in \mathbb{K}^*$, Φ est équivalente à la forme quadratique $(x, y) \mapsto bxy$ sur $E = \mathbb{K}^2$;

(iv) Φ est équivalente à la forme quadratique $(1, -1)$;

(v) pour $a \in \mathbb{K}^*$, (E, Φ) est équivalente à la forme quadratique $(a, -a)$;

(vi) il existe une base de E dans laquelle la matrice de Φ est $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

(vii) Φ est équivalente à la forme quadratique $(1, -c)$, où c est un carré de K^*.

(viii) E admet une base formée de deux éléments isotropes ;

(ix) le nombre de droites isotropes de (E, Φ) est *exactement* deux.

Démonstration. Les formes quadratiques de (ii), (iii), (iv), (v), (vi), (vii) sont toutes non dénénérées et isotropes. Donc, d'après la proposition 28 ci-dessus, les conditions (i), (ii), (iii), (iv), (v) et (vi), (vii) sont équivalentes.
Si Φ vérifie (viii), puisqu'elle n'est pas dégénérée, c'est un plan hyperbolique. Inversement, si (E, Φ) est hyperbolique, elle est équivalente à $\Psi : (x, y) \mapsto xy$ sur $E = \mathbb{K}^2$. Si $(\varepsilon_1, \varepsilon_2)$ est la base canonique de K^2, ε_1 et ε_2 sont deux éléments isotropes de Ψ.
Par équivalence Φ vérifie (viii).

Si $b \in K^*$, le nombre de droites isotropes de $\Psi : (x, y) \mapsto bxy$ (sur $E = \mathbb{K}^2$) est exactement deux : $bxy = 0$ si, et seulement si, $x = 0$ ou $y = 0$. Inversement, on écrit Ke_1 et Ke_1 les droites isotropes. On a $\Phi(e_1+e_2) = \Phi(e_1)+\Phi(e_2)+2\varphi(e_1, e_2) = 2\varphi(e_1, e_2)$. Les deux droites isotropes sont distinctes, par suite $e_1+e_2 \neq 0$. On a $\Phi(e_1+e_2) \neq 0$, sinon la droite $K(e_1 + e_2)$ serait isotrope et différente de Ke_1 et Ke_1. Alors $2\varphi(e_1, e_2) \neq 0$ et $\varphi(e_1, e_2) \neq 0$ (la caractéristique de K est supposée différente de 2). On pose $b := \varphi(e_1, e_2)$. La forme Φ est alors équivalente à $(x, y) \mapsto bxy$. Ainsi (iii) et (ix) sont équivalentes. ∎

Si (E, Φ) est un plan hyperbolique, on dit aussi que Φ est hyperbolique.

Exemples.

1. Si $K = \mathbb{C}$, tous les plans quadratiques définis sont hyperboliques.
2. Si $K = \mathbb{R}$, une forme quadratique de rang 2 est hyperbolique si, et seulement si, sa signature est $(1, -1)$ ou $(-1, 1)$.
3. Si $K = F_q$, la forme quadratique $(1, 1)$ peut être hyperbolique. Par exemple si $K = F_5$, on a $\bar{2}^2+\bar{1}^2 = \bar{0}$. Par contre, cette forme n'est pas hyperbolique si $K = F_3$.

Le résultat suivant permet de « compléter » une droite isotrope d'un espace quadratique non dégénéré (E, Φ) en un plan hyperbolique, donc régulier. C'est un cas particulier de la méthode de *complétion régulière* que nous introduirons plus loin (*cf.* le corollaire 40 de la page 108).

Proposition 30. Soit $(E; \Phi)$ un espace quadratique non dégénéré. Soit $x \in E$ un vecteur *isotrope non nul*. Alors il existe un plan F contenant x tel que $(F, \Phi_{|_F})$ soit un plan hyperbolique.

Démonstration.

On note φ la forme polaire de Φ. Puisque Φ est non dégénérée, il existe $y \in E$ tel que $\alpha := \varphi(x, y) \neq 0$. Le vecteur y n'est pas colinéaire à x (sinon $\varphi(x, y) = \lambda\Phi(x) = 0$), donc $H = Kx + Ky$ est un plan. Le déterminant de la matrice de la restriction $\Phi_{|_H}$ dans la base (x, y) est $-\alpha^2 \neq 0$ donc $\Phi_{|_H}$ est non dégénérée. Ainsi $(H, \Phi_{|_H})$ est un plan hyperbolique. ∎

Point Méthode

Les droites non isotropes et les plans hyperboliques sont les « briques de base » pour l'étude des formes quadratiques. Si x est un élément isotrope non nul, on peut « compléter » la droite Kx en un plan hyperbolique.

Exercice 9.

1. Soit (E, Φ) un espace quadratique non dégénérée sur K tel que Φ représente 0. Montrer que Φ représente tout élément de K.
2. Soient (E, Ψ) un espace quadratique non dégénéré sur K et $a \in K^*$. On note (a) la forme quadratique $Z \mapsto aZ^2$ sur K.

Montrer que les trois conditions suivantes sont équivalentes.

(i) La forme Ψ représente a.

(ii) On a $\Psi = (a) \overset{\perp}{\oplus} \Psi'$.

(iii) La forme $\Theta := \Psi \overset{\perp}{\oplus} (-a)$ représente 0.

Solution.

1. Il existe un vecteur isotrope $x \in E$ non nul . On applique la proposition 30. On obtient un plan hyperbolique H. Alors $E = H \overset{\perp}{\oplus} H^{\perp}$. La restriction de Φ à H représente tout élément de K.
2. On peut supposer $E = K^n$. Si $\Psi(x) = a$, on a $E = Ka \overset{\perp}{\oplus} F$, avec $F := (Ka)^{\perp}$. Ainsi (i) $\Rightarrow$ (ii). L'implication (ii) $\Rightarrow$ (i) est évidente, ainsi que (ii) $\Rightarrow$ (iii). Supposons que Θ a un zéro non trivial $(x_1, \dots x_n, z)$. Si $z = 0$, alors Ψ représente 0 et, d'après 1, elle représente a. Si $z \neq 0$, alors $\Psi(x_1/z, \dots, x_n/z) = a$.
   Par suite (iii) $\Rightarrow$ (i).

Définition 7. Une somme directe orthogonale de plans hyperboliques est appelée *espace hyperbolique*.

Un espace hyperbolique est déterminé, à isomorphime près, par sa dimension, qui est nécessairement *paire*. Si celle-ci est $2n$, le nombre de termes de la somme est n.
Nous prouverons plus loin le résultat suivant.

Proposition 31. Soit (E, Φ) un espace hyperbolique de dimension $2n$. Alors tous les sous-espaces totalement isotropes maximaux de E ont la même dimension. Celle-ci est égale à n.

Les sous-espaces totalement isotropes maximaux sont les analogues quadratiques des sous-espaces Lagrangiens du cas alterné.

Exercice 10.

On suppose que (E, Φ) est un espace hyperbolique de dimension $2m$ ($m \in \mathbb{N}^*$).

(i) Montrer qu'il existe une base de E dans laquelle la matrice de Φ est : $\left(\begin{array}{c|c} 0 & I_m \\ \hline I_m & 0 \end{array}\right)$.

(ii) Montrer que E est somme directe de deux sous-espaces totalement isotropes maximaux.

Solution. On suppose que $E = E_1 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} E_m$ où, pour tout $i \in [\![1, s]\!]$, $(E_i, \Phi_{|_{E_i}})$ est un plan hyperbolique. On choisit dans chaque E_i deux vecteurs isotropes non nuls distincts v_i et v_i'. On note $G := Kv_1 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} Kv_m$ et $G' := Kv_1' \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} Kv_m'$; G et G' sont des sous-espaces totalement isotropes maximaux et l'on a $E = G \oplus G'$.
On a pour tout $i \in [\![1, m]\!]$, $\varphi(v_i, v_i') \neq 0$. Quitte à multiplier chaque v_i' par un scalaire convenable non nul, on peut donc supposer que, $\varphi(v_i, v_i') = 1$, d'où le résultat.

2.4.2 Théorème de décomposition de Witt

La décomposition de Wit est un analogue quadratique du travail de classification fait plus haut dans le cas alterné. Le cas quadratique est plus difficile à cause de l'existence de sous-espaces sur lesquels la forme quadratique est *définie*. La décomposition de Witt n'est pas une classification complète, elle nous ramène seulement à la classification des formes quadratiques *définies*.

Considérons d'abord le cas *réel* : $K := \mathbb{R}$, que nous avons étudié dans [L2], I.3, 4. Un plan quadratique réel est hyperbolique si, et seulement s'il est non dénénéré et si sa signature est $(1,1)$. Plus généralement, on vérifie facilement (nous allons le voir) qu'un espace quadratique réel est hyperbolique si, et seulement s'il est non dénénéré et si sa signature est de la forme (p,p). Cette remarque conduit à « traduire » le *théorème d'inertie de Sylvester* (*cf.* [L2], I.3, 4.2, théorème 63) sous une forme indépendante du choix d'une base et ne faisant pas intervenir le fait que le corps de base est $\mathbb{R}$. Toutefois, avec cette forme, si $\neq q$, l'information $p < q$ ou $p > q$ est perdue.

Théorème 32 (Décomposition de Witt sur $\mathbb{R}$).
Soient E un $\mathbb{R}$-espace vectoriel de dimension finie et Φ une forme quadratique sur E.

(i) Il existe une décomposition en somme directe orthogonale :

$$E = E_{hyper} \overset{\perp}{\oplus} E_{def} \overset{\perp}{\oplus} \operatorname{Ker}\Phi,$$

où E_{hyper} est un espace hyperbolique et E_{def} est tel que $\Phi_{|_{E_{def}}}$ est définie. Une telle décomposition est unique à équivalence près.

(ii) Si (p,q) est la signature de Φ, on a :

$$\dim E_{hyper} = 2\min\{p,q\} \quad \text{et} \quad \dim E_{def} = |p-q|.$$

La décomposition du théorème s'appelle *décomposition de Witt*. Notons que la connaissance de $\min\{a,b\}$ et de $|a-b|$ est équivalente à celle d'un couple *non ordonné* $\{a,b\}$. La décomposition de Witt ne permet pas de savoir si $(p,q) = (a,b)$ ou $(p,q) = (b,a)$. Elle ne dit pas si la forme $\Phi_{|_{E_{def}}}$ (si elle apparaît) est définie positive ($p > q$) ou définie négative ($p < q$). Le théorème de Sylvester est plus précis. Il donne la classification des formes définies en définies positives ou définies négatives.

Démonstration. Notons (p,q) la signature de Φ. Supposons $p \geqslant q$ (le cas $q \leqslant p$ se traite de façon analogue). On a $p+q = r$ (où r est le rang de Φ) et, d'après le théorème d'inertie de Sylvester, il existe une base orthogonale $(e_1, e_2, \ldots, e_n)$ de E telle que Φ s'écrive dans cette base : $\Phi(x) = x_1^2 - x_2^2 + x_3^2 - x_4^2 + \cdots + x_{2q-1}^2 - x_{2q}^2 + x_{2q+1}^2 + \cdots + x_r^2$. On note, pour tout $i \in [\![1,q]\!]$:

$$E_i := \mathbb{R}e_{2i-1} \oplus \mathbb{R}e_{2i}, \quad E_{hyper} := \bigoplus_{i\in[\![1,q]\!]}^{\perp} E_i \quad \text{et} \quad E_{def} := \mathbb{R}e_{2q+1} \oplus \cdots \oplus \mathbb{R}e_r.$$

La décomposition $E = E_{hyper} \oplus E_{def} \oplus \operatorname{Ker}\Phi$ est orthogonale et elle répond à la question. On a $\dim E_{hyper} = 2q$ et $\dim E_{def} = r - 2q = p - q$. On en déduit facilement l'unicité a équivalence près de la décomposition : $(E_{hyper}, \Phi_{|_{E_{hyper}}})$ et $(E_{def}, \Phi_{|_{E_{def}}})$ sont uniques à équivalence près. ∎

Nous allons voir que la décomposition de Witt reste vraie sur un corps K *quelconque*.

Théorème 33 (Décomposition de Witt).
Soient E un K-espace vectoriel de dimension finie et $\Phi \in Q(E)$. Il existe une décomposition en somme directe orthogonale :

$$E = E_{hyper} \overset{\perp}{\oplus} E_{def} \overset{\perp}{\oplus} \operatorname{Ker} \Phi,$$

où E_{hyper} est un espace hyperbolique et E_{def} est tel que $\Phi_{|_{E_{def}}}$ est définie. Une telle décomposition est unique à équivalence près.

Autrement dit (E, Φ) est la somme directe orthogonale d'un espace hyperbolique, d'un espace défini (ou anisotrope) et de $(\operatorname{Ker} \Phi, 0)$. Si φ est non dégénérée, (E, Φ) est la somme directe orthogonale d'un espace hyperbolique et d'un espace défini.

Démonstration. Nous allons donner une preuve simple de *l'existence* d'une décomposition de Witt. Nous *admettons provisoirement l'unicité*. Nous la prouverons plus loin en utilisant le théorème de simplification de Witt (*cf.* le corollaire 44 de la page 110).
On se ramène au cas où Φ est *non dégénérée* en écrivant (E, Φ) comme la somme directe orthogonale d'un espace non dégénéré et de $(\operatorname{Ker} \Phi, 0)$.
Si Φ est définie, on prend $E_{hyper} = (0)$ et $E_{def} = 0$. Sinon il existe un vecteur isotrope non nul x. On applique la proposition 30. On obtient un plan H_1, contenant x, tel que $(H_1, \Phi_{|_{H_1}})$ soit un plan hyperbolique. D'après la proposition 20 de la page 92, on a une somme directe orthogonale $E = H \overset{\perp}{\oplus} H^{\perp}$ et $\dim H^{\perp} = \dim E - 2$. On termine la preuve par récurrence descendante sur la dimension. On s'arrête quand on obtient E comme somme directe orthogonale d'un espace hyperbolique et d'un sous-espace F sur lequel Φ est définie. ∎

2.4.3 Isométries, groupe orthogonal et théorème de Witt

Dans toute cette partie (E, Φ) un espace quadratique *non dégénéré*. On suppose $E \neq (0)$.

Groupe orthogonal d'une forme quadratique

Proposition et définition 34. Soit (E, Φ) un espace quadratique *non dégénéré*.

(i) Les isométries de (E, Φ) forment un sous-groupe du groupe $\mathcal{GL}(E)$. On l'appelle le *groupe orthogonal* de Φ et on le note $O(\Phi)$.

(ii) Le déterminant d'un élément de $O(\Phi)$ est ± 1. les isométries de déterminant 1 forment un sous-groupe noté $O^+(\Phi)$ (ou $SO(\Phi)$). il est distingué et d'indice 2 dans $O(\Phi)$. On note $O^-(\Phi) := \{u \in O(\Phi) \mid \det u = -1\}$.

(iii) Si $u \in O(\Phi)$ est une homothétie de rapport non nul, alors $u = \pm Id$.

Démonstration.

(i) La composée de deux isométries est une isométrie. L'inverse d'une isométrie est une isométrie.

(ii) Nous admettons provisoirement que le déterminant d'une isométrie est ± 1. On le montre en utilisant la représentation matricielle des isométries (*cf.* la proposition 45 de la page 110). On a une suite exacte de groupes : $O^+(\Phi) \to O(\Phi) \to \{\pm 1\}$; la première flêche étant l'injection canonique et la seconde $\det$.

(iii) Soit $u \in O(\Phi)$ est une homothétie de rapport $\lambda \in \mathbb{K}^*$. soit $x \in E$, non isotrope. On a $\lambda^2 \Phi(x) = \Phi(\lambda x) = \Phi(u(x)) = \Phi(x)$, donc $\lambda^2 = 1$ et $\lambda = \pm 1$. ∎

Nous donnerons plus loin la version matricielle du groupe orthogonal (*cf.* la page 110).
Dans le cas euclidien réel, nous avons étudié en détails le groupe orthogonal dans le livre [L2] (*cf.* I.4, 1.6 et I.6,2).

Exemples.

1. Soient $E := \mathbb{R}^4$ et $\Phi := (1, 1, 1, -1)$, c'est-à-dire :
$$\Phi : (x, y, z, t) \mapsto \Phi(x, y, z, t) := x^2 + y^2 + z^2 - t^2;$$
$\mathcal{O}(\Phi)$ est le *groupe de Lorentz* (cf. [L2] I.6, 2.4.4 et *Un peu d'histoire*, page 371).
2. Si, plus généralement, $K = \mathbb{R}$ est Φ est la forme $(1, \ldots, 1, -1, \ldots, -1)$ de signature (p, q) sur $\mathbb{R}^n$, son groupe orthogonal est noté $\mathrm{O}(p, q)(\mathbb{R})$.

Symétries orthogonales

Dans le cas général (non dégénéré) un exemple important (et trés utile...) d'isométries est celui des *symétries orthogonales*.
Rappelons que si $u \in \mathcal{GL}(E)$ est tel que $u^2 = Id$, il existe deux sous-espaces $E^+(u) \subset E$ et $E^-(u) \subset E$ (on notera aussi plus simplement E^+ et E^-) tels que :
$$E = E^+ \oplus E^-, \quad u_{|_{E^+}} = Id_{E^+}, \quad \text{et} \quad u_{|_{E^-}} = -Id_{E^-}.$$
On dit que u est une *symétrie* de E.

Proposition et définition 35. Soit (E, Φ) un espace quadratique non dégénéré.

(i) Soit $u \in \mathcal{GL}(E)$ tel que $u^2 = Id$. L'automorphisme u est une isométrie si, et seulement si, E^+ et E^- sont *orthogonaux*. Dans ce cas, on dit que u est une *symétrie orthogonale*. On a alors $E^+ = (E^-)^\perp$ et $E^- = (E^+)^\perp$; en particulier E^+ et E^- sont *réguliers*.

(ii) Inversement si $F \subset E$ est un sous-espace *régulier*, il existe une et une seule symétrie orthogonale telle que $F = E^+(u)$. On l'appelle *symétrie orthogonale par rapport à* F.

(iii) Si $\dim E- = 1$, on dit que u est une *réflexion orthogonale* (de droite $D = E^-$) ou une *symétrie hyperplane orthogonale* (d'hyperplan E^+). Si $\dim E- = 2$, on dit que u est un *renversement orthogonal* (de plan E^-).

D'après (iii), il y a une bijection naturelle entre les symétries orthogonales et les sous-espaces réguliers propres (c'est-à-dire différents de (0) et E) de E.

Démonstration.

(i) Supposons que u soit une isométrie. Soient $x \in E^+$ et $y \in E^-$. On a :

$$\varphi(x,y) = \varphi\left(u(x), u(y)\right) = \varphi(x,-y) = -\varphi(x,y),$$

donc $2\varphi(x,y) = 0$ et $\varphi(x,y) = 0$ (puisque la caractéristique de K est différente de 2).

Inversement, supposons $E^+ \perp E^-$. Soient $x, y \in E$ et $x = x' + x''$, $y = y' + y''$ leurs décompositions sur E^+ et E^-. On a $u(x) = x' - x''$, $u(y) = y' - y''$. Par suite $\varphi(x,y) = \varphi(x',y') + \varphi(x'',y'') = \varphi\left(u(x), u(y)\right)$. Ainsi u est une isométrie.

(ii) Soit $F \subset E$ un sous-espace *régulier*. On a $E = F \overset{\perp}{\oplus} F^\perp$. On définit u par $u_{|_F} := Id_F$ et $u_{|_{F^\perp}} := -Id_{F^\perp}$.

Alors u est involutive, c'est une symétrie orthogonale et $E^+(u) = F$. ■

Nous avons étudié les *symétries hyperplanes orthogonales* dans le cas euclidien réel dans le livre [L2]. Nous avons donné une formule pour ces symétries (*cf.* I.4, 1.7.2, proposition 45). Nous allons voir qu'elle reste valable dans le cas général. On note φ la forme polaire de Φ. Soit $y \in E$ non isotrope, l'hyperplan orthogonal à y est $H = y^\perp$. On considère l'application $s_y : E \to E$ définie par :

$$s_y(x) = x - \frac{\varphi(x,y)}{\Phi(x)}. \tag{5}$$

L'application s_y est une involution ($s_y^2 = Id$ et $s_y \neq Id$), elle induit l'identité sur H et $s(y) = -y$, c'est donc la symétrie hyperplane définie par H. Notons que s_y n'est définie que si $\Phi(y) \neq 0$. Dans le cas euclidien réel (et plus généralement dans le cas où Φ est *définie*) cette condition correspond à $y \neq 0$ et c'est plus simple...

Remarques.

1. Une *réflexion orthogonale* u est déterminée par sa droite $E^-(u)$ ou son hyperplan $E^+(u)$. Il y a une bijection entre l'ensemble des réflexions orthogonales et l'ensemble des droites non-isotropes (resp. celui des hyperplans réguliers). Le déterminant d'une réflexion est -1.

2. Un *renversement orthogonal* u est déterminée par son plan $E^-(u)$. Il y a une bijection entre l'ensemble des renversements orthogonaux et l'ensemble des plans réguliers.

La notion de symétrie orthogonale est invariante par conjugaison par une isométrie. Plus précisément, on a le résultat suivant.

Proposition 36.

(i) Soient E un K-espace vectoriel de dimension finie et u une symétrie de E. Pour tout $v \in \mathcal{GL}(E)$, vuv^{-1} est une symétrie de E et $E^+(v) = v\left(E^+(u)\right)$, $E^-(v) = v\left(E^-(u)\right)$.

(ii) Soient (E, Φ) un espace quadratique non dégénéré et u une symétrie orthogonale de E. Pour tout $v \in O(E)$, vuv^{-1} est une symétrie orthogonale de E

(iii) Soient (E, Φ) un espace quadratique non dégénéré et $F \subset E$ un sous-espace *régulier*. On note τ la symétrie orthogonale par rapport à F. Alors, pour tout $v \in O(E)$, $v(F)$ est régulier et $v\tau v^{-1}$ est la symétrie orthogonale par rapport à $v(F)$.

Démonstration.
L'assertion (i) est un exercice facile d'algèbre linéaire. Montrons (ii). L'isométrie v conserve l'orthogonalité. On a $E^+(u) \perp E^-(u)$, donc, puisque $E^\pm(v) = v\left(E^\pm(u)\right)$, on a $E^+(v) \perp E^-(v)$. d'après la proposition-définition 35, la symétrie vuv^{-1} est orthogonale. L'assertion (iii) résulte trivialement de (i) et (ii). ∎

Proposition 37. Soit (E, Φ) un espace quadratique non dégénéré. Soient $x, x' \in E$, tels que $\Phi(x) = \Phi(x') \neq 0$. Il existe une isométrie $u \in \mathcal{O}(\Phi)$ telle que $u(x) = x'$. De plus on peut choisir pour u un produit d'au plus deux réflexions.

Démonstration.

- Si $y := x - x'$ n'est pas isotrope, on prend $u := s_y$.
- Supposons $y = x - x'$ isotrope. On pose $z := x + x'$. On montre, comme dans le cas euclidien (*cf.* [L2], I.4, 1.1.3), *l'identité du parallélogramme* :
$$\Phi(y) + \Phi(z) = 2\left(\Phi(x) + \Phi(x')\right).$$
On en déduit $\Phi(z) = 2\Phi(x) \neq 0$. On peut définir s_z et l'on a $s_z(x) = -x'$.
On prend $u = s_{x'} \circ s_z$. ∎

Théorème 38. Soit (E, Φ) un espace quadratique non dégénéré. Le groupe orthogonal $\mathcal{O}(\Phi)$ est *engendré par les réflexions orthogonales*.

Démonstration. On utilise la proposition ci-dessus pour faire une récurrence sur la dimension. La preuve est voisine de celle faite dans [L2] pour le cas euclidien (*cf.* I.4, 1.7.1, corolaire 52). ∎

Remarques.

1. Dans le cas euclidien, tout élément de $\mathcal{O}(\Phi)$ est le produit d'au plus $\dim E$ réflexions. C'est encore vrai dans le cas général, c'est le *théorème de Cartan-Dieudonné* (*cf.* l'exercice I.2.21 de la page 124). Mais ceci ne résulte pas de la preuve donnée ci-dessus qui donne seulement une majoration par $2\dim E$.
2. On peut montrer que le groupe $\mathcal{O}^+(\Phi)$ est *engendré par les renversements orthogonaux*, *cf.* l'exercice I.2.20 de la page 123.

Théorème de Witt. Indice de Witt

Le résultat suivant est dit *complétion régulière d'un sous-espace*.

Proposition 39. Soient (E, Φ) un espace quadratique non dégénéré et $F \subset E$ un sous-espace *singulier*. On note $F_0 \subset F$ le noyau de $\Phi_{|_F}$. Soient $(e_1, \dots, e_p)$ une base de F_0 et U un sous-espace supplémentaire de F_0 dans F. Alors on peut trouver $f_1, \dots, f_p \in E$ tels que, pour $i \in [\![1, p]\!]$, $H_i := Ke_i + Kf_i$ soit un plan hyperbolique, et que $\widetilde{F} := H_1 + H_2 + \cdots + H_p + U$ soit une somme directe orthogonale dans E : $\widetilde{F} = H_1 \overset{\perp}{\oplus} H_2 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} H_p \overset{\perp}{\oplus} U$.

Démonstration. On note φ la forme polaire de Φ. On a $\dim F_0 \geqslant 1$ (F est singulier par hypothèse). On fait la démonstration par récurrence sur $p = \dim F_0$. Supposons $p = 1$. On a une inclusion stricte $U \subset F$, donc une inclusion stricte $F^\perp \subset U^\perp$. Soit $f_1 \in U^\perp \setminus F^\perp$. Alors $\varphi(f_1, e_1) \neq 0$ et $H_1 := Ke_1 + Kf_1$ est un plan hyperbolique orthogonal à U (on utilise $U \perp F_0$).

Supposons le résultat vrai à l'ordre $p - 1$ et montrons le à l'ordre p. On pose :

$$U_1 := (Ke_2 + \cdots + Ke_p) + U = (Ke_2 + \cdots + Ke_p) \overset{\perp}{\oplus} U.$$

On a une inclusion stricte $U_1 \subset F$, donc une inclusion stricte $F^\perp \subset U_1^\perp$.

Soit $f_1 \in U_1^\perp \setminus F^\perp$. Alors $\varphi(f_1, e_1) \neq 0$ et $H_1 := Ke_1 + Kf_1$ est un plan hyperbolique orthogonal à U_1.

On pose $F_1 := H_1 \overset{\perp}{\oplus} U_1 = H_1 \overset{\perp}{\oplus} (Ke_2 + \cdots Ke_p) \overset{\perp}{\oplus} U$. La forme $\Phi_{|_{H_1 \overset{\perp}{\oplus} U}}$ est non dégénérée, donc $\operatorname{Ker} \Phi_{|_{F_1}} = Ke_2 + \cdots Ke_p$. Par suite, l'on peut appliquer à F_1 et $U' := H_1 \overset{\perp}{\oplus} U$ l'hypothèse de récurrence. Elle nous dit qu'il existe $f_2, \dots, f_p$ tels que, pour tout $i \in [\![2, \dots, p]\!]$, $H_i := Ke_i + Kf_i$ soit un plan hyperbolique et que $H_2 + \cdots + H_p + U'$ soit une somme directe orthogonale. Alors $\widetilde{F} := H_1 + H_2 + \cdots + H_p + U$ est une somme directe orthogonale. ∎

On en déduit un résultat de « gonflement » d'un sous-espace totalement isotrope, qui généralise la proposition 30 de la page 101.

Corollaire 40. Soient (E, Φ) un espace quadratique non dégénéré et $F \subset E$ un sous-espace *totalement isotrope*. Il existe un sous-espace hyperbolique $G \subset E$, contenant F et tel que $\dim G = 2 \dim F$.

On dit que H est une *extension hyperbolique* de F.

Exercice 11.

Soit (E, Φ) un espace quadratique non dégénéré. Montrer qu'il est hyperbolique si, et seulement s'il existe un sous-espace F totalement isotrope tel que $\dim E = 2 \dim F$.

Solution.

Notons $2n := \dim E$. Supposons que E est hyperbolique : $E = H_1 \overset{\perp}{\oplus} H_2 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} H_n$. On choisit un vecteur isotrope non nul e_i dans chaque plan hyperbolique H_i ($i \in [\![1, n]\!]$).

Si $i \neq j$, $e_i \perp e_j$, donc la somme $F := Ke_1 + \cdots + Ke_n$ est une somme directe orthogonale et F est totalement isotrope et de dimension n.
Inversement. Soit $F \subset E$ un sous-espace totalement isotrope de dimension n. On le complète en un sous-espace hyperbolique $G \subset E$ de dimension $2n$ en utilisant le corollaire 40 ci-dessus. On a $\dim G = \dim E$, donc $G = E$.

Dans le cas euclidien réel le groupe orthogonal opère de façon transitive sur les sous-espaces de dimension donnée (*cf.* [L2], I.6, 1.2.1). Il n'en est plus de même dans le cas général à cause des vecteurs isotropes. Le Théorème de Witt donne un résultat partiel.

Théorème 41 (Théorème de Witt).
Soit (E, Φ) un espace quadratique non dégénéré. Soient $F, F' \subset E$ deux sous-espaces. On suppose qu'il existe une isométrie $v : F \to F'$ entre $(F, \Phi_{|_F})$ et $(F', \Phi'_{|_F})$. Alors il existe une isométrie $u \in \mathcal{O}(\Phi)$ telle que $u_{|_F} = v$.

Démonstration. Les outils de la preuve sont les propositions 37 et 39.

1. On suppose d'abord que $F \subset E$ est un sous-espace *régulier*. On prouve le résultat par récurrence sur $\dim F$. Le résultat est trivial pour $f = (0)$. Supposons $\dim F > 0$ et le résultat vrai pour toute isométrie $u_1 : F_1 \to F_1'$ de F_1 sur F_1', où F_1, F_1' sont des sous-espaces de E, F_1 étant régulier et tel que $\dim F_1 = \dim F - 1$.
   Soit $x \in F$ non isotrope. On note $x' := v(x) \in F'$. D'après la proposition 37, il existe une isométrie w de (E, Φ) telle que $w(x) = x'$. On remplace alors F' par $w^{-1}(F')$ et u par $w^{-1} \circ u$, ce qui permet de supposer $u(x) = x$. En effet si $r : E \to E$ est une isométrie prolongeant $w^{-1} \circ u$, $v := w \circ r$ et une isométrie prolongeant u.
   Le sous-espace $F \cap x^\perp$ de E est *régulier* ($Kx \subset F$ est un sous-espace régulier). On considère alors le couple formé par $E_1 := x^\perp$ et son sous-espace régulier $F_1 := F \cap x^\perp$. On a $\dim(F \cap x^\perp) = \dim F - 1$. Puisque $u(x) = x$, on a $u(F \cap x^\perp) \subset x^\perp = E_1$. On peut donc appliquer l'hypothèse de récurrence et l'on prolonge $u_{|_{F_1}}$ en une isométrie $v_1 : E_1 \to E_1$. On prolonge ensuite v_1 en une isométrie $v : E \to E$, en posant $v(x) = x$ ($E = E_1 \overset{\perp}{\oplus} Kx$).
2. On suppose maintenant que F *n'est pas régulier*. On se ramène au cas précédent en utilisant une complétion régulière $\widetilde{F}$ de F (*cf.* la proposition 39). Il suffit de voir que u s'étend en une isométrie de $\widetilde{F}$ sur une complétion régulière $\widetilde{F}'$ de F'. C'est presque évident, le lecteur le vérifiera facilement. ∎

Exercice 12.
Soit (E, Φ) un espace quadratique non dégénéré. On dit que des sous-espaces $F, F' \subset E$ sont *isométriques* si les formes quadratiques $\Phi_{|_F}$ et $\Phi_{|_{F'}}$ sont équivalentes. Montrer que si F et F' sont isométriques, alors $F^\perp$ et $F'^\perp$ sont isométriques.
Solution. D'après le théorème de Wiit, il existe $u \in O(\Phi)$ telle que $u(F) = F'$. On a alors $u(F^\perp) = F'^\perp$, donc $F^\perp$ et $F'^\perp$ sont isométriques.

Proposition et définition 42. Soit (E, Φ) un espace quadratique non dégénéré. Les sous-espaces totalement isotropes maximaux de E ont tous la même dimension. On la note $\nu(\Phi)$ et on l'appelle *indice de Witt de* (E, Φ) (ou de Φ). Cet indice est invariant par équivalence.

Démonstration. Soient $F \subset E$ un sous-espace totalement isotrope et $F' \subset E$ un sous-espace totalement isotrope *maximal*. Nous allons montrer que $\dim F \leqslant \dim F'$. L'énoncé en résultera. On raisonne par l'absurde. On suppose que $\dim F > \dim F'$. Soit alors $F_1 \subset F$ un sous-espace de dimension $\dim F'$. Les restrictions de Φ à F_1 et F' sont nulles, donc, d'après le théorème de Witt, il existe une isométrie u de E, telle que $u(F_1) = F'$. Alors $u(F)$ est, comme F, totalement isotrope. On a $\dim u(F) = \dim F > \dim F'$, donc $u(F)$ contient *strictement* F', et F' n'est pas maximal. On a une contradiction.
On peut aussi établir ce résultat sans utiliser le théorème de Witt (cf l'exercice I.2.22). L'invariance par équivalence est évidente. ■

Si $K = \mathbb{R}$, l'indice de Witt d'une forme quadratique Φ de signature (p, q) est $\min(p, q)$ (*cf.* le théorème 32 de la page 103).
On montre le résultat suivant en se ramenant au cas de l'exercice 12.

Théorème 43 (Théorème de simplification de Witt).
Soient $(E_1, \Phi_1) = (F_1, \Psi_1) \overset{\perp}{\oplus} (G_1, \Theta_1)$ et $(E_2, \Phi_2) = (F_2, \Psi_2) \overset{\perp}{\oplus} (G_2, \Theta_2)$ deux espaces quadratiques non dégénérés. Si $(E_1, \Phi_1) \sim (E_2, \Phi_2)$ et $(F_1, \Psi_1) \sim (F_2, \Psi_2)$, on a $(G_1, \Theta_1) \sim (G_2, \Theta_2)$.

On déduit facilement du théorème ci-dessus le résultat suivant.

Corollaire 44. La décomposition de Witt du théorème 33 est unique à isomorphisme près.

Version matricielle des groupes orthogonaux

Soit (E, Φ) un espace quadratique non dégénéré de rang n. En choisissant une base $\mathcal{B}$ de E, on obtient une version matricielle du groupe orthogonal. C'est un sous-groupe de $GL_n(K)$.
Soit A la matrice de Φ dans la base $\mathcal{B}$. Elle est *symétrique* et inversible. On vérifie comme dans le cas euclidien (*cf.* la proposition 37 de I.4 1.6.3 de [L2]) et dans le cas symplectique (*cf.* la proposition 13), que M est la matrice d'une isométrie de $\mathcal{O}(\varphi)$ dans la base $\mathcal{B}$ si, et seulement si, : ${}^tMAM = A$. Une condition équivalente est ${}^tMA = AM^{-1}$.
On peut choisir la base $\mathcal{B}$ Φ-orthogonale. Alors $A = \mathrm{Diag}(a_1, \ldots, a_n)$ et la condition s'écrit ${}^tP\,\mathrm{Diag}(a_1, \ldots, a_n)P = \mathrm{Diag}(a_1, \ldots, a_n)$

Proposition et définition 45. Soit $A \in S_n(K)$ inversible. Une matrice $M \in M_n(K)$ est dite *A-orthogonale* si ${}^tMAM = A$. Les matrices A-orthogonales forment un sous-groupe de $GL_n(K)$.

Le déterminant d'une matrice A-orthogonale est ± 1. En effet :

$$\det A = \det({}^tMAM) = (\det A)(\det M)^2, \text{ donc } \det M = \pm 1.$$

2.5 Introduction à l'étude des formes quadratiques sur $\mathbb{Q}$

L'étude et la classification des formes quadratiques sur $\mathbb{Q}$ sont *difficiles*. Nous donnerons les définitions nécessaires et les énoncés essentiels. Nous ne démontrerons pas le résultat le plus difficile, le théorème de Hasse-Minkowski. Nous avons aussi omis d'autres preuves, beaucoup moins délicates, mais assez techniques. Pour les preuves manquantes, le lecteur intéressé pourra se reporter à [Serre] ou [BCh]. Nous souhaitons surtout présenter la démarche générale et l'une de nos motivations est de mettre en évidence le rôle essentiel des corps p-adiques, introduits dans le module III.3.

2.5.1 Du local au global

Le corps $\mathbb{R}$ et les corps p-adiques sont des complétions de $\mathbb{Q}$ associées à diférentes normes : $|\,|_\infty$ et les $|\,|_p$, pour p un nombre premier (*cf.* le module III.3). On dit que ce sont des *corps locaux*[18]. Le corps $\mathbb{Q}$ a une infinité de normes non deux à deux équivalentes, on dit que c'est un *corps global*. Partant d'un problème sur $\mathbb{Q}$, la première idée est de le *localiser* : on considère le même problème sur $\mathbb{R}$ et sur chaque $\mathbb{Q}_p$. Supposons que l'on sache résoudre tous ces problèmes locaux. L'étape suivante est d'essayer d'en déduire la solution du problème initial, que l'on appelle le problème *global*. C'est,l'idée du *passage du local au global*. Nous allons voir que la méthode est efficace dans le cas des formes quadratiques. C'est le *principe de Hasse*.

Soit $\Phi := (a_1, a_2, \ldots, a_n)$ une forme quadratique diagonale sur $\mathbb{Q}$. En utilisant les injections naturelles de corps : $\mathbb{Q} \to \mathbb{R}$ et, pour tout nombre premier p, $\mathbb{Q} \to \mathbb{Q}_p$ (où $\mathbb{Q}_p$ est un corps p-adique, *cf.* le module III.3), on obtient des formes quadratiques diagonales Φ_∞ et Φ_p sur $\mathbb{R}$ et sur les $\mathbb{Q}_p$. On dit que l'on a *localisé* la forme quadratique Φ.

- Si Φ représente 0, alors Φ_∞ et les Φ_p représentent (évidemment) 0.
- Si Φ et $\Psi := (b_1, b_2, \ldots, b_n)$ sont équivalentes sur $\mathbb{Q}$, (c'est-à-dire si les matrices diagonales $\mathrm{Diag}(a_1, a_2, \ldots, a_n)$ et $\mathrm{Diag}(b_1, b_2, \ldots, b_n)$ sont congruentes sur $\mathbb{Q}$), alors Φ_∞ et Ψ_∞ sont équivalentes sur $\mathbb{R}$ et, pour tout nombre premier p, Φ_p et Ψ_p sont équivalentes sur $\mathbb{Q}_p$.

On s'intéresse aux *réciproques éventuelles des deux assertions ci-dessus*. En dimension deux, on prouve assez facilement la réciproque de la première assertion.

Proposition 46. Soit $\Phi = (a_1, a_2)$ une forme quadratique non dégénérée sur $\mathbb{Q}^2$. Alors les deux conditions suivantes sont équivalentes :

(i) Φ représente 0 ;

(ii) Φ_∞ et les Φ_p, pour tout nombre premier p, représentent 0.

C'est le théorème de Hasse-Minkowski en dimension 2.

Démonstration. L'implication (i) $\Rightarrow$ (ii) est évidente. Montrons (ii) $\Rightarrow$ (i). On peut se ramener au cas où $a_1 = 1$. On note alors $a_2 = -a$. La forme quadratique $(1, -a)$ sur $\mathbb{R}$ représente 0 et elle est non-dégénérée, donc $a > 0$. On écrit $a \in \mathbb{Q}_+^*$ sous la forme : $a = \prod_p p^{v_p(a)}$ (le produit portant

[18] Avec un abus de langage pour le cas de $\mathbb{R}$.

sur un nombre fini de nombres premiers), Pour chaque p du produit, la forme quadratique $(1, -a)$ sur $\mathbb{Q}_p$ représente 0. Par suite a est *un carré dans* $\mathbb{Q}_p$ et $v_p(a)$ est *pair*. Ainsi a est *un carré dans* $\mathbb{Q}$ et la forme quadratique $(1, -a)$ sur $\mathbb{Q}$ représente 0. ■

Il résulte immédiatement de la proposition ci-dessus que $(\mathbb{Q}^2, \Phi)$ est un plan hyperbolique si, et seulement si, $(\mathbb{R}^2, \Phi_\infty)$ et tous les $(\mathbb{Q}_p^2, \Phi_p)$, pour p premier, sont des plans hyperboliques.

Le cas général est *très difficile*. On a le résultat suivant, que nous admettons (*cf.* [Serre], 3.2 Théorème 8).

Théorème 47 (Hasse-Minkowski). Soit Φ une forme quadratique sur un $\mathbb{Q}$-espace vectoriel de dimension finie. Pour que Φ représente 0, il faut et il suffit que Φ_∞ et chaque Φ_p, pour p premier, représentent 0.

En dimension 3 le théorème est dû à Legendre (avec une formulation différente...). Dans le cas général, la preuve de Hasse (en 1923) est par récurrence sur la dimension. Le passage de la dimension 3 à la dimension 4 est difficile. (Hasse utilise le théorème de Dirichlet sur les nombres premiers dans une progression arithmétique, qui est un théorème *d'analyse complexe et d'arithmétique modulaire*[19].

Corollaire 48. Soit Φ une forme quadratique non dégénérée sur $\mathbb{Q}$. Soit $a \in \mathbb{Q}^*$. Pour que Φ représente a dans $\mathbb{Q}$, il faut et il suffit qu'il en soit ainsi dans $\mathbb{R}$ et dans tous les $\mathbb{Q}_p$.

Démonstration. On peut supposer que $\Phi = (a_1, \ldots, a_n)$. On applique le théorème 47 à la forme quadratique $(a, -a_1, \ldots, -a_n)$ sur $\mathbb{Q}^{n+1}$. ■

On en déduit le théorème de classification des formes quadratiques sur $\mathbb{Q}$.

Théorème 49. Soient Φ et Ψ deux formes quadratiques non dégénérées sur $\mathbb{Q}$. Elles sont équivalentes sur $\mathbb{Q}$ si, et seulement si, elles le sont sur $\mathbb{R}$ et sur chaque $\mathbb{Q}_p$, pour p premier.

Démonstration. La nécessité est triviale. On montre la suffisance par récurrence sur le rang n. Supposons que les formes sont équivalentes sur $\mathbb{R}$ et sur chaque $\mathbb{Q}_p$. On note (a) la forme quadratique $x \mapsto ax^2$ sur $\mathbb{Q}$. Le résultat est évident pour $n = 0$ Supposons $n > 0$. Soit $a \in \mathbb{Q}^*$ représenté par Φ. Il est aussi représenté par Ψ_∞ et les Ψ_p, donc, d'après le corollaire 48 ci-dessus, il est représenté par Ψ. On a donc : $\Phi \sim (a) \overset{\perp}{\oplus} \Phi'$ et $\Psi \sim (a) \overset{\perp}{\oplus} \Psi'$, où Φ' et Ψ' sont des formes quadratiques sur $\mathbb{Q}$ (de rang $n-1$). En utilisant l'hypothèse et le théorème de simplification de Witt (*cf.* 43), on en déduit que les formes quadratiques Φ' et Ψ' sont équivalentes sur $\mathbb{R}$ et sur tous les $\mathbb{Q}_p$. L'hypothèse de récurrence montre alors que Φ' et Ψ' sont équivalentes sur $\mathbb{Q}$. Par suite Φ et Ψ sont équivalentes sur $\mathbb{Q}$. ■

2.5.2 Symbole de Hilbert. Invariants de Hasse

Nous avons vu dans le paragraphe précédent que la classification des formes quadratiques sur $\mathbb{Q}$ se ramène aux classifications sur $\mathbb{R}$ et les $\mathbb{Q}_p$. Nous connaissons la classification sur $\mathbb{R}$: à l'aide du rang et de la *signature*. Il reste à trouver un[20] invariant « remplaçant » la signature dans le cas de $\mathbb{Q}_p$. C'est *l'invariant de Hasse*, que l'on définit à partir du *symbole de Hilbert*.

[19] Pour une preuve, de niveau accessible à un étudiant de licence courageux, *cf.* [Serre], Chapitre VI.

[20] A priori on pourrait avoir besoin de *plusieurs* invariants.

L'idée de départ est très simple. Soient (E, Φ) et (E, Ψ) deux plans quadratiques non dégénérés. Si Φ représente 1 et si $\Psi \sim \Phi$, alors Φ représente 1. Inversement, D'après l'exercice 5, si Φ et Ψ représentent 1 elles sont équivalentes. D'après l'exercice 9 de la page 101, la forme diagonale $(x, y) \to ax^2+by^2$ représente 1 si, et seulement si, la forme diagonale $(x, y, z) \to -ax^2-by^2+z^2$ représente 0. Il est donc naturel d'introduire un nouvel invariant, *l'invariant de Hasse*, pour les formes quadratiques non dégénérées de rang 2 défini de la façon suivante. Considérons la forme $ax^2 + by^2$ sur K^2. On pose $(a, b)_p = 1$ si la forme quadratique $z^2 - ax^2 - by^2 = 0$, sur K^3, représente 0, sinon on pose $(a, b)_p = -1$. Le nombre $(a, b)_p \in \{\pm 1\}$ est invariant par équivalence de formes diagonales sur K^2, il permet donc de définir un invariant ε_p sur les formes non dégénérées de rang 2 : on pose $\varepsilon_p(\Phi) = (a, b)_p$ si Φ est équivalente à la forme $ax^2 + by^2$. Nous verrons plus loin que, si $K = \mathbb{Q}_p$, les deux invariants $\delta(\Phi)$ et $\varepsilon_p(\Phi)$ caractérisent une telle forme à équivalence près (*cf.* l'exercice 13).

Dans la suite de ce paragraphe K est soit $\mathbb{R}$ soit $\mathbb{Q}_p$ (p étant un nombre premier).

Définition 8. Soient $a, b \in K^*$.
On pose $(a, b) = 1$ si la forme quadratique $Q(z, x, y) : (x, y, z) \mapsto z^2 - ax^2 - by^2 = 0$, sur K^3, représente 0, c'est-à-dire si elle a une solution différente de $(0, 0, 0)$ dans K^3. On pose $(a, b) = -1$ sinon. Le nombre $(a, b) \in \{\pm 1\}$ s'appelle le *symbole de Hilbert* de a et b relativement à K.

Le symbole de Hilbert ne change pas si l'on multiplie a et b par des carrés. Il définit donc une application de $K^*/(K^*)^2 \times K^*/(K^*)^2$ dans $\{\pm 1\}$. S'il est nécessaire de préciser le corps K, on note $(a, b)_\infty$ pour $K = \mathbb{R}$ et $(a, b)_p$ pour $K = \mathbb{Q}_p$.
Soit $b \in K^*$. S'il n'a pas de racine carré dans K, $K_b := K[X]/(X^2 - b)K[X]$ est un corps engendré par l'image β de X, qui est une racine carrée de b. La norme de $\xi = x + \beta y \in K_b$ ($x, y \in K$) est, *par définition*, $N(\xi) := x^2 - by^2$.

Proposition 50. Soient $a, b \in K^*$ et $K_b := K(\sqrt{b})$ le corps obtenu en adjoignant à K une racine carrée de b. Pour que $(a, b) = 1$, il faut et il suffit que a soit *la norme* d'un élément de K_b^*.

Démonstration.
Les normes des éléments de K_b^* forment un groupe, le groupe des normes, noté NK_b^*.

- Si b est un carré : $b = c^2$ ($c \in K$). On a $K_b = K$ et $NK_b^* = K^*$. L'équation $z^2 - ax^2 - by^2$ admet pour solution $(c, 0, 1)$, donc $(a, b) = 1$.
- Supposons que b ne soit pas un carré. On note β une racine carrée de b. Tout élément $\xi \in K_b$ s'écrit $\xi = x + \beta y$, avec $x, y \in K$. Sa norme est, par définition :
  $N(\xi) := (x + \beta y)(x - \beta y) = x^2 - by^2$. On a $a \in NK_b^*$ si, et seulement s'il existe $y, z \in K$ tels que $a = z^2 - by^2$. Dans ce cas $z^2 - a - by^2 = 0$ et $(z, 1, y)$ est une solution de Q, donc $(a, b) = 1$. Inversement si $(a, b) = 1$, Q a une solution non triviale (z, x, y). On a $x \neq 0$ (sinon b serait un carré) et a est la norme de $z/x + \beta y/x$. ∎

On montre facilement les résultats suivants. Le lecteur le vérifiera à titre d'exercice. Pour l'assertion (iii), il utilisera la proposition 50 ci-dessus.

Proposition 51. On a (avec les hypothèses évidentes sur a, a', b, c) les formules suivantes :

(i) $(a,b) = (b,a)$ et $(a,c^2) = 1$;

(ii) $(a,-a) = 1$ et $(a,1-a) = 1$;

(iii) $(a,b) = 1 \Rightarrow (aa',b) = (a',b)$;

(iv) $(a,b) = (a,-ab) = (a,(1-a)b)$

Si $K = \mathbb{R}$. on a $(a,b)_\infty = 1$ si $a > 0$ ou $b > 0$ et $(a,b)_\infty = -1$ si $a < 0$ et $b < 0$. Pour $K = \mathbb{Q}_p$, on calcule (a,b) en utilisant le théorème suivant, que nous admettons (pour une preuve, *cf.* [Serre], 1.2 Théorème 1). On utilise le théorème 10 de la page 447 pour écrire $a = p^\alpha u$ et $b = p^\beta v$.

Théorème 52. On suppose $K := \mathbb{Q}_p$. Soient $a, b \in \mathbb{Q}_p^*$. On les écrit (de manière unique) sous la forme $a = p^\alpha u$ et $b = p^\beta v$, où $\alpha, \beta \in \mathbb{Z}$ et u, v appartiennent au groupe $\mathbb{U}$ des unités de $\mathbb{Q}_p^*$. On a :

$$(a,b)_p = (-1)^{\alpha\beta\varepsilon(p)} \left(\frac{u}{p}\right)^\beta \left(\frac{v}{p}\right)^\alpha, \quad \text{si} \quad p \neq 2; \tag{6}$$

$$(a,b)_2 = (-1)^{\varepsilon(u)\varepsilon(v)+\alpha\omega(v)+\beta\omega(u)}, \quad \text{si} \quad p = 2. \tag{7}$$

Dans les formules ci-dessus, on a noté respectivement $\varepsilon(x)$ et $\omega(x)$ les classes modulo 2 de $\frac{x-1}{2}$ et $\frac{x^2-1}{8}$.

Théorème 53. Le symbole de Hilbert est *une forme bilinéaire symétrique non dégénérée* sur le F_2-espace vectoriel $K^*/(K^*)^2$.

La bilinéarité se traduit par la formule $(aa',b) = (a,b)(a',b)$. La condition de non dégénérescence veut dire que, si $b \in K^*$ est tel que $(a,b) = 1$ pour tout $a \in K^*$, alors on a $b \in (K^*)^2$.

Démonstration. Le cas $K = \mathbb{R}$ est évident. Nous allons prouver le théorème dans le cas où $K = \mathbb{Q}_p$ avec $p \neq 2$. Le cas $p = 2$ est, sans être vraiment difficile, beaucoup plus technique et nous admettons le résultat (pour une preuve, *cf.* [Serre]). Soit $p \neq 2$. La bilinéarité se déduit de la formule (6). Pour montrer la non dégénérescence, il suffit de trouver, pour tout $a \in K^*/(K^*)^2$ distinct de l'élément neutre, un élément b tel que $(a,b) = -1$. On utilise pour cela la description de $\mathbb{Q}_p^*/(\mathbb{Q}_P^*)^2$ donnée dans le module III.3 (*cf.* l'exercice III.3.28 de la page 526). Ce groupe a 4 éléments et admet un système de représentants $\{1, p, u, up\}$, où $u \in \mathbb{U}$ est tel que $\left(\frac{u}{p}\right) = -1$. On peut donc prendre $a = p,\ u$ ou up. On choisit alors respectivement $b = u,\ p$ et u. ∎

On a une formule du produit pour les symboles de Hilbert. Elle est essentiellement équivalente à la loi de réciprocité quadratique. On note $\mathcal{P}$ l'ensemble des nombres premiers.

Théorème 54 (Hilbert). Soient $a, b \in \mathbb{Q}^*$. On a $(a,b)_p = 1$, sauf pour un nombre fini de $p \in \mathcal{P}$, et :

$$(a,b)_\infty \prod_{p\in\mathcal{P}} (a,b)_p = 1.$$

Démonstration. Compte tenu de la bilinéarité du symbole de Hilbert (*cf.* le théorème 53), il suffit de considérer les cas où a et b sont égaux à -1 ou à un nombre premier. On traite un par un tous

les cas en distinguant les nombres premiers pair et impairs. On utilise la proposition 51 et diverses formules sur le symbole de Legendre, en particulier la loi de réciprocité quadratique (ℓ, ℓ' étant des nombres premiers distincts) : $\left(\frac{\ell'}{\ell}\right)\left(\frac{\ell}{\ell'}\right) = (-1)^{\varepsilon(\ell)\varepsilon(\ell')}$ (*cf.* I.1, 3.3, théorème 24) et la formule $\left(\frac{2}{n}\right) = (-1)^{\omega(n)}$ (*cf.* I.1, 3.4, propostion 25). Le lecteur pourra trouver les détails dans [Serre] (pages 44 et 45) ou traiter patiemment tous les cas à titre d'exercice. ■

Proposition 55. Soit (E, Φ) une forme quadratique plane non dégénérée sur $\mathbb{Q}_p$. Soient $a, a' \in \mathbb{Q}_p$ représentés par Φ. Alors $(a, -\delta(\Phi))_p = (a', -\delta(\Phi))_p$.

Démonstration. Puisque Φ représente a, il existe une base orthogonale dans laquelle Φ s'écrit $\Psi(x, y) = ax^2 + by^2$. Alors Ψ représente a' : $a' = ax_0^2 + by_0^2$. Donc $aa' - aby_0^2 - (ax_0)^2 = 0$. Ainsi la forme $aa'x^2 - aby^2 - z^2$ représente 0 et $(aa', -ab) = 1$. Alors $(aa', -\delta(\Phi)) = 1$, et, d'après le théorème 53, $(a, -\delta(\Phi))(a', -\delta(\Phi)) = (aa', -\delta(\Phi))$, donc $(a, -\delta(\Phi)) = (a', -\delta(\Phi))$. ■

Exercice 13.

Soient $\Phi := (a_1, a_2)$ et $\Psi := (b_1, b_2)$ deux formes quadratiques diagonales non dégénérées sur $\mathbb{Q}_p$. Montrer qu'elles sont équivalentes si, et seulement si, elles ont même discriminant et si $(a_1, a_2)_p = (b_1, b_2)_p$.

Solution. La nécessité des deux conditions est évidente. Supposons inversement que $\delta(\Phi) = \delta(\Psi)$ et $(a_1, a_2)_p = (b_1, b_2)_p$. Si $a_1 a_2 = -d^2$, on a des plans hyperboliques et ils sont équivalents (*cf.* l'exercice 5). On suppose que l'on n'est pas dans ce cas. Montrons qu'il existe un nombre $c \in \mathbb{Q}_p$ représentant les deux formes. Leur équivalence résultera alors de l'exercice 5.

Supposons que Ψ représente c. D'après les propositions 51 et 55, on a :

$$(a_1, -a_1a_2)_p = (a_1, a_2)_p = (b_1, b_2)_p = (b_1, -b_1b_2)_p = (c, -b_1b_2)_p = (c, -a_1a_2)_p.$$

Par suite $(a_1, -a_1a_2)_p = (c, -a_1a_2)_p$ et $(ca_1^{-1}, -a_1a_2)_p = 1$, par bilinéarité. D'après la définition du symbole de Hilbert, il existe (x, y, z) tel que : $ca_1^{-1}x^2 - a_1a_2y^2 - z^2 = 0$. On a $x \neq 0$ (puisque l'on n'est pas dans le cas hyperbolique), donc $c = a_1\left(\frac{z}{x}\right)^2 + a_2\left(\frac{y}{x}\right)^2$ et la forme Φ représente c.

D'après le résultat de cet exercice, on a les assertions suivantes.

- Si $\Phi \sim (a_1, a_2)_p$, le symbole de Hilbert $(a_1, a_2)_p$ est indépendant de (a_1, a_2). C'est un invariant de Φ. On l'appelle *l'invariant de Hasse* de Φ et on le note $\varepsilon_p(\Phi)$.
- Deux formes quadratiques planes Φ et Ψ sur $\mathbb{Q}_p$ sont équivalentes si, et seulement si, $\delta(\Phi) = \delta(\Psi)$ *et* $\varepsilon_p(\Phi) = \varepsilon_p(\Psi)$.

Ces résultats s'étendent en toute dimension.

Proposition et définition 56. Soit $K = \mathbb{Q}_p$. Soit Φ une forme quadratique sur $\mathbb{Q}_p$. On suppose $\Phi \sim (a_1, \ldots, a_n)$. Le produit $\prod_{i,j \in [\![1,n]\!]; i<j} (a_i, a_j)_p$ est indépendant de $(a_1, \ldots, a_p)$. On l'appelle *l'invariant de Hasse* de Φ. On le note $\varepsilon_p(\Phi)$.

Pour la preuve de l'invariance par congruence, *cf.* [Serre], ch. 4, 2.1 Théorème 5. On a évidemment $\varepsilon_p(\Phi) = \pm 1$. Cet invariant se calcule avec une base Φ-orthogonale arbitraire.

On peut généraliser le résultat de l'exercice 13. Nous énonçons le résultat sans démonstration (*cf.* Serre, 2.3 Théorème 7).

Théorème 57 (Classification des formes quadratiques sur un corps p-adique). Deux formes quadratiques non dégénérées, définies sur $\mathbb{Q}_p$, sont équivalentes si, et seulement si, elles ont *même rang, même discriminant et même invariant de Hasse* ε_p.

On a, comme dans les cas de $\mathbb{R}$, $\mathbb{C}$ et des corps finis F_q, un *nombre fini* d'invariants, explicitement calculables.

Exercice type corrigé

I.2.0 Le but de l'exercice est de donner une preuve de la *loi de réciprocité quadratique* (*cf.* I.1, 3.3, théorème 24).

Dans tout l'exercice p et q sont des nombres premiers *impairs*. Si $a \in \mathbb{Z}$, on note $\left(\frac{a}{p}\right) := \left(\frac{\bar{a}}{p}\right)$ le *symbole de Legendre* (*cf.* I.1, 3.2, définition 3).

1. Soit $a \in \mathbb{Z}^*$. Montrer que le cardinal de l'ensemble $\{x \in F_p^* \mid \bar{a}x^2 = 1\}$ est $1 + \left(\frac{a}{p}\right)$.
2. On définit deux formes quadratiques Φ et Ψ sur le F_q-espace vectoriel F_q^p par :

$$\Phi : x := (x_1, \ldots, x_p) \mapsto \Phi(x) := x_1^2 + \cdots + x_p^2,$$

$$\Psi : y := (y_1, \ldots, y_p) \mapsto \Psi(y) := 2y_1y_2 + 2y_3y_4 + \cdots + 2y_{p-2}y_{p-1} + (-1)^{\frac{p-1}{2}} y_p^2$$

et l'on note $S := \{x \in F_q^p \mid \Phi(x) = 1\}$ et $S' := \{y \in F_q^p \mid \Psi(y) = 1\}$.

a) Calculer le discriminant de Ψ et montrer que Φ et Ψ sont équivalentes.

b) Vérifier que l'on définit une action du groupe additif $\mathbb{Z}/p\mathbb{Z}$ sur S par :

$$\forall k \in \mathbb{Z}, \quad k : (x_1, \ldots, x_p) \mapsto k(x_1, \ldots, x_p) := (x_{1+k}, \ldots, x_{n+k})$$

(k et les indices étant vus modulo p). Décrire les orbites de cette action et montrer, en utilisant l'équation aux classes (*cf.* la proposition 21 de I.2 3-4 du livre [L2]), que $\operatorname{card} S \equiv 1 + \left(\frac{p}{q}\right) \mod p$.

c) Soit $m \in \mathbb{N}^*$. Quel est le nombre d'éléments d'un hyperplan affine de F_q^m ?

d) Calculer $\operatorname{card} S' \mod p$.
On distinguera les $y \in S'$ tels que $y_1 = y_3 = \ldots = y_{p-2} = 0$ et ceux tels qu'il existe $k \in \{1, 3, \ldots, p-2\}$ tel que $y_k \neq 0$ et l'on utilisera dans un cas la question 1 et dans l'autre la question précédente.

3. Montrer que $\operatorname{card} S = \operatorname{card} S'$. En déduire :

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}.$$

Solution.

1. On distingue deux cas.
   - On suppose que $\bar{a}$ est un carré de F_q : $\bar{a} = \alpha^2$. Alors $ax^2 = 1$ si, et seulement si, $x = \pm\alpha^{-1}$. L'équation a deux solutions. On a $2 = 1 + 1 = 1 + \left(\frac{a}{p}\right)$.

- On suppose que $\bar{a}$ n'est pas un carré de F_q. Alors $\bar{a}x^2$ non plus et on ne peut pas avoir $\bar{a}x^2 = 1$. L'équation a donc 0 solution. On a $0 = 1 - 1 = 1 + \left(\frac{a}{p}\right)$.

Dans tout les cas le nombre de solutions de l'équation est $1 + \left(\frac{a}{p}\right)$.

2. a) Le discriminant de Ψ est $(-1)^{\frac{p-1}{2}}(-1)^{\frac{p-1}{2}} = 1$. Celui de Φ est évidemment 1, donc, d'après le corollaire 27 de la page 99, Φ et Ψ sont équivalentes.

   b) Si $(x_1, \ldots, x_p) \in S$ et si $\sigma \in \mathfrak{S}$, $(x_{\sigma(1)}, \ldots, x_{\sigma(p)}) \in S$; l'action de $\mathbb{Z}/p\mathbb{Z}$ sur F_q^p induit donc une action sur S. Les sous-groupes de $\mathbb{Z}/p\mathbb{Z}$ sont $\{\bar{0}\}$ et $\mathbb{Z}/p\mathbb{Z}$. Il y a donc deux sortes d'orbites : celles dont le stabilisateur est $\{\bar{0}\}$, c'est-à-dire celles contenant un $x := (x_1, x_2, \ldots, x_p)$, où les x_i sont deux à deux distincts et celles dont le stabilisateur est $\mathbb{Z}/p\mathbb{Z}$ c'est-à-dire celles de la forme $\{x = (u, u, \ldots, u)\}$. Les premières sont de cardinal p et les secondes de cardinal 1 (points fixes). Supposons que le nombre des premières est r et celui des secondes s. L'équation aux classes s'écrit : $\operatorname{card} S = rp+s$ et l'on a $\operatorname{card} S \equiv s \mod p$. On a $\Phi(u, u, \ldots, u) = pu^2 = 1$; on en déduit que s est le nombre de solutions dans $\mathbb{Z}/p\mathbb{Z}$ (donc dans F_p^*) de $pu^2 = 1$. D'après la question 1, ce nombre est $1 + \left(\frac{p}{q}\right)$. Par suite $\operatorname{card} S = 1 + \left(\frac{p}{q}\right) \mod p$.

   c) On montre en utilisant une translation que le cardinal d'un hyperplan *affine* de F_q^m est celui d'un hyperplan *vectoriel* de F_q^m, c'est-à-dire celui de F_q^{m-1}, qui est q^{m-1}.

   d) On note $d := \frac{p-1}{2}$ et $a := (-1)^d = (-1)^{\frac{p-1}{2}}$.

   Soit $y = (y_1, y_2, \ldots, y_{p-1}, y_p) \in S'$. On considère deux cas.

   * On suppose $y_1 = y_3 = \ldots = y_{p-2} = 0$. On peut alors choisir $y_2, \ldots, y_{p-1}$ arbitrairement et y_p solution de $ay^2 = 1$. Il y a q^d choix de $y_2, \ldots, y_{p-1}$ et $1 + \left(\frac{a}{q}\right)$ choix de y_p ; c'est-à-dire $q^d\left(1 + \left(\frac{a}{q}\right)\right)$ éléments de cette forme.

   * On suppose que l'un des $y_1, y_3, \ldots, y_{p-2}$ n'est pas nul. Alors $(y_1, y_3, \ldots, y_{p-2})$ est un vecteur non nul de F_q^d (il y a $q^d - 1$ choix possibles). On choisit ensuite y_p arbitrairement (il y a q choix possibles). Si $(y_1, y_3, \ldots, y_{p-2})$ et y_p sont fixés, $y_2, \ldots, y_{p-1}$ varient dans un hyperplan affine. D'après la question b), il y a q^{d-1} possibiltés. Ainsi il y a $(q^d - 1)qq^{d-1} = q^d(q^d - 1)$ éléments de cette forme.

   Par suite $\operatorname{card} S' = q^d\left(\left(1 + \left(\frac{a}{q}\right) + q^d - 1\right) = q^d\left(\left(\frac{a}{q}\right) + q^d\right)\right.$

3. En utilisant l'équivalence des formes Φ et Ψ, on montre $\operatorname{card} S = \operatorname{card} S'$, donc a fortiori $\operatorname{card} S = \operatorname{card} S' \mod p$. En utilisant les résultats de 2 a) et 2 c), on en déduit : $1 + \left(\frac{p}{q}\right) \equiv q^d\left(\left(\frac{a}{q}\right) + q^d\right) \mod p$. Par suite :

$$1 + \left(\frac{p}{q}\right) \equiv q^d\left(\frac{a}{q}\right) + q^{p-1} \equiv q^d\left(\frac{a}{q}\right) + 1 \mod p \quad \text{et} \quad \left(\frac{p}{q}\right) \equiv q^d\left(\frac{a}{q}\right).$$

En utilisant le théorème 19 de I.1.3.2, et la multiplicativité du symbole de Legendre, on obtient : $q^d = q^{\frac{p-1}{2}} = \left(\frac{q}{p}\right)$ et $\left(\frac{a}{q}\right) = \left(\frac{-1}{q}\right)^{\frac{p-1}{2}} = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}$. Donc : $\left(\frac{p}{q}\right) = q^d \left(\frac{a}{q}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}} \left(\frac{q}{p}\right) \mod p$. Ainsi, en multipliant par $\left(\frac{q}{p}\right)$:

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}.$$

Nous avons *a priori* seulement prouvé l'égalité modulo p, mais les deux membres prennent leurs valeurs dans ± 1 ; elle est donc vraie.

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.2.1 Soit $n \in \mathbb{N}^*$. On considère le groupe $Sp_{2n}(\mathbb{R})$ des matrices symplectiques réelles.

1. Soit $a > 0$. Soit $M : \begin{array}{rcl} [0,a[& \longrightarrow & M_{2n}(\mathbb{R}) \\ t & \longmapsto & M(t) \end{array}$ une application (interprétée comme une application vectorielle) de classe $\mathcal{C}^1$, telle que $M(0) = I_n$ et que, $\operatorname{Im} M \subset Sp_{2n}(\mathbb{R})$. On note $L := (dM/dt)_{t=0}$. Montrer que ${}^tLJ_n + J_nL = 0$.
2. On note $sp_{2n}(\mathbb{R}) := \{L \in M_{2n}(\mathbb{R}) \mid {}^tLJ_n + J_nL = 0\}$. On écrit $L \in M_{2n}(\mathbb{R})$ sous forme de matrice 2×2 par blocs $n \times n$: $L = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)$. Donner des conditions nécessaires et suffisantes sur les blocs A, B, C, D pour que $L \in sp_{2n}(\mathbb{R})$. Montrer que $sp_{2n}(\mathbb{R})$ est un sous-espace vectoriel de dimension $n(2n+1)$ de $M_{2n}(\mathbb{R})$.
3. Soit $L \in sp_{2n}(\mathbb{R})$. Montrer que la matrice e^L est symplectique : $\mathrm{e}^L \in Sp_{2n}(\mathbb{R})$.
4. Comparer avec le cas euclidien.

I.2.2 *Transvections symplectiques.* Montrer la proposition-définition 17 de la page 87.

I.2.3 *Transvections symplectiques et groupe symlectique.* Le but de l'exercice est de montrer que le groupe symplectique est engendré par les transvections symplectiques. Le corps K est arbitraire. Soit (E, φ) un espace symplectique. On note $G \subset Sp(\varphi)$ le sous-groupe de $Sp(\varphi)$ engendré par les transvections symplectiques. On appelle *couple hyperbolique* tout couple $(a, b) \in E^2$ tel que $\varphi(a, b) = 1$.

1. Montrer que G agit transitivement sur $E \setminus \{0\}$.
2. Montrer que G agit transitivement sur les couples hyperboliques.
3. Montrer, par récurrence sur $\dim E/2$, que $G = Sp(\varphi)$.
4. Montrer qu'une transformation symplectique est de déterminant 1.

I.2.4 $**$ *Invariant de Arf et transvections symplectiques*[21].
On suppose dans cet exercice que K est un corps commutatif de caractéristique 2. On utilisera le théorème 19 de la page 88.

1. Montrer que l'application $\sigma : K \to K$ définie par $\sigma : t \mapsto t^2 + t$ est un homomorphisme de groupes additifs. On suppose K *fini*. Quel est le cardinal de $K/\sigma(K)$?
2. Soit Φ une forme quadratique non dégénérée de rang $2n$. Soit $\mathcal{B} = (e_1, e_2, \ldots, e_{2n})$ une base symplectique pour B_Φ (*cf.* le corollaire 5). On a $B_\Phi(x, y) = {}^t X J_n Y$. On pose :

$$A(\Phi, \mathcal{B}) := \sum_{j=1}^{n} \Phi(e_i)\Phi(e_{n+i}).$$

   a) Soient $\mu, \lambda \in K$ et $q \in E$. Soit $\widetilde{\Phi} : x \mapsto \Phi(x) + \mu\,(B_\Phi(x, q))^2$. Comparer $A(\Phi, \mathcal{B})$ et $A(\widetilde{\Phi}, \mathcal{B})$, modulo $\sigma(K)$.
   On considère la transvection symplectique de (E, B_Φ), définie par $T : x \mapsto x + \lambda B_\Phi(x, q)\, q$.
   Soit $\widehat{\Phi} := \Phi \circ T$. Montrer que $A(\Phi, \mathcal{B}) \equiv A(\widehat{\Phi}, \mathcal{B})$, modulo $\sigma(K)$.
   b) Montrer que l'image de $A(\Phi, \mathcal{B})$ dans le groupe quotient $K/\sigma(K)$ est indépendante du choix de la base symplectique $\mathcal{B}$. On note $\mathbf{A}(\Phi)$ cette image. On l'appelle *invariant de Arf* de Φ.
   c) Montrer que si Φ et Ψ sont deux formes quadratiques non-dégénérées équivalentes, alors $\mathbf{A}(\Phi) = \mathbf{A}(\Psi)$.
   d) On suppose dans cette question que l'application de Frobenius $x \in K \to x^2 \in K$ est surjective. Soient deux formes quadratiques non-dégénérées Φ et Ψ telles que $\mathbf{A}(\Phi) = \mathbf{A}(\Psi)$. Montrer qu'elle sont équivalentes. On se ramènera au cas où $B_\Phi = B_\Psi$ et l'on cherchera une transvection symplectique T telle que $\Psi = \Phi \circ T$.
   Soit (E, Φ) un space quadratique non dégénéré de rang $2n$. Montrer que l'on peut choisir une base de E telle que :

$$\Phi(x_1, \ldots, x_{2n}) = \Psi_\nu(x_1, \ldots, x_{2n}) \sum_{i=1}^{n} x_i x_{i+n} + \nu(x_n^2 + x_{2n}^2),\ \nu \in K.$$

3. Donner une classification des formes quadratiques sur un corps fini de caractéristique 2.

Exercices sur la section 2

I.2.5 On suppose $K = \mathbb{R}$. Soit Φ une forme quadratique. Montrer que :

$$B_\Phi(x, y) = \lim_{t \to 0} \frac{1}{t} \left(\Phi(x + ty) - \Phi(x)\right).$$

I.2.6 Soient K un corps de caractéristique 2 et (E, Φ) un espace quadratique sur K. On suppose que l'application $x \in K \mapsto x^2 \in K$ est *surjective*. On suppose que $\operatorname{Ker} \Phi = (0)$.
Montrer que $\operatorname{Ker} B_\Phi$ est de dimension $\leqslant 1$.

[21] Cet exercice nous a été suggéré par un court article de Roger H. Dye, *On the Arf invariant*, Journal of Algebra 53, 1978, p. 36-39.

I.2.7 $*$ Soit $\Phi(x,y) = ax^2 + bxy + cz^2$ une forme quadratique réelle définie positive sur $\mathbb{R}^2$. On note $D = b^2 - 4ac$.

1. Montrer que Φ atteint sa borne inférieure sur l'ensemble $\mathbb{Z}^2 \setminus \{(0,0)\}$ en un point noté (ξ,η). Montrer que ξ et η sont premiers entre eux. Montrer qu'il existe $u \in SL_2(\mathbb{Z})$ tel que $\Psi := \Phi \circ u$ est de la forme $Ax^2 + By^2 + Cy^2$, avec $A = \Phi(\xi,\eta)$. Montrer que A est le minimum de Ψ sur $\mathbb{Z}^2 \setminus \{(0,0)\}$.
2. Montrer qu'il existe un couple $(x_0,y_0) \in \mathbb{Z}^2 \setminus \{(0,0)\}$, tel que $0 < \Phi(x_0,y_0) \leqslant \sqrt{-D/3}$ (on utilisera le minimum de $\Psi(x,y)$ quand $y = 1$ et $x \in \mathbb{Z}$).
3. Soit p un nombre premier tel que $p \equiv 1,\ 3 \mod 8$. Montrer qu'il existe un entier a tel que $a^2 \equiv -2 \mod p$. Montrer, en utilisant la forme quadratique $2x^2 + (ax+py)^2$, que p est de la forme $p = x_0^2 + 2y_0^2$, avec $(x_0,y_0) \in \mathbb{Z}^2$. Quelle est la condition nécessaire et suffisante sur un nombre premier q pour qu'il s'écrive $q = x^2 + 2y^2$, avec $(x,y) \in \mathbb{Z}^2$.

I.2.8 $*$ *Classes de formes quadratiques binaires entières.* On utilise les définitions de la remarque 1 de la page 89. Une *forme quadratique binaire entière* est une application :

$$f : \begin{array}{ccc} \mathbb{Z}^2 & \longrightarrow & \mathbb{Z} \\ (x,y) & \longmapsto & ax^2 + bxy + cy^2 \end{array} \qquad \text{avec} \qquad a,b,c \in \mathbb{Z}.$$

On la note $[a,b,c]$. Son *discriminant* est $D_f := b^2 - 4ac \in \mathbb{Z}$. On note $Q(D)$ l'ensemble des formes quadratiques binaires de discriminant D. Si Si $D_f < 0$ *et* $a > 0$, on dit que f est *imaginaire* . Le groupe $GL_2(\mathbb{Z})$ agit à droite sur $Q(D)$ par $f \mapsto f^P := f \circ P$, pour $P \in GL_2(\mathbb{Z})$. On pose $S := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ et $T := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Dans tout l'exercice, on suppose que les formes binaires considérées sont *imaginaires*. Une forme imaginaire $[a,b,c]$ est dite *réduite* si $|b| \leqslant a < c$ *ou* $0 \leqslant b \leqslant a = c$.

1. Montrer que le discriminant D_f est invariant sous l'action de $GL_2(\mathbb{Z})$.
2. Montrer, par un algorithme, que toute forme binaire est équivalente à une forme réduite (on utilisera S et T). Trouver une forme réduite équivalente à $[29,-34,10]$.
3. Trouver un nombre réel κ (indépendant de D) tel que, si $[a,b,c]$ est réduite, alors $a \leqslant \kappa\sqrt{-D}$. Montrer que le nombre de classes de D (c'est-à-dire le nombre d'orbites de $Q(D)$ sous l'action de $SL_2(\mathbb{Z})$) est *fini*.
4. Soit $f = [a,b,c]$ une forme réduite. Montrer que, pour $xy \neq 0$, on a $f(x,y) \geqslant a - |b| + c$. Calculer :
$$\inf_{\substack{(x,y)\in\mathbb{Z}^2 \\ (x,y)\neq(0,0)}} f(x,y) \qquad \text{et} \qquad \inf_{(x,x',x',y')\in\mathcal{J}} f(x,y)f(x',y'),$$
où $\mathcal{J} := \{x,y,x',y' \in \mathbb{Z} \mid xy' - x'y \neq 0\}$. Montrer que si deux formes réduites $[a,b,c]$ et $[a',b',c']$ sont équivalentes, alors $a = a'$, $c = c'$ et $b = \pm b'$.

I.2.9 $*$ *Plans quadratiques en caractéristique* 2.

À partir de la question 2, on suppose que le corps K est de caractéristique 2.

1. Soit $n \in \mathbb{N}^*$. Nous avons vu dans le livre [L1] (cf. I.3, 3.1.2) que si K est un corps commutatif de caractéristique différente de 2, alors tout polynôme P homogène de degré 2 sur K^n

définit une forme quadratique sur K^n. Montrer que si l'on utilise la définition usuelle, ce résultat n'est pas vrai si le corps K est de caractéristique 2. Montrer qu'il reste vrai avec la nouvelle définition.

2. On définit $\sigma : K \to K$ par $\sigma : t \mapsto t^2+t$. Montrer que σ est un homomorphisme de groupes additifs. Calculer le cardinal de $\sigma(K)$ quand K est fini.
3. On considère les formes quadratiques suivantes ($a \in K^*$ étant donné) :
   (a) $\Phi_a : (x,y) \mapsto x^2 + xy + ay^2$ et $\Psi : (x,y) \mapsto xy$ sur K^2 ;
   (b) $\Phi'_a : (x,y) \mapsto x^2 + xy + ay^2 + z^2$ et $\Psi' : (x,y) \mapsto xy + z^2$ sur K^3.

   Comparer B_{Φ_a} et B_Ψ. Étudier, pour a) et b), l'équivalence éventuelle des formes (dans le cas, on utilisera σ pour pour donner une condition nécessaire d'équivalence).
4. Montrer que tout plan quadratique non dégénéré est isomorphe à $(K^2, \Phi_{b,c})$, où
$$\Phi_{b,c} : (x,y) \mapsto bx^2 + xy + cy^2 \qquad \text{avec} \qquad b,c \in K.$$
5. a) Soit (E,Φ) un plan quadratique non dégénéré sur K. À toute base $\mathcal{B} = (e_1, e_2)$ symplectique pour B_Φ, on associe la classe $\mathbf{A}(\Phi)$ de $\Phi(e_1)\Phi(e_2)$ dans K modulo $\sigma(K)$. Montrer que $\mathbf{A}(\Phi)$ est indépendant du choix de la base $\mathcal{B}$. On l'appelle *invariant de Arf* de Φ.
   b) Montrer que si les plans quadratiques non dégénérés (E,Φ) et (E,Ψ) sont isomorphes, alors $\mathbf{A}(\Phi) = \mathbf{A}(\Psi)$. Montrer que si $\Phi_{b,c}$ et $\Phi_{b',c'}$ sont équivalentes, alors $bc = b'c'$ modulo l'image $\sigma(K)$ de σ.
6. On suppose que l'application $x \in K \mapsto x^2 \in K$ est *surjective*. Montrer que si $bc = b'c'$ modulo $\sigma(K)$, alors $\Phi_{b,c}$ et $\Phi_{b',c'}$ sont équivalentes.
7. Donner une classification des plans quadratiques non dégénérés sur un corps fini de caractéristique 2 à l'aide de l'invariant de Arf.

I.2.10 Soient K un corps de caractéristique 2 et (E,Φ) un espace quadratique.

1. Montrer que Φ est diagonalisable si, et seulement si, B_Φ est identiquement nulle.
2. On suppose que l'application de Frobenius $x \mapsto x^2$ est surjective et que $B_\Phi = (0)$. Montrer que Φ est équivalente à la forme quadratique définie sur K^m ($m := \dim E$) par $(x_1, \ldots, x_m) \mapsto x_1^2$.
3. Montrer que l'on a une somme directe orthogonale :
   $E = (E_1, \Phi_{|_{E_1}}) \overset{\perp}{\oplus} (E_2, \Phi_{|_{E_2}}) \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} (E_m, \Phi_{|_{E_m}}) \overset{\perp}{\oplus} (\operatorname{Ker} B_\Phi, \Phi_{|_{\operatorname{Ker} B_\Phi,}})$ où les E_i sont des plans symplectiques pour B_Φ. Montrer que m de dépend pas de la somme. Montrer que si $B_\Phi = (0)$, on a une décomposition orthogonale de E en deux sous-espaces quadratiques tels que sur l'un Φ est définie et sur l'autre elle est identiquement nulle.
4. Soit K un corps *fini* de caractéristique 2. Montrer que toute forme quadratique sur un K-espace vectoriel E de dimension 3 est équivalente à une forme quadratique sur K^3 de la forme $(x,y,z) \mapsto ax^2 + by^2 + cz^2$ $(a,b,c \in K)$ ou à la forme quadratique $(x,y,z) \mapsto xy + z^2$. On utilisera l'exercice I.2.9. Montrer que toute forme quadratique sur un K-espace vectoriel E de dimension supérieure ou égale à 3 représente 0.

I.2.11 Quel est le nombre de points d'une conique projective sur un corps fini F_q (de caractéristique arbitraire). On pourra utiliser l'exercice I.2.10 ci-dessus.

I.2.12 Décrire les carrés de $\mathbb{Q}$.

I.2.13

1. Soit (E, Ψ) un plan quadratique sur F_q. Montrer qu'il est hyperbolique si, et seulement si, -1 représente $\delta(\Psi)$, c'est-à-dire $\delta(\Psi) = \left(\frac{-1}{q}\right) = (-1)^{(q-1)/2}$.
2. Montrer que la forme quadratique $\Phi := (1,1)$ sur F_q^2 est *définie* si, et seulement si, (-1) n'est pas un carré de F_q.
3. Calculer l'invariant de Witt d'une forme quadratique sur F_q en fonction du rang, du discriminant et de q.

I.2.14 Vérifier que 987 est un nombre premier. Soit $K = F_{987}$. On considère les formes quadratiques $\Phi := (1,1)$ et $\Psi := (1,37)$ sur K^2 Ces deux formes sont-elles équivalentes ? Est-ce que la forme quadratique Ψ est définie ?

I.2.15 $*$ *Formes binaires réelles définies positives et demi-plan de Poincaré.* On reprend les notations de la remarque de la page 95. Soit $P := \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in SL_2(\mathbb{R})$. Pour tout $z \in \mathcal{H}$ on définit $Pz := \frac{\alpha z+\beta}{\gamma z+\delta} \in \mathbb{C}$. On note : $[a,b,c] : (x,y) \mapsto ax^2 + bxy + cy^2$. On note $\mathcal{Q}^+$ l'ensemble des formes quadratiques binaires réelles *définies positives* et $[a,b,c]^P = [a,b,c] \circ P$.

On pose $S := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ et $T := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

1. Montrer que si $P \in SL_2(\mathbb{R})$ et $z \in \mathcal{H}$, on a $\operatorname{Im} Pz = \frac{\operatorname{Im} z}{|\gamma z+\delta|^2}$.

   Montrer que $\mathcal{H}$ est stable par l'action de $SL_2(\mathbb{R})$.
2. Soit $\Gamma := SL_2(\mathbb{Z})/\{\pm I\}$. On note $\Gamma' \subset \Gamma$ le sous-groupe engendré par S et T. On définit $\mathcal{F} \subset \mathbb{C}$, appelé *domaine fondamental* : $\mathcal{F} := \{z \in \mathcal{H} \mid |z| \geqslant 1 \text{ et } |\operatorname{Re} z| \leqslant 1/2\}$.

   Montrer que tout point $z \in \mathcal{H}$ est Γ'-équivalent à un point de $\mathcal{F}$ (on pourra considérer un point de l'orbite de z sous l'action de Γ', dont la partie imaginaire est maximale).
3. On admet que Γ est engendré par S et T, c'est-à-dire $\Gamma' = \Gamma$.

   Montrer que pour tous $[a,b,c] \in \mathcal{Q}^+$ et $P \in \Gamma$, on a $\xi\left([a,b,c]^P\right) = P^{-1}\xi\left([a,b,c]\right)$.
4. On dit que $[a,b,c] \in \mathcal{Q}^+$ est *quasi-réduite* si $|b| \leqslant a \leqslant c$. Montrer que $[a,b,c] \in \mathcal{Q}^+$ est *quasi-réduite* si, et seulement si, $\xi\left([a,b,c]\right) \in \mathcal{F}$. Montrer que toute classe d'équivalence de $\mathcal{Q}^+$, modulo $SL_2(\mathbb{Z})$, contient une forme quasi-réduite.

I.2.16 *Groupe orthogonal en dimension deux.* Soit Φ une forme quadratique sur un K-espace vectoriel E de dimension 2 (le corps K est supposé de caractéristique impaire).

1. Montrer que les éléments de $\mathcal{O}^-(\Phi)$ sont les réflexions.
2. Montrer que tout élément de $\mathcal{O}^+(\Phi)$ est le produit de deux réflexions.

3. On suppose $u \in \mathcal{O}^+(\Phi)$ et $v \in \mathcal{O}^-(\Phi)$. Montrer que $vuv^{-1} = u^{-1}$.
4. Montrer que le groupe $\mathcal{O}^+(\Phi)$ est *abélien*.
5. Soit $u \in \mathcal{O}^+(\Phi)$. On suppose qu'il existe $x \in E$, non nul et non isotrope, tel que $u(x) = x$. Montrer que u est l'identité ;
6. On suppose que (E, Φ) est un plan hyperbolique. Montrer que $\mathcal{O}^+(\Phi)$ est isomorphe au groupe des matrices $\left\{\mathrm{Diag}(\alpha, \alpha^{-1}) \mid \alpha \in K^*\right\}$.

   On suppose que $K := F_q$ est *un corps fini*. Calculer le cardinal de $\mathcal{O}^+(\Phi)$ et celui de $\mathcal{O}(\Phi)$. Décrire $\mathcal{O}(\Phi)$ si $K := F_3$.

I.2.17 Soit (E, Φ) une forme quadratique non dégénérée de rang 3 sur K. On suppose que son invariant de Witt est $\nu(\Phi) = 1$. Montrer que $\Phi \sim \lambda(1, 1, -1)$, avec $\lambda \in K^*$.

I.2.18 Soient (E, Φ) un espace quadratique non dégénéré et $u \in O(\Phi)$ une symétrie orthogonale. Montrer qu'il existe une base orthonormée de E dans laquelle la matrice de u est de la forme $\mathrm{Diag}(1 \ldots, 1, -1, \ldots, -1)$.

I.2.19 *Centre d'un groupe orthogonal.* Soit (E, Φ) un espace quadratique non-dégénéré (sur un corps de caractéristique différente de 2). Le but de l'exercice est de montrer que le centre du groupe orthogonal $O(\Phi)$ est $\{\pm Id\}$, sauf dans le cas d'un plan hyperbolique sur le corps fini F_3.

1. Soit $u \in \mathcal{GL}(E)$ laissant globalement invariantes toutes les droites de E. Montrer que u est une homothétie.
2. On suppose $\dim E = 2$. Soit $u \in \mathcal{GL}(E)$. On suppose qu'il existe trois vecteurs propres de u deux à deux non colinéaires. Montrer que u est une homothétie.
3. Soit u dans le centre de $O(\Phi)$. Montrer que u laisse globalement invariantes toutes les droites *non isotropes* de E. Montrer que u laisse globalement invariants tous les plans *réguliers* de E.
4. On suppose $\dim E = 2$ et que (E, Φ) n'est pas un plan hyperbolique sur le corps $K \neq F_3$. Montrer que le centre de $\mathcal{O}(\Phi)$ est $\{\pm Id\}$.
5. On suppose que $\dim E \geqslant 3$. Montrer que si D est une droite isotrope, alors elle est l'intersection de deux plans hyperboliques.
6. On suppose que $\dim E \geqslant 3$. Montrer que le centre de $\mathcal{O}(\Phi)$ est $\{\pm Id\}$.
7. On suppose que (E, Φ) est un plan hyperbolique sur le corps F_3. Montrer que $O(\Phi)$ est isomorphe à $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ (le groupe de Klein V_2), donc *abélien* et égal à son centre.

I.2.20 *Renversements.* Soit (E, Φ) un espace quadratique non dégénéré (sur un corps de caractéristique différente de 2). On suppose $\dim E \geqslant 3$.

1. Soient τ_1 et τ_2 deux réflexions.

   Montrer qu'il existe deux renversements σ_1 et σ_2 tels que $\tau_1\tau_2 = \sigma_1\sigma_2$.
2. Montrer que $O^+(\Phi)$ est engendré par les renversements.

I.2.21 $*$ *Théorème de Cartan-Dieudonné.* Le but de l'exercice est de prouver le *théorème de Cartan-Dieudonné*. Soit (E, Φ) un espace quadratique non dégénéré (sur un corps de caractéristique différente de 2). Ce théorème dit que toute transformation de $O(\Phi)$ est le produit d'au plus $\dim E$ réflexions.

1. Soit $u \in O(\Phi)$. On note $v := u - Id$. On suppose que $\operatorname{Ker} v$ et $\operatorname{Im} v$ sont totalement isotropes (condition (C)). Montrer que $\operatorname{Ker} v$ et $\operatorname{Im} v$ sont deux sous-espaces totalement isotropes maximaux, puis qu'ils sont égaux et que $\dim E$ est pair. On note $\dim E = 2p$. Montrer que, dans une base convenable, la matrice de u est de la forme : $\left(\begin{array}{c|c} I_p & A \\ \hline 0_p & I_p \end{array}\right)$, avec $A \in M_p(K)$. Montrer que $\det u = 1$.
2. Soit $n \geqslant 3$. On suppose que, pour tout espace quadratique (F, Ψ) avec $\dim F = n - 1$, tout $w \in O(\Psi)$ est le produit d'au plus $n - 1$ réflexions. Montrer que, pour tout espace quadratique (E, Φ) avec $\dim E = n$, tout $u \in O(\Phi)$ est le produit d'au plus n réflexions. On distinguera trois cas, en montrant que l'un au moins est réalisé :
   a) Il existe $x \in E$, non isotrope, tel que $u(x) = x$.
   b) Il existe $x \in E$, non isotrope, tel que $v(x)$ soit non isotrope.
   c) L'automorphisme u satisfait la condition (C) de la question 1.

   On ramènera le cas c) à a) ou b) en considérant la composée de u et d'une réflexion.
3. Montrer le théorème de Cartan-Dieudonné.

I.2.22 $*$ *Sous-espaces totalement isotropes maximaux.* Soit (E, Φ) un espace quadratique. Le but de l'exercice est de montrer de façon élémentaire, sans utiliser le théorème de Witt, que les sous-espaces totalement isotropes maximaux de E ont tous la même dimension. Soient $F_1, F_2 \subset E$ deux sous-espaces totalement isotropes maximaux. On suppose $F_1 \neq F_2$. On note $F := F_1 \cap F_2$ et l'on choisit un supplémentaire F_1' (resp. F_2') de F dans F_1 (resp F_2).

1. Montrer que $F_1' \cap {F_2'}^{\perp} \subset F_2^{\perp}$.
2. Soit $x \in F_1' \cap {F_2'}^{\perp}$. On note $G := F_2 + Kx$. Montrer que $G = F_2$.
3. Montrer que $F_1' \cap {F_2'}^{\perp} = F_2' \cap {F_1'}^{\perp} = (0)$. En déduire $\dim F_1 = \dim F_2$.

I.2.23 On suppose $K = \mathbb{R}$.

1. Montrer, à partir du théorème de Sylvester, que les *plans* quadratiques réels non dégénérés sont classifiés par la donnée du discriminant et de l'invariant de Hasse ε_∞.
2. Soit (E, Φ) un espace quadratique non-dégénéré de signature (p, q). Calculer son invariant de Hasse $\varepsilon_\infty(\Phi)$.

Géométrie

Partie II

Les mathématiques sont nées du point et du nombre ; aujourd'hui le second (via ses extensions) est souverain, mais ce fut longtemps le contraire. La tradition grecque régnait encore sans partage en 1794 quand Legendre proposait une édition d'Euclide, proche de l'original, qui restera une Bible pour un siècle (l'enseignement secondaire en était encore influencé en 1950). Au nom d'Euclide, il faut ajouter ceux d'Archimède et d'Apollonius, qui n'eurent guère de postérité créative pendant quinze siècles. Ce si long règne s'explique par deux points essentiels : une technique axiomatique clarifiant les règles du jeu d'une part, le fait que le sujet repose sur des bases intuitives fortes liées à l'expérience spatiale d'autre part. Toutefois les axiomes euclidiens étaient insuffisants et la géométrie finira par souffrir de la comparaison avec l'algèbre et l'analyse, aux fondements plus sûrs.
Les découvertes simultanées de la géométrie analytique par Descartes en 1637 dans le *Discours de la Méthode* et Fermat, puis son exploitation intensive par Euler – découvrant par le *calcul* la droite qui porte son nom ! –, ouvrirent de nouveaux horizons, notamment par la création d'objets d'étude à l'infini. La géométrie devint ainsi un outil pour la physique et la mécanique céleste, vers la maîtrise de l'Univers.
Un autre atout de la géométrie moderne est le calcul vectoriel, issu des quaternions découverts par Hamilton en 1843. Dès lors, la percée de l'algèbre linéaire pendant les cinquante ans qui suivirent changea profondément la donne. Aujourd'hui les axiomatiques de Hilbert, corrigeant Euclide, sont « marginalisées » par l'inclusion de la géométrie dans cette nouvelle et puissante discipline. L'enseignement en fut, un temps, fortement bouleversé.
Pourtant, loin d'être morte, la géométrie est toujours secrètement présente. Une vision géométrique sous-tend la théorie des nombres et une partie de l'analyse. Elle féconde les travaux les plus actuels, comme par exemple la recherche de symétries cachées des mathématiques ou de la physique théorique. Le point reprend sa souveraineté, même si les représentations intuitives sous-jacentes diffèrent largement des figures euclidiennes de base.

Dans ce cours, nous présentons brièvement les notions de base d'une théorie des surfaces de l'espace usuel, sans tomber dans un encyclopédisme aujourd'hui obsolète à propos d'un thème qui fut fort riche dans les siècles passés, et qui servit souvent de matrice aux champs plus abstraits de l'algèbre et de l'analyse.

Surfaces II.1

Dans ce module, nous allons étudier des surfaces d'un espace affine $\mathcal{E}$ de dimension 3. Un point de l'espace affine $\mathcal{E}$ est repéré par ses coordonnées dans un repère $(O;\vec{\imath},\vec{\jmath},\vec{k})$. Le point M de coordonnées (x,y,z) est le point $M = O + x\vec{\imath} + y\vec{\jmath} + z\vec{k}$. Nous pouvons donc identifier l'espace affine $\mathcal{E}$ avec $\mathbb{R}^3$, ce qui nous amène à étudier les surfaces de $\mathbb{R}^3$.
Une surface est la représentation mathématique d'objets tels que les bulles de savon, la frontière commune entre un milieu liquide et un milieu gazeux, un relief avec des montagnes, des sommets, des cols, des vallées, etc.
Il existe essentiellement deux manières de décrire une surface : comme l'image d'un domaine U de $\mathbb{R}^2$ par une application continue $f : U \to \mathcal{E}$ ou comme un ensemble de points vérifiant une équation : $\{M \in \mathcal{E} \mid F(M) = 0\}$. Dans le premier cas, nous parlerons de *nappe paramétrée*. Dans le second, nous parlerons de *surface implicite*.
Nous verrons, entre autres choses, que lorsque les surfaces considérées sont suffisamment régulières (lisses), les deux approches décrivent localement les mêmes objets. Le pont entre les deux notions est établi au moyen du théorème des fonctions implicites. Il est indispensable que le lecteur ait une connaissance détaillée du module sur les fonctions de plusieurs variables du livre [L2].
Nous introduirons la notion d'espace et de plan tangent, et nous étudierons la position de la surface par rapport à son plan tangent. Nous serons pour cela amenés à étudier le développement de Taylor à l'ordre 2 de certains paramétrages locaux de la surface. La partie d'ordre 2 du développement de Taylor est une forme quadratique. Nous supposerons que le lecteur est familier avec les modules d'algèbre bilinéaire du livre [L2]. En particulier, le dernier chapitre de ce module est consacré à la présentation d'exemples, dont celui des quadriques : ce sont des surfaces définies par une équation de la forme $\Phi(\overrightarrow{OM}) + L(\overrightarrow{OM}) + b = 0$ avec Φ une forme quadratique sur $\mathbb{R}^3$, L une forme linéaire sur $\mathbb{R}^3$ et b un réel.
À plusieurs reprises, l'espace sera muni d'un produit scalaire. Nous travaillerons alors dans des repères orthonormés ce qui nous permettra d'identifier le produit scalaire avec le produit scalaire usuel :

$$\left(\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} \middle| \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} \right) = x_1x_2 + y_1y_2 + z_1z_2.$$

Nous serons fréquemment amenés à considérer des applications de classe $\mathcal{C}^p$, $p \geqslant 1$, ou des applications de classe $\mathcal{C}^\infty$. Tous les résultats que nous énoncerons pour des applications de classe $\mathcal{C}^p$ seront valides pour $p \in \{1, 2, 3, \ldots, +\infty\}$.

1 Nappes paramétrées

Dans toute cette partie, U désigne un domaine (un ouvert connexe) non vide de $\mathbb{R}^2$.

1.1 Définitions

Comme dans le cas des courbes du plan et de l'espace (*cf.* module « Courbes paramétrées » du livre [L1]), il faut manipuler la notion de surface de $\mathbb{R}^3$ avec beaucoup de précautions. Nous commencerons par étudier le cas des surfaces paramétrées. Le terme consacré est celui de *nappe paramétrée*.

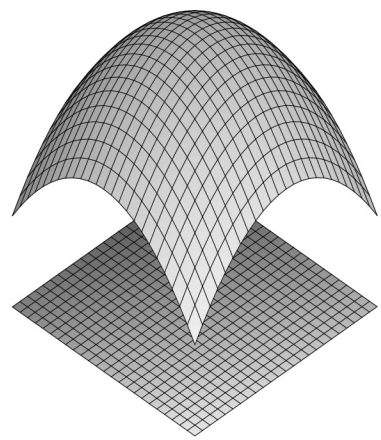
Définition 1. Une *nappe paramétrée* est une application continue $\boldsymbol{f} : U \to \mathbb{R}^3$.
Le *support de la nappe* est $\boldsymbol{f}(U)$.

Exemples.

1. Le graphe d'une fonction continue $F : U \to \mathbb{R}$ peut être interprété comme une nappe paramétrée, avec $\boldsymbol{f} : U \to \mathbb{R}^3$ définie par :

$$\boldsymbol{f}(x, y) := \big(x, y, F(x, y)\big).$$

Une telle nappe s'appelle une *nappe cartésienne*. Le dessin ci-contre montre un exemple de nappe cartésienne avec $F :]-2, 2[\times]-2, 2[\to \mathbb{R}$ définie par $F(x, y) := 8 - x^2 - y^2$.

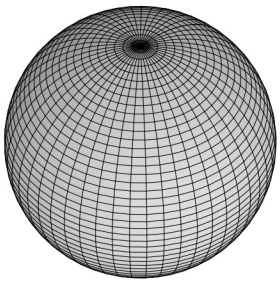


2. Le support de la nappe paramétrée $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ définie par :

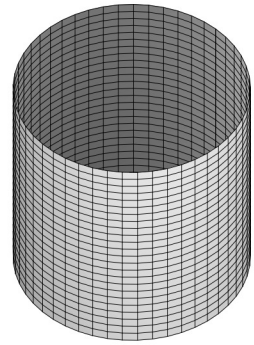
$$\boldsymbol{f}(\theta, \varphi) := (\cos\theta\cos\varphi, \sin\theta\cos\varphi, \sin\varphi)$$

est la sphère Σ de centre O et de rayon 1. En effet, si $M := (x, y, z)$ est un point du support, alors :

$$x^2 + y^2 + z^2 = \cos^2\theta\cos^2\varphi + \sin^2\theta\cos^2\varphi + \sin^2\varphi = \cos^2\varphi + \sin^2\varphi = 1.$$



Le point M appartient donc à la sphère Σ. D'autre part, si M est un point de la sphère Σ et si l'on pose $r^2 := x^2 + y^2$, alors, $r^2 + z^2 = 1$, ce qui montre qu'il existe $\varphi \in \mathbb{R}$ tel que $r = \cos\varphi$ et $z = \sin\varphi$. Alors, comme $x^2 + y^2 = r^2$, il existe $\theta \in \mathbb{R}$ tel que $x = r\cos\theta$ et $y = r\sin\theta$. Donc tout point de la sphère Σ est l'image d'un point $(\theta, \varphi) \in \mathbb{R}^2$ par l'application $\boldsymbol{f}$.



3. Le support de la nappe paramétrée $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}(\theta, z) := (2\cos\theta, 2\sin\theta, z)$ est le cylindre d'axe (Oz) et de rayon 2.

Une nappe paramétrée n'est pas seulement un sous-ensemble de $\mathbb{R}^3$. Une nappe paramétrée vient avec un paramétrage. Un point d'une nappe n'est donc pas seulement un point du support : il est *marqué* par le paramétrage.

Définition 2. Un *point* $\boldsymbol{m}$ d'une nappe paramétrée $\boldsymbol{f} : U \to \mathbb{R}^3$ est un couple $(\boldsymbol{u}, M)$ avec $\boldsymbol{u} \in U$ et $M = \boldsymbol{f}(\boldsymbol{u}) \in \mathbb{R}^3$.

Pour étudier les propriétés géométriques des nappes paramétrées, nous aurons souvent besoin d'imposer des conditions de régularité.

Définition 3. Une nappe paramétrée $\boldsymbol{f} : U \to \mathbb{R}^3$ est *de classe* $\mathcal{C}^p$ si l'application $\boldsymbol{f}$ est *de classe* $\mathcal{C}^p$.

Ce qui nous empêche généralement d'identifier les points d'une nappe paramétrée avec les points de son support est le fait qu'il puisse exister des points *multiples*.

Définition 4. Soit $\boldsymbol{f} : U \to \mathbb{R}^3$ une nappe paramétrée. Un point $M \in \boldsymbol{f}(U)$ est :

- un *point simple* si $\boldsymbol{f}^{-1}(M)$ ne contient qu'un seul point,
- un *point multiple* si $\boldsymbol{f}^{-1}(M)$ contient au moins deux points.

La nappe paramétrée est une *nappe simple* si tous les points du support sont des points simples.

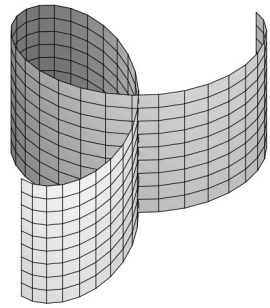
Remarque. Une nappe paramétrée $\boldsymbol{f} : U \to \mathbb{R}^3$ est simple si, et seulement si, $\boldsymbol{f}$ est injective. On peut alors identifier les points de la nappe et les points du support puisqu'ils sont en bijection.

Exemples.

1. La nappe paramétrée $\boldsymbol{f} :]0, \pi[\times]0, 1[\to \mathbb{R}^3$ définie par :

$$\boldsymbol{f}(\theta, s) := \big(\cos(2\theta)\cos(\theta), \cos(2\theta)\sin(\theta), s\big)$$

n'est pas une nappe simple. En effet, les points $(\pi/4, s)$ et $(3\pi/4, s)$ ont la même image $(0, s)$.



2. Toute nappe cartésienne est simple : si $\boldsymbol{f}(x,y) = \big(x, y, F(x,y)\big)$, alors :

$$\big(x_1, y_1, F(x_1,y_1)\big) = \big(x_2, y_2, F(x_2,y_2)\big) \quad \Longleftrightarrow \quad (x_1,y_1) = (x_2,y_2).$$

Nous allons maintenant introduire la notion de *courbe tracée sur une nappe*. Signalons que nos représentations graphiques du support des nappes paramétrées sont obtenues en traçant de telles courbes. Nous renvoyons le lecteur au livre [L1]pour les notions sur les courbes paramétrées.
Dans ce qui suit, $I \subset \mathbb{R}$ est un intervalle d'intérieur non vide. Observons que lorsque $\boldsymbol{f} : U \to \mathbb{R}^3$ est une nappe paramétrée et $\gamma_1 : I \to U$ est une courbe paramétrée, alors, $\boldsymbol{f} \circ \gamma_1 : I \to \mathbb{R}^3$ est une courbe paramétrée et le support de $\boldsymbol{f} \circ \gamma_1$ est contenu dans le support de $\boldsymbol{f}$.

Définition 5. Une courbe paramétrée $\gamma : I \to \mathbb{R}^3$ est *tracée* sur la nappe paramétrée $\boldsymbol{f} : U \to \mathbb{R}^3$ s'il existe une courbe paramétrée $\gamma_1 : I \to U$ telle que $\gamma = \boldsymbol{f} \circ \gamma_1$. La courbe passe par le point $\boldsymbol{m} := (\boldsymbol{u}, M)$ au temps $t_0 \in I$ si $\gamma_1(t_0) = \boldsymbol{u}$ et $\gamma(t_0) = M$.

Exemples.

1. Si $\boldsymbol{f} : (x,y) \in \mathbb{R}^2 \mapsto \big(x, y, F(x,y)\big) \in \mathbb{R}^3$ est une nappe cartésienne, alors les courbes $x \mapsto \big(x, y_0, F(x,y_0)\big)$ et $y \mapsto \big(x_0, y, F(x_0,y)\big)$ représentent les graphes des fonctions $F(\cdot, y_0)$ et $F(x_0, \cdot)$. Le support de ces courbes coïncident avec l'intersection du support de $\boldsymbol{f}$ et du plan d'équation $y = y_0$ ainsi qu'avec l'intersection du support de $\boldsymbol{f}$ et du plan d'équation $x = x_0$. On les appelle les *lignes coordonnées* de la nappe relativement au paramétrage $\boldsymbol{f}$.

2. Considérons le paramétrage de la sphère de centre O et de rayon 1 donné précédemment : $\boldsymbol{f}(\theta,\varphi) = (\cos\theta\cos\varphi, \sin\theta\cos\varphi, \sin\varphi)$. Les courbes $\varphi \mapsto \boldsymbol{f}(\theta_0,\varphi)$ sont appelées les *méridiens*. Les courbes $\theta \mapsto \boldsymbol{f}(\theta,\varphi_0)$ sont appelées les *parallèles*. Le support de ces courbes sont des cercles tracés sur la sphère.

Si l'on s'impose des conditions de régularité, on a la définition suivante.

Définition 6. Soit $\boldsymbol{f} : U \to \mathbb{R}^3$ une nappe paramétrée de classe $\mathcal{C}^p$. Une courbe paramétrée $\gamma : I \to \mathbb{R}^3$ est $\mathcal{C}^p$*-tracée* sur la nappe $\boldsymbol{f}$ s'il existe une courbe paramétrée $\gamma_1 : I \to U$ de classe $\mathcal{C}^p$ telle que $\gamma = \boldsymbol{f} \circ \gamma_1$.

La courbe paramétrée $\gamma : \mathbb{R} \to \mathbb{R}^3$ définie par $\gamma(t) := (t, 0, 0)$ est de classe $\mathcal{C}^\infty$. C'est une courbe tracée sur la nappe paramétrée $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}(x,y) := (x^3, y, 0)$. En effet, $\gamma = \boldsymbol{f} \circ \gamma_1$ avec $\gamma_1(t) := \big(\sqrt[3]{t}, 0, 0\big)$. Il y a unicité d'une telle courbe γ_1 puisque la nappe $\boldsymbol{f}$ est simple. La nappe $\boldsymbol{f}$ est de classe $\mathcal{C}^\infty$. Cependant, la courbe γ n'est pas $\mathcal{C}^\infty$-tracée (ni même $\mathcal{C}^1$-tracée) sur la nappe $\boldsymbol{f}$, puisque γ_1 n'est pas de classe $\mathcal{C}^1$ en $t = 0$.

1.2 Nappes géométriques

En réalité, l'objet géométrique qui nous intéresse véritablement est le support de la courbe paramétrée. Nous ne pouvons pas complètement nous passer de la notion de paramétrage, mais nous pouvons affaiblir la notion de nappe de la manière suivante.

Observons d'abord que si U_1 et U_2 sont deux domaines de $\mathbb{R}^2$ et si $\varphi : U_2 \to U_1$ est un homéomorphisme, alors $\boldsymbol{f}_1 : U_1 \to \mathbb{R}^3$ est une nappe paramétrée si, et seulement si, $\boldsymbol{f}_2 := \boldsymbol{f}_1 \circ \varphi : U_2 \to \mathbb{R}^3$ est une nappe paramétrée. De plus, si $\varphi : U_2 \to U_1$ est un difféomorphisme de classe $\mathcal{C}^p$, c'est-à-dire si φ et φ^{-1} sont des applications de classe $\mathcal{C}^p$, alors $\boldsymbol{f}_1$ est une nappe paramétrée de classe $\mathcal{C}^p$ si, et seulement si, $\boldsymbol{f}_2$ est une nappe paramétrée de classe $\mathcal{C}^p$. Cela nous conduit à la définition suivante.

Définition 7. Deux nappes paramétrées $\boldsymbol{f}_1 : U_1 \to \mathbb{R}^3$ et $\boldsymbol{f}_2 : U_2 \to \mathbb{R}^3$ sont *équivalentes* (respectivement $\mathcal{C}^p$*-équivalentes*) s'il existe un homéomorphisme (respectivement un difféomorphisme de classe $\mathcal{C}^p$) $\varphi : U_2 \to U_1$ tel que $\boldsymbol{f}_2 = \boldsymbol{f}_1 \circ \varphi$.

Exercice 1.

Montrer que la relation « $\boldsymbol{f}_1 : U_1 \to \mathbb{R}^3$ et $\boldsymbol{f}_2 : U_2 \to \mathbb{R}^3$ sont équivalentes (respectivement $\mathcal{C}^p$-équivalentes) » est une relation d'équivalence.

Solution.

- Pour la réflexivité, prendre $\varphi = \mathrm{Id}$.
- Pour la transitivité, utiliser le fait que la composée de deux homéomorphismes (respectivement de deux difféomorphismes de classe $\mathcal{C}^p$) est un homéomorphisme (respectivement un difféomorphisme de classe $\mathcal{C}^p$).
- Pour la symétrie, utiliser le fait que l'inverse d'un homéomorphisme (respectivement d'un difféomorphisme de classe $\mathcal{C}^p$) est un homéomorphisme (respectivement un difféomorphisme de classe $\mathcal{C}^p$).

Définition 8. Deux nappes paramétrées $\boldsymbol{f}_1 : U_1 \to \mathbb{R}^3$ et $\boldsymbol{f}_2 : U_2 \to \mathbb{R}^3$ représentent la même *nappe géométrique* (respectivement la même *nappe géométrique de classe* $\mathcal{C}^p$) si elles sont équivalentes (respectivement $\mathcal{C}^p$-équivalentes).

Remarque. En d'autres termes, une nappe géométrique est une classe d'équivalence de nappes paramétrées au sens de la définition 7.

Notons que deux paramétrages d'une même nappe géométrique ont même support.

Définition 9. Le *support* d'une nappe géométrique est le support de n'importe quel paramétrage.

Dans le module « Courbes paramétrées » du livre [L1], nous avons introduit la notion de courbe géométrique. Les courbes géométriques sont aux courbes paramétrées ce que les nappes géométriques sont aux nappes paramétrées : ce sont des classes d'équivalence de courbes paramétrées.

De même que l'on a parlé de courbe paramétrée tracée sur une nappe paramétrée, on peut parler de courbe géométrique tracée sur une nappe géométrique.

Définition 10. Une courbe géométrique Γ est *tracée sur une nappe géométrique* Σ s'il existe un paramétrage $\gamma : I \to \mathbb{R}^3$ de Γ tracé sur un paramétrage $\boldsymbol{f} : U \to \mathbb{R}^3$ de Σ.

Bien entendu, le support d'une courbe géométrique tracée sur une nappe géométrique est contenu dans le support de la nappe.
Notons finalement que si $\varphi : U_2 \to U_1$ est un homéomorphisme entre deux domaines de $\mathbb{R}^2$, alors φ est un difféomorphisme de classe $\mathcal{C}^p$ si, et seulement si, φ est une application de classe $\mathcal{C}^p$ dont la différentielle est inversible en tout point de U_2 (voir par exemple l'exercice 10 de la page 810 du livre [L2]). La différentielle $D_{\boldsymbol{u}}\varphi$ de φ en un point $\boldsymbol{u} \in U_2$ est inversible si, et seulement si, le jacobien de φ au point $\boldsymbol{u}$ est non nul (le jacobien est le déterminant de la matrice jacobienne). Si φ est un difféomorphisme de classe $\mathcal{C}^p$, son jacobien est une application continue du connexe U_2 à valeurs dans $\mathbb{R}^*$. Il garde donc un signe constant.

Définition 11. Un difféomorphisme $\varphi : U_2 \to U_1$ de classe $\mathcal{C}^p$ entre deux domaines de $\mathbb{R}^2$ est *direct* si son jacobien est positif et *indirect* sinon.

Exemples.

1. Si U est un domaine de $\mathbb{R}^2$, l'identité de U dans U est un difféomorphisme direct de classe $\mathcal{C}^\infty$.
2. L'application $(x, y) \mapsto (x, -y)$ de $\mathbb{R}^2$ dans $\mathbb{R}^2$ est un difféomorphisme indirect de classe $\mathcal{C}^\infty$.

Définition 12. Soit Σ une nappe géométrique de classe $\mathcal{C}^p$. Deux paramétrages $\boldsymbol{f}_1 : U_1 \to \mathbb{R}^3$ et $\boldsymbol{f}_2 : U_2 \to \mathbb{R}^3$ de Σ sont *$\mathcal{C}^p$-positivement équivalents* s'il existe un difféomorphisme $\varphi : U_2 \to U_1$ de classe $\mathcal{C}^p$ dont le jacobien est positif sur U_2.

L'inverse d'un difféomorphisme direct de classe $\mathcal{C}^p$ est un difféomorphisme direct de classe $\mathcal{C}^p$ (les jacobiens sont inverses l'un de l'autre, donc de même signe). La composée de deux difféomorphismes directs de classe $\mathcal{C}^p$ est un difféomorphisme direct de classe $\mathcal{C}^p$ (le jacobien de la composée est le produit des jacobiens). Il s'en suit que la relation « $\boldsymbol{f}_1 : U_1 \to \mathbb{R}^3$ et $\boldsymbol{f}_2 : U_2 \to \mathbb{R}^3$ sont $\mathcal{C}^p$-positivement équivalents » est une relation d'équivalence sur la nappe Σ. Cette relation d'équivalence a au plus deux classes d'équivalence. S'il y a deux classes distinctes, on dit que la nappe Σ est *orientable*. Orienter Σ consiste alors à choisir une des deux classes d'équivalence et à appeler ses représentants des *paramétrages directs* de la nappe Σ. Lorsqu'il n'y a qu'une seule classe, on dit que Σ est non orientable.

Exercice 2.
Soient Σ une nappe géométrique de classe $\mathcal{C}^p$ et $\boldsymbol{f} : U \to \mathbb{R}^3$ un paramétrage de Σ. Montrer que la nappe Σ est non orientable si, et seulement s'il existe un difféomorphisme indirect $\varphi : U \to U$ de classe $\mathcal{C}^p$ tel que $\boldsymbol{f} = \boldsymbol{f} \circ \varphi$.

Solution. S'il existe un tel difféomorphisme indirect $\varphi : U \to U$, il n'y a évidemment qu'une seule classe d'équivalence.
Réciproquement, supposons qu'il n'y ait qu'une seule classe d'équivalence. Soit $\sigma : \mathbb{R}^2 \to \mathbb{R}^2$ l'application définie par $\sigma(x, y) = (x, -y)$ et soit $V := \sigma(U)$. L'application $\sigma : V \to U$ est un difféomorphisme de classe $\mathcal{C}^\infty$ dont le jacobien est constant, égal à -1. C'est donc un difféomorphisme indirect. L'application $\boldsymbol{g} := \boldsymbol{f} \circ \sigma : V \to \mathbb{R}^3$ est donc un paramétrage de Σ, qui est, par hypothèse, $\mathcal{C}^p$-positivement équivalent à $\boldsymbol{f} : U \to \mathbb{R}^3$. Il existe donc un difféomorphisme direct $\psi : U \to V$ de classe $\mathcal{C}^p$ tel que $\boldsymbol{f} = \boldsymbol{g} \circ \psi = \boldsymbol{f} \circ (\sigma \circ \psi)$. Cela conclut la démonstration puisque $\varphi := \sigma \circ \psi : U \to U$ est un difféomorphisme indirect de classe $\mathcal{C}^p$.

Exercice 3.

Montrer que toute nappe géométrique simple Σ est orientable.
Solution. Soit $\boldsymbol{f} : U \to \mathbb{R}^3$ un paramétrage de Σ et soit $\varphi : U \to U$ un difféomorphisme tel que $\boldsymbol{f} = \boldsymbol{f} \circ \varphi$. Si $\boldsymbol{f}$ est injective, alors φ est l'identité, donc un difféomorphisme direct. D'après l'exercice précédent, Σ est orientable.

1.3 Plan tangent, espace tangent

Nous allons maintenant définir la notion de plan tangent à une nappe. Nous ferons pour cela l'hypothèse que la nappe est suffisamment régulière. À partir de maintenant, toutes les nappes considérées seront au minimum de classe $\mathcal{C}^1$.
Nous souhaitons étudier le comportement des nappes au voisinage de leurs points. Par exemple, un des problèmes qui nous intéresse est celui de l'existence d'un plan tangent. Un autre est celui de la position de la nappe par rapport au plan tangent. Nous supposerons pour cela que la nappe est régulière au voisinage de ce point. Cette hypothèse est assez restrictive mais c'est celle qui est habituellement adoptée.

Définition 13. Soit $\boldsymbol{f} : U \to \mathbb{R}^3$ une nappe paramétrée. Le point $\boldsymbol{m} := (\boldsymbol{u}, M)$ est un *point régulier* de la nappe si la différentielle $D_{\boldsymbol{u}}\boldsymbol{f}$ est de rang 2. Sinon, $\boldsymbol{m}$ est un *point singulier* de la nappe.

Exemple. Les points d'une nappe cartésienne de classe $\mathcal{C}^1$ sont des points réguliers. En effet, la matrice jacobienne :

$$J_{(x,y)}\boldsymbol{f} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ * & * \end{pmatrix}$$

est de rang 2 en tout point $(x, y) \in U$.

Exercice 4.

Déterminer les points réguliers et les points singuliers de la nappe paramétrée $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}(\theta, \varphi) := (\cos\theta\cos\varphi, \sin\theta\cos\varphi, \sin\varphi)$, dont le support est la sphère de centre $(0, 0, 0)$ et de rayon 1.

Solution. On a :

$$J_{(\theta,\varphi)}\boldsymbol{f} = \begin{pmatrix} -\sin\theta\cos\varphi & -\cos\theta\sin\varphi \\ \cos\theta\cos\varphi & -\sin\theta\sin\varphi \\ 0 & \cos\varphi \end{pmatrix}.$$

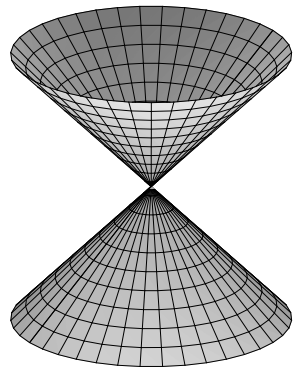
Le rang de la matrice est 2 si, et seulement si :

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \neq \begin{pmatrix} -\sin\theta\cos\varphi \\ \cos\theta\cos\varphi \\ 0 \end{pmatrix} \wedge \begin{pmatrix} -\cos\theta\sin\varphi \\ -\sin\theta\sin\varphi \\ \cos\varphi \end{pmatrix} = \begin{pmatrix} \cos\theta\cos^2\varphi \\ \sin\theta\cos^2\varphi \\ \sin\varphi\cos\varphi \end{pmatrix}.$$

Comme $\cos\theta$ et $\sin\theta$ ne peuvent pas s'annuler simultanément, le point $\boldsymbol{m} := \big(\theta, \varphi, \boldsymbol{f}(\theta,\varphi)\big)$ est singulier si, et seulement si, $\cos\varphi = 0$, donc si, et seulement si, $\varphi = \pi/2 \mod \pi$. Ces points correspondent aux « pôle nord et pôle sud » de la sphère où tous les méridiens se croisent.

Exercice 5.

Soit $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ la nappe paramétrée définie par $\boldsymbol{f}(t,\theta) = (t\cos\theta, t\sin\theta, t)$ (dont le support est un *cône de révolution* représenté sur la figure ci-contre).
Déterminer les point réguliers et les points singuliers de la nappe.



Solution. Si $\boldsymbol{m} := (\boldsymbol{u}, M)$ est un point de la nappe et si $M \neq (0,0,0)$, alors $\boldsymbol{m}$ est un point régulier. En effet :

$$J_{(t,\theta)}\boldsymbol{f} = \begin{pmatrix} \cos\theta & -t\sin\theta \\ \sin\theta & t\cos\theta \\ 1 & 0 \end{pmatrix}$$

et le déterminant de la matrice $\begin{pmatrix} \cos\theta & -t\sin\theta \\ \sin\theta & t\cos\theta \end{pmatrix}$ est égal à $t \neq 0$.
Sinon, $t = 0$ et le rang de la matrice jacobienne est égal à 1.

Venons-en maintenant à la notion de plan tangent.

Définition 14. Soient $\boldsymbol{f} : U \to \mathbb{R}^3$ une nappe paramétrée et $\boldsymbol{m} := (\boldsymbol{u}, M)$ un point régulier de la nappe. Le *plan tangent* à la nappe au point $\boldsymbol{m}$ est le plan affine P passant par M et dirigé par l'image de $D_{\boldsymbol{u}}\boldsymbol{f}$:

$$P := M + \operatorname{Im} D_{\boldsymbol{u}}\boldsymbol{f}.$$

Toute droite affine passant par M et contenue dans le plan tangent est une *tangente* à la nappe au point $\boldsymbol{m}$.

Exercice 6.

Soit $\boldsymbol{f} : (x,y) \in U \mapsto \big(x, y, F(x,y)\big) \in \mathbb{R}^3$ une nappe cartésienne de classe $\mathcal{C}^1$. Donner une équation paramétrique du plan tangent à la nappe en un point $\boldsymbol{m} := (\boldsymbol{u}, M)$ de la nappe.

Solution. Nous avons vu (exemple de la page précédente) que tous les points de la nappe sont des points réguliers. Le plan tangent en un point $\boldsymbol{m} := (\boldsymbol{u}, M)$ de la nappe est le plan passant par M et dirigé par :

$$\operatorname{Im} D_{\boldsymbol{u}}\boldsymbol{f} = \operatorname{Vect}\left(\frac{\partial \boldsymbol{f}}{\partial x}(\boldsymbol{u}), \frac{\partial \boldsymbol{f}}{\partial y}(\boldsymbol{u})\right) = \operatorname{Vect}\left(\begin{pmatrix} 1 \\ 0 \\ \dfrac{\partial F}{\partial x}(\boldsymbol{u}) \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \dfrac{\partial F}{\partial y}(\boldsymbol{u}) \end{pmatrix}\right).$$

Si $\boldsymbol{u} := (x_0, y_0)$, une équation paramétrique du plan tangent à la nappe au point $\boldsymbol{m}$ est donc :

$$\begin{cases} x = x_0 + r \\ y = y_0 + s \\ z = F(x_0, y_0) + r \cdot \dfrac{\partial F}{\partial x}(x_0, y_0) + s \cdot \dfrac{\partial F}{\partial y}(x_0, y_0), \end{cases} \qquad (r, s) \in \mathbb{R}^2.$$

Soit $\boldsymbol{f} := (f_1, f_2, f_3) : U \to \mathbb{R}^3$ une nappe régulière au point $\boldsymbol{m} := \big(\boldsymbol{u}, M_0\big)$. Pour qu'un point M de coordonnées (x, y, z) appartienne au plan tangent à la nappe au point $\boldsymbol{m}$, il faut et il suffit que la famille :

$$\left(\overrightarrow{M_0M}, \frac{\partial \boldsymbol{f}}{\partial u_1}(\boldsymbol{u}), \frac{\partial \boldsymbol{f}}{\partial u_2}(\boldsymbol{u})\right)$$

soit liée. Par conséquent, une équation du plan tangent à la nappe au point $\boldsymbol{m}$ est :

$$\det\begin{pmatrix} x - f_1(\boldsymbol{u}) & \dfrac{\partial f_1}{\partial u_1}(\boldsymbol{u}) & \dfrac{\partial f_1}{\partial u_2}(\boldsymbol{u}) \\ y - f_2(\boldsymbol{u}) & \dfrac{\partial f_2}{\partial u_1}(\boldsymbol{u}) & \dfrac{\partial f_2}{\partial u_2}(\boldsymbol{u}) \\ z - f_3(\boldsymbol{u}) & \dfrac{\partial f_3}{\partial u_1}(\boldsymbol{u}) & \dfrac{\partial f_3}{\partial u_2}(\boldsymbol{u}) \end{pmatrix} = 0.$$

Exercice 7.

Déterminer une équation des plans tangents dans le cas d'une nappe cartésienne de classe $\mathcal{C}^1$.

Solution. Si $\boldsymbol{f} : (x, y) \in U \mapsto \big(x, y, F(x, y)\big) \in \mathbb{R}^3$ est une nappe cartésienne, une équation du plan tangent au point $\big((x_0, y_0), M_0\big)$ avec $M_0 = (x_0, y_0, z_0)$ est :

$$\det\begin{pmatrix} x - x_0 & 1 & 0 \\ y - y_0 & 0 & 1 \\ z - z_0 & \dfrac{\partial F}{\partial x}(x_0, y_0) & \dfrac{\partial F}{\partial y}(x_0, y_0) \end{pmatrix} = 0.$$

Autrement dit :

$$z - z_0 = \frac{\partial F}{\partial x}(x_0, y_0) \cdot (x - x_0) + \frac{\partial F}{\partial y}(x_0, y_0) \cdot (y - y_0).$$

Il existe un lien fort entre les tangentes à une nappe en un point $\boldsymbol{m}$ de la nappe et les tangentes des courbes tracées sur la nappe passant par $\boldsymbol{m}$: les tangentes à la nappe sont les tangentes aux courbes $\mathcal{C}^1$ tracées sur la nappe. C'est ce qu'établit la proposition suivante. Rappelons auparavant qu'une courbe $\gamma : I \to \mathbb{R}^3$ qui est $\mathcal{C}^1$-tracée sur une nappe de classe $\mathcal{C}^1$ est elle-même de classe $\mathcal{C}^1$. Si elle admet un point régulier au temps t_0, alors elle admet une tangente au temps t_0 passant par le point $M := \gamma(t_0)$ et dirigée par $\mathrm{Vect}\big(\vec{\gamma}'(t_0)\big)$.

Proposition 1. Soient $\boldsymbol{f} : U \to \mathbb{R}^3$ une nappe paramétrée de classe $\mathcal{C}^1$ et $\boldsymbol{m}$ un point régulier de la nappe.

1. Si $\gamma : I \to \mathbb{R}^3$ est une courbe régulière $\mathcal{C}^1$-tracée sur la nappe et passant par le point $\boldsymbol{m}$ au temps $t_0 \in I$, alors la tangente à γ au temps t_0 est une tangente à la nappe au point $\boldsymbol{m}$.
2. Si D est une tangente à la nappe au point $\boldsymbol{m}$, alors il existe une courbe régulière $\gamma : I \to \mathbb{R}^3$ qui est $\mathcal{C}^1$-tracée sur la nappe, qui passe par le point $\boldsymbol{m}$ au temps $t_0 \in I$ et dont la tangente au temps t_0 est la droite D.

Démonstration. Notons $\boldsymbol{m} = (\boldsymbol{u}, M)$ avec $\boldsymbol{u} \in U$ et $M := \boldsymbol{f}(\boldsymbol{u})$.

1. Écrivons $\gamma = \boldsymbol{f} \circ \gamma_1$ avec $\gamma_1 : I \to U$ une courbe de classe $\mathcal{C}^1$. Alors :
$$\gamma_1(t_0) = \boldsymbol{u} \quad \text{et} \quad \vec{\gamma}'(t_0) = D_{\boldsymbol{u}}\boldsymbol{f}\big(\vec{\gamma}_1'(t_0)\big).$$
On voit donc que $\vec{\gamma}'(t_0)$ appartient à l'image de la différentielle $D_{\boldsymbol{u}}\boldsymbol{f}$. Par conséquent, la tangente à γ au temps t_0 est contenue dans le plan tangent à la nappe au point $\boldsymbol{m}$.
2. Soit D une tangente à la nappe au point $\boldsymbol{m}$ et soit $\vec{v} \in \operatorname{Im} D_{\boldsymbol{u}}\boldsymbol{f}$ un vecteur qui dirige D. Alors, $\vec{v} = D_{\boldsymbol{u}}\boldsymbol{f}(\vec{w})$ avec $\vec{w} \in \mathbb{R}^2$. Soit $\varepsilon > 0$ suffisamment petit pour que l'intervalle $\big\{\boldsymbol{u} + t\vec{w} \;\big|\; t \in]-\varepsilon, \varepsilon[\big\}$ soit contenu dans l'ouvert U. Posons $I :=]-\varepsilon, \varepsilon[$ et définissons $\gamma : I \to \mathbb{R}^3$ par :
$$\gamma(t) = \boldsymbol{f}\big(\boldsymbol{u} + t\vec{w}\big).$$
Alors γ est un arc $\mathcal{C}^1$-tracé sur la nappe, qui passe par le point $\boldsymbol{m}$ au temps t_0.
De plus :
$$\gamma(t_0) = \boldsymbol{f}(\boldsymbol{u}) = M \quad \text{et} \quad \vec{\gamma}'(t_0) = D_{\boldsymbol{u}}\boldsymbol{f}(\vec{w}) = \vec{v}.$$
Par conséquent, la tangente à γ au temps t_0 est la droite D. ∎

Il est également très utile d'introduire la notion d'*espace tangent*.

Définition 15. L'*espace tangent* à une nappe paramétrée en un point régulier $\boldsymbol{m}$ est l'espace vectoriel qui dirige le plan tangent.

Il est plus habituel de définir l'espace tangent à une nappe paramétrée en un point régulier $\boldsymbol{m}$, comme l'ensemble de tous les vecteurs $\vec{v} \in \mathbb{R}^3$ qui sont de la forme $\vec{v} = \vec{\gamma}'(t_0)$ avec $\gamma : I \to \mathbb{R}^3$ une courbe paramétrée régulière $\mathcal{C}^1$-tracée sur la nappe et qui passe

par le point $\boldsymbol{m}$ au temps t_0. L'équivalence des définitions est justifiée par la proposition précédente.
Observons que la notion de point régulier ou singulier ne dépend que de la nappe géométrique, pas du paramétrage. En effet, soient $\boldsymbol{f}_1 : U_1 \to \mathbb{R}^3$ et $\boldsymbol{f}_2 : U_2 \to \mathbb{R}^3$ deux nappes paramétrées et soit $\varphi : U_2 \to U_1$ un difféomorphisme de classe $\mathcal{C}^1$ tel que $\boldsymbol{f}_2 = \boldsymbol{f}_1 \circ \varphi$. Soient $\boldsymbol{u}_1$ un point de U_1 et $\boldsymbol{u}_2 := \varphi^{-1}(\boldsymbol{u}_1)$. Le rang de la différentielle $D_{\boldsymbol{u}_1}\boldsymbol{f}_1$ est égal au rang de la différentielle $D_{\boldsymbol{u}_2}\boldsymbol{f}_2$ puisque, d'après la règle de la chaîne, on a :

$$D_{\boldsymbol{u}_2}\boldsymbol{f}_2 := D_{\boldsymbol{u}_1}\boldsymbol{f}_1 \circ D_{\boldsymbol{u}_2}\varphi \quad \text{et} \quad D_{\boldsymbol{u}_2}\varphi \text{ est inversible.}$$

De plus, l'égalité précédente montre que l'image de la différentielle de $\boldsymbol{f}_1$ en $\boldsymbol{u}_1$ coïncide avec l'image de la différentielle de $\boldsymbol{f}_2$ en $\boldsymbol{u}_2$. Ces observations nous autorisent à donner la définition suivante.

Définition 16. Une nappe géométrique Σ est *régulière* si pour un (et donc pour tout) paramétrage $\boldsymbol{f} : U \to \mathbb{R}^3$, tous les points sont réguliers. Si Σ est une nappe géométrique régulière et simple et si M est un point du support de Σ, alors :

- L'*espace tangent* $T_M\Sigma$ à Σ au point M est l'espace tangent de n'importe quel paramétrage $\boldsymbol{f} : U \to \mathbb{R}^3$ au point $\left(\boldsymbol{f}^{-1}(M), M\right)$.
- Le *plan tangent* à Σ au point M est le plan affine $M + T_M\Sigma$.

1.4 Position par rapport au plan tangent

Nous allons maintenant voir qu'au voisinage d'un point régulier, une nappe de classe $\mathcal{C}^p$ est le graphe d'une fonction de classe $\mathcal{C}^p$ au dessus de son plan tangent.

Théorème 2. Soient $\boldsymbol{f} : U \to \mathbb{R}$ une nappe paramétrée de classe $\mathcal{C}^p$ et $\boldsymbol{m} := (\boldsymbol{u}, M)$ un point régulier de la nappe. Soit $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ une base de l'espace vectoriel $\mathbb{R}^3$ telle que l'espace tangent à la nappe en $\boldsymbol{m}$ ne contienne pas $\vec{e}_3$. Alors, il existe un domaine $U_1 \subset U$ contenant $\boldsymbol{u}$, un domaine $V \subset \mathbb{R}^2$ contenant $(0,0)$ et une nappe paramétrée $\boldsymbol{g} : V \to \mathbb{R}^3$ tels que :

- $\boldsymbol{f} : U_1 \to \mathbb{R}^3$ et $\boldsymbol{g} : V \to \mathbb{R}^3$ représentent la même nappe géométrique de classe $\mathcal{C}^p$,
- $\boldsymbol{g}(x_1, x_2) = M + x_1 \cdot \vec{e}_1 + x_2 \cdot \vec{e}_2 + F(x_1, x_2) \cdot \vec{e}_3$ avec $F : V \to \mathbb{R}$ une fonction de classe $\mathcal{C}^p$ qui s'annule en $(0,0)$.

Remarque. Autrement dit, dans le repère $(M; \vec{e}_1, \vec{e}_2, \vec{e}_3)$, il existe une nappe cartésienne qui représente la même nappe géométrique que $\boldsymbol{f}$ au voisinage de M.

Démonstration. Plaçons nous dans le repère $(M; \vec{e}_1, \vec{e}_2, \vec{e}_3)$. L'application $\boldsymbol{f}$ prend alors la forme :

$$\boldsymbol{f}(u_1, u_2) = X_1(u_1, u_2) \cdot \vec{e}_1 + X_2(u_1, u_2) \cdot \vec{e}_2 + X_3(u_1, u_2) \cdot \vec{e}_3$$

avec $X_1 : U \to \mathbb{R}$, $X_2 : U \to \mathbb{R}$ et $X_3 : U \to \mathbb{R}$ des fonctions de classe $\mathcal{C}^p$ qui s'annulent en $(u_1, u_2) = (0,0)$. L'image $D_{\boldsymbol{u}}\boldsymbol{f}$ ne contenant pas le vecteur $\vec{e}_3$, le déterminant de la

matrice :

$$\begin{pmatrix} \dfrac{\partial X_1}{\partial u_1} & \dfrac{\partial X_1}{\partial u_2} & 0 \\ \dfrac{\partial X_2}{\partial u_1} & \dfrac{\partial X_2}{\partial u_2} & 0 \\ \dfrac{\partial X_3}{\partial u_1} & \dfrac{\partial X_3}{\partial u_2} & 1 \end{pmatrix}$$

est non nul. La matrice des dérivées partielles :

$$\begin{pmatrix} \dfrac{\partial X_1}{\partial u_1} & \dfrac{\partial X_1}{\partial u_2} \\ \dfrac{\partial X_2}{\partial u_1} & \dfrac{\partial X_2}{\partial u_2} \end{pmatrix}$$

est donc inversible.

D'après le théorème d'inversion locale du livre [L2], il existe un domaine $U_1 \subset U$ contenant $\boldsymbol{u}$ et un domaine V contenant $(0,0)$ tels que l'application $\varphi : U_1 \to V$ définie par $\varphi := (X_1, X_2)$ soit un difféomorphisme de classe $\mathcal{C}^1$.

On a alors $X_1 \circ \varphi^{-1}(x_1, x_2) = x_1$ et $X_2 \circ \varphi^{-1}(x_1, x_2) = x_2$. Comme φ est de classe $\mathcal{C}^p$, c'est automatiquement un difféomorphisme de classe $\mathcal{C}^p$ (voir par exemple l'exercice 10 de la page 810 du livre [L2].

Soit $\boldsymbol{g} : V \to \mathbb{R}^3$ la nappe paramétrée définie par $\boldsymbol{g} := \boldsymbol{f} \circ \varphi^{-1}$. Par construction, les nappes $\boldsymbol{f}_{|_{U_1}}$ et $\boldsymbol{g}$ sont $\mathcal{C}^p$ équivalentes. De plus, on a :

$$\boldsymbol{g} := \boldsymbol{f} \circ \varphi^{-1} = M + \left(X_1 \circ \varphi^{-1}\right) \cdot \vec{e}_1 + \left(X_2 \circ \varphi^{-1}\right) \cdot \vec{e}_2 + \left(X_3 \circ \varphi^{-1}\right) \cdot \vec{e}_3.$$

On a donc, comme requis :

$$\boldsymbol{g}(x_1, x_2) = M + x_1 \cdot \vec{e}_1 + x_2 \cdot \vec{e}_2 + F(x_1, x_2) \cdot \vec{e}_3 \quad \text{avec} \quad F := X_3 \circ \varphi^{-1}$$

une fonction de classe $\mathcal{C}^p$ sur V qui s'annule en $(0,0)$. ■

Le résultat précédent s'applique en particulier lorsque $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ est une base de l'espace vectoriel $\mathbb{R}^3$ telle que $(\vec{e}_1, \vec{e}_2)$ soit une base de l'espace tangent à la nappe au point $\boldsymbol{m}$. En travaillant dans le repère $(M; \vec{e}_1, \vec{e}_2, \vec{e}_3)$, on voit qu'étudier la position de la nappe paramétrée par rapport à son plan tangent, au voisinage du point $\boldsymbol{m}$, revient à étudier le signe de $F(x_1, x_2)$ pour (x_1, x_2) proche de $(0,0)$. Le plan tangent à la nappe au point $\boldsymbol{m}$ est le plan d'équation $X_3 = 0$. Les deux vecteurs :

$$\vec{e}_1 + \frac{\partial F}{\partial x_1}(0,0) \cdot \vec{e}_3 \quad \text{et} \quad \vec{e}_2 + \frac{\partial F}{\partial x_2}(0,0) \cdot \vec{e}_3$$

appartiennent à l'espace tangent, ce qui montre que la différentielle de F s'annule en $(0,0)$. Pour étudier la position de la nappe par rapport à son plan tangent, nous ferons l'hypothèse que la nappe paramétrée, et donc F, est de classe $\mathcal{C}^2$. On dispose alors de la formule de Taylor à l'ordre 2. Dans ce cas, l'étude du signe de F a été détaillée dans le module « Fonctions de plusieurs variables » du livre [L2]. Si l'on pose :

$$r := \frac{\partial^2 F}{\partial x_1^2}(0,0), \quad s := \frac{\partial^2 F}{\partial x_1 x_2}(0,0) \quad \text{et} \quad t := \frac{\partial^2 F}{\partial x_2^2}(0,0),$$

alors $F(x_1, x_2) = \dfrac{1}{2}\big(rx_1^2 + 2sx_1x_2 + tx_2^2\big) + \mathrm{o}\big(\|(x_1, x_2)\|^2\big)$.

- Si $rt - s^2 > 0$, alors la fonction F présente un extremum local en $(0, 0)$. Si $r + t > 0$, alors $F(x, y) > 0$ au voisinage de l'origine (excepté bien sûr à l'origine). Au voisinage du point $\boldsymbol{m}$, le support de la nappe paramétrée reste alors du même côté du plan tangent que $\vec{e}_3$. Si $r + t < 0$, alors, au voisinage du point $\boldsymbol{m}$, le support de la nappe paramétrée reste de l'autre côté du plan tangent. On dit que $\boldsymbol{m}$ est un *point elliptique* de la nappe.
- Si $rt - s^2 < 0$ alors F change de signe au voisinage de l'origine et la nappe paramétrée traverse donc le plan tangent. Nous donnerons plus loin une description plus précise de l'intersection locale du plan tangent avec la nappe paramétrée. On dit que $\boldsymbol{m}$ est un *point hyperbolique* de la nappe.
- Si $rt - s^2 = 0$, alors on ne peut rien conclure sans pousser plus en avant l'analyse, en faisant un développement de Taylor d'ordre supérieur à 2 par exemple si $\boldsymbol{f}$ est de classe $\mathcal{C}^p$ avec $p \geqslant 3$. On dit que $\boldsymbol{m}$ est un *point parabolique* de la nappe.

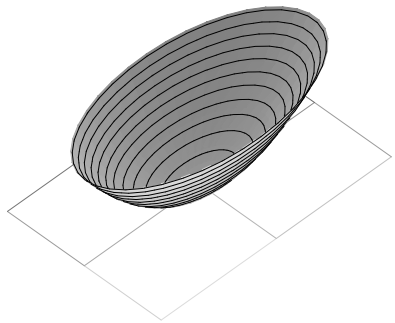
Remarque. Dans l'étude précédente, plutôt que de tester le signe de $r+t$, on pourrait tester le signe de r ou de t. En effet, dans le cas où $rt - s^2 > 0$, le signe de r, t et $r+t$ est le même. Signalons cependant que $rt - s^2$ est le déterminant de la matrice hessienne et $r + t$ est la trace de la matrice hessienne. Il est donc plus naturel de considérer le signe de $r + t$.

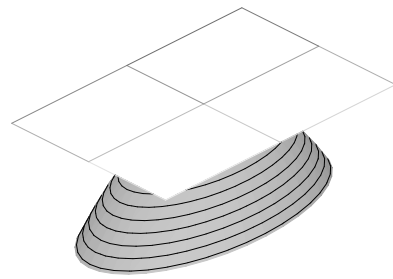
Exemples. Commençons par considérer le cas d'une nappe cartésienne de la forme :

$$\boldsymbol{f}(x, y) = \big(x, y, \Phi(x, y)\big) \quad \text{avec} \quad \Phi(x, y) = rx^2 + 2sxy + ty^2$$

une forme quadratique sur $\mathbb{R}^2$.

1. Premier cas : $rt - s^2 > 0$. La forme quadratique Φ est non dégénérée positive (respectivement négative) si $r + t > 0$ (respectivement $r + t < 0$). Sa signature est $(2, 0)$ (respectivement $(0, 2)$).

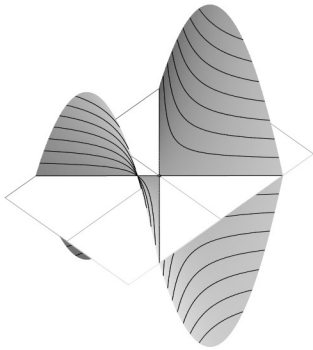




   - Si $r + t > 0$, la forme quadratique Φ présente un minimum (global) en $(0, 0)$ et le support de la nappe se trouve au dessus du plan tangent d'équation $z = 0$. L'intersection du support de la nappe et d'un plan horizontal d'équation $z = z_0$ avec $z_0 > 0$ est une courbe dont la projection sur le plan tangent parallèlement à l'axe (Oz) est la courbe d'équation $\Phi(x, y) = z_0$. Nous avons vu dans le

module « Géométrie dans les espaces affines » du livre [L1]), qu'il s'agit d'une ellipse centrée à l'origine. Le support de la nappe est un *paraboloïde elliptique*.

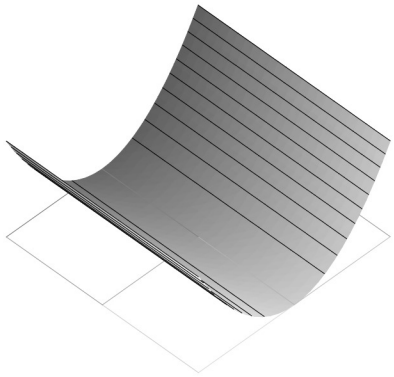
- Si $r + t < 0$, la forme quadratique Φ présente un maximum (global) en $(0, 0)$ et le support de la nappe se trouve au dessous du plan tangent d'équation $z = 0$. L'intersection du support de la nappe et d'un plan horizontal d'équation $z = z_0$ avec $z_0 < 0$ est ici encore une ellipse et le support de la nappe est un paraboloïde elliptique.

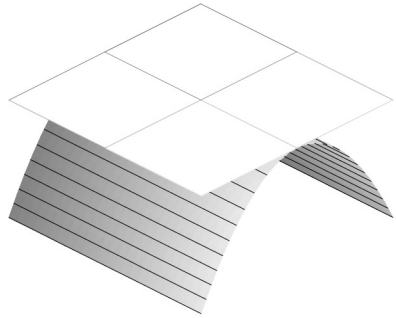
2. 

Deuxième cas : $rt - s^2 < 0$. La forme quadratique Φ est non dégénérée de signature $(1, 1)$. L'intersection du support de la nappe avec le plan tangent d'équation $z = 0$ est la réunion de deux droites distinctes passant par l'origine. L'intersection du support de la nappe et d'un plan horizontal d'équation $z = z_0$ avec $z_0 \neq 0$ est une hyperbole. Le support de la nappe est un *paraboloïde hyperbolique*.

3. Troisième cas : $rt - s^2 = 0$. La forme quadratique Φ est dégénérée nulle ou bien de signature $(1, 0)$ ou bien de signature $(0, 1)$.

   - Si la forme Φ est nulle, le support de la nappe paramétrée est le plan d'équation $z = 0$.





   - Si la signature est $(1, 0)$, $\Phi(x, y) = (ax + by)^2$ avec $(a, b) \in \mathbb{R}^2$. Le support de la nappe se trouve au dessus du plan tangent d'équation $z = 0$. L'intersection du support de la nappe avec le plan tangent est la droite d'équation $\{ax + by = 0 \text{ et } z = 0\}$. On peut faire le changement de variable $X = ax + by$ et $Y = -bx + ay$ qui est une similitude centrée en 0. La nappe paramétrée prend alors la forme $\boldsymbol{g}(X, Y) = (X, Y, X^2)$. Le support de la nappe s'appelle un *cylindre parabolique*.
   - Si la signature est $(0, 1)$, $\Phi(x, y) = -(ax + by)^2$ avec $(a, b) \in \mathbb{R}^2$. On peut faire le changement de variable $X := ax + by$ et $Y := -bx + ay$. La nappe paramétrée prend alors la forme $\boldsymbol{g}(X, Y) = (X, Y, -X^2)$.

Exercice 8.

Soit $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ la nappe cartésienne définie par $\boldsymbol{f}(x,y) := \left(x, y, \frac{x^2}{\alpha^2} - \frac{y^2}{\beta^2}\right)$.

Montrer que l'intersection du plan tangent avec le support de la nappe en un point quelconque est la réunion de deux droites sécantes.

Solution. Le plan tangent à la nappe au point $\big((x_0,y_0), M_0\big)$ avec $M := (x_0, y_0, z_0)$ est le plan d'équation :

$$(z - z_0) = \frac{2x_0}{\alpha^2}(x - x_0) - \frac{2y_0}{\beta^2}(y - y_0).$$

Un point $M := (x,y,z)$ appartient au plan tangent et au support de la nappe si, et seulement si :

$$\begin{cases} z - z_0 = \dfrac{2x_0}{\alpha^2}(x - x_0) - \dfrac{2y_0}{\beta^2}(y - y_0) \\ z - z_0 = \dfrac{x^2 - x_0^2}{\alpha^2} - \dfrac{y^2 - y_0^2}{\beta^2} \end{cases}$$

donc (en soustrayant les deux équations) si, et seulement si :

$$\begin{cases} z - z_0 = \dfrac{2x_0}{\alpha^2}(x - x_0) - \dfrac{2y_0}{\beta^2}(y - y_0) \\ 0 = \dfrac{(x - x_0)^2}{\alpha^2} - \dfrac{(y - y_0)^2}{\beta^2}. \end{cases}$$

L'intersection du support de la nappe et du plan tangent est donc la réunion des deux droites :

$$D_1 := \begin{cases} z - z_0 = \dfrac{2x_0}{\alpha^2}(x - x_0) - \dfrac{2y_0}{\beta^2}(y - y_0) \\ \dfrac{x - x_0}{\alpha} - \dfrac{y - y_0}{\beta} = 0 \end{cases}$$

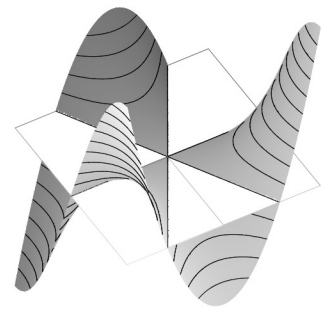
$$D_2 := \begin{cases} z - z_0 = \dfrac{2x_0}{\alpha^2}(x - x_0) - \dfrac{2y_0}{\beta^2}(y - y_0) \\ \dfrac{x - x_0}{\alpha} + \dfrac{y - y_0}{\beta} = 0. \end{cases}$$

Exercice 9.

Soit $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ la nappe cartésienne définie par $\boldsymbol{f}(x,y) = (x, y, 3x^2y - y^3)$.

Étudier la position du support de la nappe par rapport au plan d'équation $z = 0$.

Solution. Observons que $3x^2y - y^3 = \mathrm{Im}\big((x+iy)^3\big)$.



La partie imaginaire de $(x+iy)^3$ s'annule le long des trois droites $\mathbb{R}$, $e^{2i\pi/3} \cdot \mathbb{R}$ et $e^{-2i\pi/3} \cdot \mathbb{R}$.

Ces trois droites partagent le plan en six secteurs où la fonction $(x,y) \mapsto 3x^2y - y^3$ prend alternativement des valeurs positives et des valeurs négatives. Si l'argument de $x + iy$ se trouve dans l'un des trois intervalles $]0, \pi/3[$, $]2\pi/3, \pi[$, ou $]4\pi/3, 5\pi/3[$, le point $(x, y, 3x^2y - y^3)$ se situe au dessus du plan d'équation $z = 0$. Si l'argument de $x + iy$ se trouve dans l'un des trois intervalles $]\pi/3, 2\pi/3[$, $]\pi, 4\pi/3[$, ou $]5\pi/3, 2\pi[$, le point $(x, y, 3x^2y - y^3)$ se situe en dessous du plan d'équation $z = 0$.

Supposons maintenant que $\boldsymbol{f} : (x,y) \in U \to \big(x,y,F(x,y)\big) \in \mathbb{R}^3$ soit une nappe cartésienne de classe $\mathcal{C}^3$. Supposons de plus que $F(x,y) = \Phi(x,y) + \mathrm{o}\big(\|(x,y)\|^2\big)$ avec Φ une forme quadratique non dégénérée. Nous allons montrer qu'au voisinage de l'origine, l'intersection du support de la nappe avec les plans d'équation $z = z_0$ ont la même allure que les lignes de niveaux de la forme quadratique Φ, c'est-à-dire des ellipses ou des arcs d'hyperboles. Nous utiliserons pour cela le *Lemme de Morse*. Nous l'énonçons dans $\mathbb{R}^n$. Nous l'utiliserons tout de suite dans $\mathbb{R}^2$ et plus loin dans $\mathbb{R}^3$.

Proposition 3 (Lemme de Morse). Soient U un domaine de $\mathbb{R}^n$ contenant l'origine $\mathbf{0}$ et $F : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^3$ telle que $F(\boldsymbol{u}) = \Phi(\boldsymbol{u}) + \mathrm{o}\big(\|\boldsymbol{u}\|^2\big)$ avec Φ une forme quadratique non dégénérée. Alors, il existe des domaines $V \subset U$ et $W \subset \mathbb{R}^n$ contenant $\mathbf{0}$ et un difféomorphisme $\psi : V \to W$ de classe $\mathcal{C}^1$ tels que :

$$F_{|_V} = \Phi \circ \psi.$$

Dans le cas où $F : U \subset \mathbb{R}^2 \to \mathbb{R}$ est une fonction de deux variables, ψ est un homéomorphisme entre les deux ensembles :

$$\big\{(x,y) \in V \;\big|\; F(x,y) = z_0\big\} \quad \text{et} \quad \big\{(X,Y) \in W \;\big|\; \Phi(X,Y) = z_0\big\}.$$

Cela montre qu'au voisinage de l'origine, l'intersection du support de la nappe cartésienne $(x,y) \mapsto \big(x,y,F(x,y)\big)$ avec le plan d'équation $z = z_0$ est homéomorphe à l'intersection de la nappe cartésienne $(x,y) \mapsto \big(x,y,\Phi(x,y)\big)$ avec le même plan d'équation $z = z_0$.
En particulier, si la forme quadratique est de signature $(1,1)$, au voisinage de l'origine, l'intersection du support de la nappe cartésienne avec le plan tangent d'équation $z = 0$ est la réunion de deux arcs réguliers qui se coupent en $(0,0)$ avec des tangentes distinctes.
Le lecteur pourra dans un premier temps laisser la démonstration de côté.

Démonstration. Rappelons que $M_n(\mathbb{R}) = S_n(\mathbb{R}) \oplus A_n(\mathbb{R})$ où $S_n(\mathbb{R})$ (respectivement $A_n(\mathbb{R})$) est le sous-espace vectoriel de $M_n(\mathbb{R})$ composé des matrices symétriques (respectivement antisymétriques).
Soit $U' \subset U$ une boule centrée en $\mathbf{0}$.

Lemme 4. Il existe une fonction $A : U' \to S_n(\mathbb{R})$ de classe $\mathcal{C}^1$ telle que $A(\mathbf{0})$ soit la hessienne de F en $\mathbf{0}$ et telle que pour tout $\boldsymbol{u} \in U'$, on ait $F(\boldsymbol{u}) = {}^t\boldsymbol{u} \cdot A(\boldsymbol{u}) \cdot \boldsymbol{u}$.

Démonstration. Étant donné $\boldsymbol{u} \in U'$, soit $g : [0,1] \to \mathbb{R}$ la fonction définie par $g(t) = F(t\boldsymbol{u})$. La fonction g est de classe $\mathcal{C}^3$ sur $[0,1]$. De plus, pour $t_0 \in \,]0,1[$, on a :

$$g(t_0 + t) = F(t_0\boldsymbol{u} + t\boldsymbol{u}) = F(t_0\boldsymbol{u}) + D_{t_0\boldsymbol{u}}F(t\boldsymbol{u}) + \frac{1}{2}Q_{t_0\boldsymbol{u}}F(t\boldsymbol{u}) + \mathrm{o}\big(\|t\boldsymbol{u}\|^2\big)$$

avec $Q_{t_0\boldsymbol{u}}F$ la forme quadratique dont la matrice est la hessienne $H_{t_0\boldsymbol{u}}F$ des dérivées secondes de F :

$$\big(H_{t_0\boldsymbol{u}}F\big)_{ij} := \frac{\partial^2 F}{\partial u_i \partial u_j}(t_0\boldsymbol{u}).$$

On a donc :

$$g(0) = F(0 \cdot \boldsymbol{u}) = 0, \quad g'(0) = D_{\mathbf{0}}F(0 \cdot \boldsymbol{u}) = 0 \quad \text{et} \quad g''(t) = {}^t\boldsymbol{u} \cdot H_{t\boldsymbol{u}}F \cdot \boldsymbol{u}.$$

La formule de Taylor avec reste intégral donne :

$$F(\boldsymbol{u}) = g(1) = g(0) + g'(0) + \int_0^1 (1-t)g''(t)\,\mathrm{d}t = {}^t\boldsymbol{u} \cdot A(\boldsymbol{u}) \cdot \boldsymbol{u}$$

avec $A : U' \to S_n(\mathbb{R})$ l'application définie par :

$$A(\boldsymbol{u}) = \int_0^1 (1-t)H_{t\boldsymbol{u}}F\,\mathrm{d}t.$$

Étant donné que l'application F est de classe $\mathcal{C}^3$, l'application $(\boldsymbol{u}, t) \mapsto (1-t)H_{t\boldsymbol{u}}F$ est continue sur $U' \times [0,1]$ et de classe $\mathcal{C}^1$ par rapport à $\boldsymbol{u}$, avec des dérivées partielles continues sur $U' \times [0,1]$. On peut donc dériver sous le signe intégral (voir le corollaire 131 de la page 720 du livre [L2]). L'application $A : U' \to S_n(\mathbb{R})$ est donc de classe $\mathcal{C}^1$. ■

Lemme 5. Il existe un domaine $U'' \subset U'$ contenant $\mathbf{0}$ et une application $B : U'' \to M_n(\mathbb{R})$ de classe $\mathcal{C}^1$ telle que $B(\mathbf{0}) = \mathrm{I}_n$ et telle que pour tout $\boldsymbol{u} \in U''$, on ait $A(\boldsymbol{u}) = {}^tB(\boldsymbol{u}) \cdot A(\mathbf{0}) \cdot B(\boldsymbol{u})$.

Démonstration. Observons que $A_0 := A(\mathbf{0})$ est une matrice symétrique inversible (la forme quadratique Φ est non dégénérée). Soit $L : M_n(\mathbb{R}) \to S_n(\mathbb{R}) \times A_n(\mathbb{R})$ l'application définie par :

$$L(M) := \big({}^tM \cdot A_0 \cdot M, {}^t(A_0M) - A_0M\big).$$

C'est une application de classe $\mathcal{C}^\infty$. On a $L(\mathrm{I}_n) = \big(A_0, \mathbf{0}\big)$ et :

$$D_{\mathrm{I}_n}L(M) = \big({}^tM \cdot A_0 + A_0 \cdot {}^tM, {}^t(AM) - AM\big) = \big({}^t(A_0M) + A_0M, {}^t(A_0M) - A_0M\big).$$

La matrice M est donc dans le noyau de $D_{\mathrm{I}_n}L$ si, et seulement si, ${}^t(A_0M) = A_0M = -A_0M$, donc si, et seulement si, $A_0M = \mathbf{0}$, donc si, et seulement si, $M = \mathbf{0}$. La différentielle de L en I_n est donc une application linéaire injective entre deux espaces vectoriels de même dimension. Elle est donc inversible et l'on peut appliquer le théorème d'inversion locale. Par conséquent, il existe un domaine $U'' \subset U'$ contenant $\mathbf{0}$ et une application $B : U \to M_n(\mathbb{R})$ telle que pour tout $\boldsymbol{u} \in U''$, on ait $L\big(B(\boldsymbol{u})\big) = \big(A(\boldsymbol{u}), \mathbf{0}\big)$. On a alors, comme requis, $A(\boldsymbol{u}) = {}^tB(\boldsymbol{u}) \cdot A_0 \cdot B(\boldsymbol{u})$. ■

Considérons maintenant l'application $\psi : U'' \to \mathbb{R}^2$ définie par $\psi(\boldsymbol{u}) = B(\boldsymbol{u}) \cdot \boldsymbol{u}$. Comme B est de classe $\mathcal{C}^1$, il en est de même pour ψ. Comme $B(\mathbf{0}) = \mathrm{I}_n$, la différentielle de ψ en $(0,0)$ est $D_{\mathbf{0}}\psi = \mathrm{I}_n$. Le théorème d'inversion locale implique alors qu'il existe des domaines $V \subset U''$ et $W \subset \mathbb{R}^2$ contenant $\mathbf{0}$, tels que $\psi : V \to W$ soit un difféomorphisme de classe $\mathcal{C}^1$. Pour tout $\boldsymbol{u} \in V$, on a alors :

$$F(\boldsymbol{u}) = {}^t\boldsymbol{u} \cdot A(\boldsymbol{u}) \cdot \boldsymbol{u} = {}^t\boldsymbol{u}{}^tB(\boldsymbol{u}) \cdot A(\mathbf{0}) \cdot B(\boldsymbol{u}) \cdot \boldsymbol{u} = {}^t\psi(\boldsymbol{u}) \cdot A(\mathbf{0}) \cdot \psi(\boldsymbol{u}) = \Phi \circ \psi(\boldsymbol{u}).$$

2 Surfaces implicites

Un des principaux avantages à décrire une surface à l'aide d'un paramétrage est que l'on peut facilement trouver des points qui appartiennent à la surface. On peut par exemple représenter la surface en traçant des courbes sur la surface.
Un des inconvénients est qu'il est en général difficile de vérifier si un point donné appartient à la surface. Déterminer l'intersection de deux surfaces n'est pas toujours aisé si les deux surfaces sont les supports de nappes paramétrées.
Une autre approche consiste à décrire les surfaces de $\mathbb{R}^3$ comme le lieu où une fonction $F : U \subset \mathbb{R}^3 \to \mathbb{R}$ s'annule. Dans toute cette partie, U désigne un domaine de $\mathbb{R}^3$.

2.1 Définitions

Définition 17. Si $F : U \to \mathbb{R}$ est une fonction continue et si $\Sigma := F^{-1}\{0\}$, alors le couple (F, Σ) s'appelle la *surface d'équation* $F(M) = 0$. Si F est de classe $\mathcal{C}^p$ on dit que la surface est de classe $\mathcal{C}^p$. L'ensemble Σ s'appelle le support de la surface.

Remarques.

1. Le support peut être vide, comme par exemple pour la surface d'équation :
$$x^2 + y^2 + z^2 = -1.$$
2. Le support peut être égal à $\mathbb{R}^3$ si F est identiquement nulle.
3. Le support peut être le même pour deux fonctions F_1 et F_2 distinctes. Par exemple, si $F_1 : U \to \mathbb{R}$ est une fonction continue et si $F_2 = F_1^2$, alors $\Sigma = F_1^{-1}\{0\} = F_2^{-1}\{0\}$. Cependant, les deux surfaces (F_1, Σ) et (F_2, Σ) sont distinctes puisque les fonctions F_1 et F_2 sont distinctes.

Si l'on ne se donne pas des conditions de régularité, le support peut ne pas ressembler à une surface. Par exemple, comme nous venons de le signaler, il peut être égal à $\mathbb{R}^3$ tout entier.
Rappelons que si $F : U \to \mathbb{R}$ est une fonction de classe $\mathcal{C}^p$, alors la différentielle $D_M F$ de F en un point $M \in U$ est une forme linéaire sur l'espace vectoriel $\mathbb{R}^3$.

Définition 18. Soit (F, Σ) une surface de classe $\mathcal{C}^p$. Le point $M \in \Sigma$ est un *point régulier* si $D_M F$ n'est pas la forme nulle. Le point M est un point singulier si $D_M F$ est la forme nulle.

Le point $M \in \Sigma$ de coordonnées (x_0, y_0, z_0) est un point singulier si, et seulement si :
$$\frac{\partial F}{\partial x}(x_0, y_0, z_0) = \frac{\partial F}{\partial y}(x_0, y_0, z_0) = \frac{\partial F}{\partial z}(x_0, y_0, z_0) = 0.$$
Pour montrer que le point M est régulier, il suffit donc de montrer que l'une au moins des trois dérivées partielles premières est non nulle.

Exercice 10.
Déterminer les points singuliers de la surface d'équation $x^2 + y^2 - z^2 = 0$. Cette surface s'appelle un cône de révolution d'axe (Oz).

Solution.
Les dérivées partielles de la fonction $F : \mathbb{R}^3 \to \mathbb{R}$ définie par $F(x, y, z) := x^2 + y^2 - z^2$ sont :

$$\frac{\partial F}{\partial x}(x, y, z) = 2x, \quad \frac{\partial F}{\partial y}(x, y, z) = 2y \quad \text{et} \quad \frac{\partial F}{\partial z}(x, y, z) = -2z.$$

Les trois dérivées partielles s'annulent simultanément si, et seulement si, $x = y = z = 0$. Ce point appartient à la surface d'équation $x^2 + y^2 + z^2 = 0$. Par conséquent, le seul point singulier de cette surface est le point $(0, 0, 0)$.

Définition 19. La surface (F, Σ) est *lisse* si elle est de classe $\mathcal{C}^p$ et si tous les points $M \in \Sigma$ sont des points réguliers.

Exemple. La surface d'équation $x^2 + y^2 + z^2 - 1 = 0$ est une surface lisse. Son support Σ est la sphère de centre $(0, 0, 0)$ et de rayon 1. En effet, en tout point $M := (x, y, z) \in \Sigma$, on a :

$$D_M F \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2x & 2y & 2z \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = 2x^2 + 2y^2 + 2z^2 = 2 \neq 0.$$

La forme $D_M F$ est donc non nulle.

2.2 Sous-variétés lisses

Une surface lisse est un cas particulier de sous-variété lisse de dimension 2 de $\mathbb{R}^3$. La notion de sous-variété lisse a été introduite dans le module « Fonctions de plusieurs variables » du livre [L2]. Nous reformulons la définition dans le cas des sous-variétés lisses de $\mathbb{R}^3$ de dimension 2.

Définition 20. Un ensemble $\Sigma \subset \mathbb{R}^3$ est une *sous-variété lisse* de $\mathbb{R}^3$ de dimension 2 si pour tout $M \in \Sigma$, il existe un ouvert $U \subset \mathbb{R}^3$ contenant M et une application $F : U \to \mathbb{R}$ de classe $\mathcal{C}^1$ telle que

- la forme linéaire $D_M F$ soit non nulle et
- $\Sigma \cap U$ soit l'ensemble des solutions de l'équation $F(M) = 0$.

Exercice 11.
Montrer que le support d'une nappe cartésienne de classe $\mathcal{C}^1$ est une sous-variété lisse de $\mathbb{R}^3$.

Solution. Si la nappe cartésienne est donnée par $\boldsymbol{g}(x, y) = \big(x, y, G(x, y)\big)$ avec V un ouvert de $\mathbb{R}^2$ et $G : V \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$, alors $U := V \times \mathbb{R}$ est un ouvert de $\mathbb{R}^3$. La fonction $F : U \to \mathbb{R}$ définie par $F(x, y, z) = G(x, y) - z$ est de classe $\mathcal{C}^1$. Le support de la nappe cartésienne est l'ensemble Σ des points $(x, y, z) \in U$ tels que

$F(x,y,z)=0$. La différentielle de F en un point de $M := (x,y,z) \in U$ vérifie :

$$D_{(x,y,z)}G \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \frac{\partial G}{\partial x}(x,y)\cdot 0 + \frac{\partial G}{\partial y}(x,y)\cdot 0 - 1\cdot 1 = -1 \neq 0.$$

La différentielle $D_M F$ n'est donc jamais nulle.

Nous avons vu (proposition 9 de la page 798 du livre [L2]) qu'une sous-variété lisse de $\mathbb{R}^3$ de dimension 2 est localement le graphe d'une application exprimant une des variables en fonction des deux autres. Nous reformulons la proposition de la manière suivante.

Proposition 6. Soit $\Sigma \subset \mathbb{R}^3$ une sous-variété lisse de dimension 2. Pour tout point régulier $M \in \Sigma$, il existe un voisinage U de M dans $\mathbb{R}^3$ et une base $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$ de l'espace vectoriel $\mathbb{R}^3$ tels que $\Sigma \cap U$ soit le support d'une nappe cartésienne de classe $\mathcal{C}^1$:

$$\boldsymbol{f} : (x_1, x_2) \mapsto x_1 \cdot \vec{e}_1 + x_2 \cdot \vec{e}_2 + F(x_1, x_2) \cdot \vec{e}_3.$$

En particulier, au voisinage de chacun de ses points, une sous-variété lisse de $\mathbb{R}^3$ de dimension 2 coïncide avec le support d'une nappe paramétrée simple et régulière.

2.3 Espace et plan tangent

Dans le module II.5 du livre [L2], nous avons alors observé (proposition 10 de la page 801) que si, au voisinage d'un point M_0, Σ est à la fois donnée par une équation $F(M) = 0$ avec $D_{M_0}F$ non nulle, et comme le support d'une nappe régulière $\boldsymbol{f}$ avec $\boldsymbol{f}(\boldsymbol{u}_0) = M_0$, alors il y a égalité des espaces vectoriels $\operatorname{Ker} D_{M_0}F = \operatorname{Im} D_{\boldsymbol{u}_0}\boldsymbol{f}$. Il est donc naturel de donner la définition suivante.

Définition 21. Si $\Sigma \subset \mathbb{R}^3$ est une sous-variété lisse de dimension 2 donnée au voisinage d'un point $M_0 \in \Sigma$ par l'équation $F(M) = 0$ avec $D_{M_0}F$ non nulle, alors :

- L'*espace tangent* à Σ au point M_0, noté $T_{M_0}\Sigma$, est le plan vectoriel $\operatorname{Ker} D_{M_0}F$.
- Le *plan tangent* à Σ au point M_0 est le plan affine $M_0 + T_{M_0}\Sigma$.

Il est parfois utile d'introduire la notion de *normale* en un point d'une sous-variété lisse de $\mathbb{R}^3$ de dimension 2. Pour cela, on suppose que l'on travaille dans un espace euclidien muni d'un repère orthonormé, ou encore que l'espace vectoriel $\mathbb{R}^3$ est muni du produit scalaire usuel.

Si $F : U \subset \mathbb{R}^3 \to \mathbb{R}$ est une application de classe $\mathcal{C}^1$, si $M \in U$ et si $\boldsymbol{h} \in \mathbb{R}^3$, on a alors :

$$D_M F(\boldsymbol{h}) = \left(\overrightarrow{\operatorname{grad}}\, F(M) \mid \boldsymbol{h} \right) \quad \text{avec} \quad \overrightarrow{\operatorname{grad}}\, F(M) := \begin{pmatrix} \dfrac{\partial F}{\partial x}(M) \\ \dfrac{\partial F}{\partial y}(M) \\ \dfrac{\partial F}{\partial z}(M) \end{pmatrix}.$$

Définition 22. Si Σ est une sous-variété lisse de $\mathbb{R}^3$ de dimension 2 et si $M \in \Sigma$, la droite $M + \mathrm{Vect}\big(\overrightarrow{\mathrm{grad}}\, F(M)\big)$ s'appelle la *normale* à Σ en M.

Étant donné que la forme linéaire $D_M F$ s'annule précisément sur l'espace tangent à Σ en M, on voit que pour tout point $M \in \Sigma$:

$$T_M\Sigma = \big(\overrightarrow{\mathrm{grad}}\, F(M)\big)^{\perp}.$$

En particulier, la normale à Σ en M est la droite passant par M et perpendiculaire au plan tangent à Σ en M.

Exercice 12.
Déterminer une équation du plan tangent à la surface d'équation $z^3 - xy = 0$ au point $M_0 := (1,1,1)$.
Solution. Soit $F : \mathbb{R}^3 \to \mathbb{R}$ définie par $F(x,y,z) := z^3 - xy$. Les dérivées partielles premières de F sont :

$$\frac{\partial F}{\partial x}(x,y,z) = -y, \quad \frac{\partial F}{\partial y}(x,y,z) = -x \quad \text{et} \quad \frac{\partial F}{\partial z}(x,y,z) = 3z^2.$$

Elles s'annulent simultanément si, et seulement si, $x = y = z = 0$. La surface a donc un point singulier en $(0,0,0)$. Les autres points sont des points réguliers.
Soit $M_0 := (x_0, y_0, z_0) \neq (0,0,0)$ un point de la surface.
Alors, le point $M := (x,y,z)$ appartient au plan tangent à la surface en M_0 si, et seulement si, $\big(\overrightarrow{\mathrm{grad}}(F)(M_0) \mid \overrightarrow{M_0M}\big) = 0$, donc si, et seulement si :

$$\left(\begin{pmatrix} -y_0 \\ -x_0 \\ 3z_0^2 \end{pmatrix} \middle| \begin{pmatrix} x - x_0 \\ y - y_0 \\ z - z_0 \end{pmatrix} \right) = 0.$$

Une équation du plan tangent à la surface en M_0 est donc :

$$-y_0(x - x_0) - x_0(y - y_0) + 3z_0^2(z - z_0) = 0.$$

Dans le cas où $M_0 := (1,1,1)$, l'équation devient $-(x-1) - (y-1) + 3(z-1) = 0$, c'est-à-dire : $x + y - 3z + 1 = 0$.

En fait, on vient de voir qu'une équation du plan tangent à la surface d'équation $F(x,y,z) = 0$ en point régulier $M_0 := (x_0, y_0, z_0)$ est :

$$\frac{\partial F}{\partial x}(M_0) \cdot (x - x_0) + \frac{\partial F}{\partial y}(M_0) \cdot (y - y_0) + \frac{\partial F}{\partial z}(M_0) \cdot (z - z_0) = 0.$$

2.4 Intersection de deux surfaces

Pour compléter cette partie, nous allons étudier l'allure de l'intersection de deux sous-variétés lisses Σ_1 et Σ_2 de $\mathbb{R}^3$ de dimension 2, au voisinage d'un point commun $M_0 \in \Sigma_1 \cap \Sigma_2$.

Définition 23. Soient Σ_1 et Σ_2 deux sous-variétés lisses de $\mathbb{R}^3$ de dimension 2 et soit $M_0 \in \Sigma_1 \cap \Sigma_2$.

- Si $T_{M_0}\Sigma_1 = T_{M_0}\Sigma_2$, on dit Σ_1 et Σ_2 sont *tangentes* en M_0.
- Si $T_{M_0}\Sigma_1 \neq T_{M_0}\Sigma_2$, on dit que Σ_1 et Σ_2 sont *transverses* en M_0.

Proposition 7. Soient Σ_1 et Σ_2 deux sous-variétés lisses de $\mathbb{R}^3$ de dimension 2, d'équations respectives $F_1(M) = 0$ et $F_2(M) = 0$ au voisinage d'un point $M_0 \in \Sigma_1 \cap \Sigma_2$. Alors, Σ_1 et Σ_2 sont tangentes en M_0 si, et seulement si, on a :

$$\overrightarrow{\operatorname{grad}}\, F_1(M_0) \wedge \overrightarrow{\operatorname{grad}}\, F_2(M_0) = 0.$$

Démonstration.

Notons que les plans vectoriels $T_{M_0}\Sigma_1$ et $T_{M_0}\Sigma_2$ sont confondus si, et seulement si, les vecteurs $\overrightarrow{\operatorname{grad}}\, F_1(M_0) \in (T_{M_0}\Sigma_1)^\perp$ et $\overrightarrow{\operatorname{grad}}\, F_2(M_0) \in (T_{M_0}\Sigma_2)^\perp$ sont colinéaires. ■

Exercice 13.

Montrer que les deux surfaces d'équations respectives :

$$x^2 + y^2 + z^2 = 6 \quad \text{et} \quad x^3 + y^3 + z^3 = 8$$

sont transverses en $(1, 2, -1)$.

Solution. Le point $M := (1, 2, -1)$ appartient aux deux surfaces. De plus :

$$\overrightarrow{\operatorname{grad}}\, F_1(M_0) = \begin{pmatrix} 2 \cdot 1 \\ 2 \cdot 2 \\ 2 \cdot (-1) \end{pmatrix} \quad \text{et} \quad \overrightarrow{\operatorname{grad}}\, F_2(M_0) = \begin{pmatrix} 3 \cdot 1^2 \\ 3 \cdot 2^2 \\ 3 \cdot (-1)^2 \end{pmatrix}.$$

Or les vecteurs :

$$\begin{pmatrix} 2 \\ 4 \\ -2 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 3 \\ 12 \\ 3 \end{pmatrix}$$

ne sont pas colinéaires. Les surfaces ne sont donc pas tangentes en M_0.

La situation la plus simple à étudier est celle où Σ_1 et Σ_2 sont transverses en M_0. Dans ce cas, au voisinage de M_0, $\Sigma_1 \cap \Sigma_2$ est une sous-variété lisse de $\mathbb{R}^3$ de dimension 1, ou de manière équivalente, le support d'une courbe paramétrée régulière de classe $\mathcal{C}^1$.

Proposition 8. Soient Σ_1 et Σ_2 deux sous-variétés lisses de $\mathbb{R}^3$ de dimension 2. Supposons que Σ_1 et Σ_2 soient transverses en un point $M_0 \in \Sigma_1 \cap \Sigma_2$. Alors, il existe un voisinage $V \subset \mathbb{R}^3$ de M_0 et une courbe paramétrée régulière et simple $\gamma : I \to \mathbb{R}^3$ de classe $\mathcal{C}^1$ telle que :

- $V \cap \Sigma_1 \cap \Sigma_2$ soit le support de γ et
- $M_0 = \gamma(t_0)$ avec $t_0 \in I$ et $T_{M_0}\Sigma_1 \cap T_{M_0}\Sigma_2 = \mathrm{Vect}\big(\vec{\gamma}'(t_0)\big)$.

En particulier, la tangente à γ au temps t_0 coïncide avec l'intersection des plans tangents à Σ_1 et Σ_2 en M_0.

Démonstration. On choisit un voisinage U de M_0 tel que $\Sigma_1 \cap U$ et $\Sigma_2 \cap U$ soient données par les équations $F_1(M) = 0$ et $F_2(M) = 0$ avec $F_1 : U \to \mathbb{R}$ et $F_2 : U \to \mathbb{R}$ des fonctions de classe $\mathcal{C}^1$ dont les différentielles en M_0 sont non nulles. L'application $\boldsymbol{F} := (F_1, F_2) : U \to \mathbb{R}^2$ est de classe $\mathcal{C}^1$ et sa différentielle en M_0 est donnée par :

$$D_{M_0}\boldsymbol{F}(\boldsymbol{h}) = \big(D_{M_0}F_1(\boldsymbol{h}), D_{M_0}F_2(\boldsymbol{h})\big).$$

On a donc :

$$\mathrm{Ker}\, D_{M_0}\boldsymbol{F} = \mathrm{Ker}\, D_{M_0}F_1 \cap \mathrm{Ker}\, D_{M_0}F_2 = T_{M_0}\Sigma_1 \cap T_{M_0}\Sigma_2.$$

Étant donné que Σ_1 et Σ_2 sont transverses en M_0, le noyau de $D_{M_0}\boldsymbol{F}$ est de dimension 1 et le rang de $D_{M_0}\boldsymbol{F}$ est donc égal à 2. Quitte à permuter les variables, on peut supposer que la matrice suivante est inversible :

$$\begin{pmatrix} \dfrac{\partial F_1}{\partial x} & \dfrac{\partial F_1}{\partial y} \\ \dfrac{\partial F_2}{\partial x} & \dfrac{\partial F_2}{\partial y} \end{pmatrix}.$$

D'après le théorème des fonctions implicites, il existe un voisinage $V \subset \mathbb{R}^3$ contenant M_0 et un intervalle ouvert $I \subset \mathbb{R}$ tel que :

$$(x, y, z) \in U \text{ et } \boldsymbol{F}(x, y, z) = (0, 0) \quad \Longleftrightarrow \quad z \in I \text{ et } (x, y) = \boldsymbol{f}(z)$$

avec $\boldsymbol{f} : I \to \mathbb{R}^2$ de classe $\mathcal{C}^1$. On note (f_1, f_2) les coordonnées de $\boldsymbol{f}$, $f_3 := \mathrm{Id}$ et l'on définit $\gamma : I \to \mathbb{R}^3$ par :

$$\gamma := (f_1, f_2, f_3).$$

Alors, $M \in V \cap \Sigma_1 \cap \Sigma_2$ si, et seulement si, $M = \gamma(t)$ avec $t \in I$. Si $M_0 := (x_0, y_0, z_0)$, alors, en posant $t_0 := z_0$, on a $\gamma(t_0) = M_0$. Comme $\boldsymbol{F} \circ \gamma$ est identiquement nulle, $\vec{\gamma}'(t_0)$ appartient au noyau $D_{M_0}\boldsymbol{F}$, donc à $T_{M_0}\Sigma_1 \cap T_{M_0}\Sigma_2$. Une des coordonnées de $\vec{\gamma}'(t)$ est toujours égale à 1, ce qui montre que $\vec{\gamma}'(t_0) \neq 0$ et que γ est une courbe paramétrée simple et régulière. Finalement $T_{M_0}\Sigma_1 \cap T_{M_0}\Sigma_2$ est de dimension 1 et donc, $T_{M_0}\Sigma_1 \cap T_{M_0}\Sigma_2 = \mathrm{Vect}\big(\vec{\gamma}'(t_0)\big)$. ∎

L'intersection de deux sous-variétés lisses de $\mathbb{R}^3$ de dimension 2 qui sont tangentes en un point M_0 est plus délicate à étudier. Nous en avons eu un aperçu dans la partie précédente où le lemme de Morse nous a permis d'étudier l'intersection d'une nappe cartésienne avec son plan tangent.

Si Σ_1 et Σ_2 sont des sous-variétés tangentes en M_0, on peut localement les considérer comme les supports de deux nappes cartésiennes définies sur un voisinage de M_0 dans leur plan tangent commun. On se ramène donc au cas où Σ_1 et Σ_2 sont les supports de deux nappes cartésiennes $\boldsymbol{g}_1 : V \to \mathbb{R}^3$ et $\boldsymbol{g}_2 : V \to \mathbb{R}^3$, où V est un voisinage de $\mathbb{R}^2$ contenant $(0,0)$ et où :

$$\boldsymbol{g}_1(x,y) = \big(x, y, G_1(x,y)\big) \quad \text{et} \quad \boldsymbol{g}_2(x,y) = \big(x, y, G_2(x,y)\big)$$

avec :

$$G_1(x,y) = \Phi_1(x,y) + \mathrm{o}\big(\|(x,y)\|^2\big) \quad \text{et} \quad G_2(x,y) = \Phi_2(x,y) + \mathrm{o}\big(\|(x,y)\|^2\big)$$

pour des formes quadratiques Φ_1 et Φ_2 sur $\mathbb{R}^2$. On a alors le résultat suivant :

- Premier cas : $\Phi_2 - \Phi_1$ est une forme quadratique non dégénérée de signature $(2,0)$. Alors, au voisinage de M_0, $\Sigma_1 \cap \Sigma_2 = \{0\}$ et Σ_2 se trouve au dessus de Σ_1.
- Deuxième cas : $\Phi_2 - \Phi_1$ est une forme quadratique non dégénérée de signature $(0,2)$. Alors, au voisinage de M_0, $\Sigma_1 \cap \Sigma_2 = \{0\}$ et Σ_2 se trouve au dessous de Σ_1.
- Troisième cas : $\Phi_2 - \Phi_1$ est une forme quadratique non dégénérée de signature $(1,1)$. Alors, au voisinage de M_0, $\Sigma_1 \cap \Sigma_2$ est la réunion de deux courbes régulières qui se coupent en M_0 avec des tangentes distinctes.
- Quatrième cas : $\Phi_2 - \Phi_1$ est une forme quadratique nulle ou dégénérée de signature $(0,1)$ ou $(1,0)$. On ne peut *a priori* rien conclure.

Exercice 14.

La courbe de Viviani est obtenue en coupant une sphère de rayon 1 par un cylindre de révolution de diamètre 1 dont une génératrice passe par le centre de la sphère :

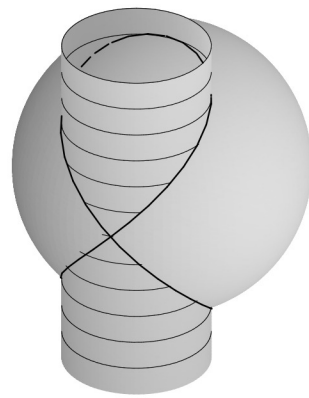
$$\begin{cases} x^2 + y^2 + z^2 - 1 = 0 \\ x^2 + y^2 - x = 0. \end{cases}$$

Étudier l'allure de la courbe au voisinage de $(1,0,0)$.

Solution. Si l'on pose $F_1(x,y,z) := x^2 + y^2 + z^2 - 1$ et $F_2(x,y,z) := x^2 + y^2 - x$, alors $\overrightarrow{\operatorname{grad}} F_1(1,0,0) = (2,0,0)$ et $\overrightarrow{\operatorname{grad}} F_2(1,0,0) = (1,0,0)$. Ces vecteurs sont colinéaires. On en déduit que la sphère et le cylindre sont tangents au point $(1,0,0)$, le plan tangent étant le plan d'équation $x = 1$.

Un point M de coordonnées (x,y,z) appartient à la courbe de Viviani si, et seulement si :

$$\begin{cases} x = 1 - z^2 \\ x = (1-z^2)^2 + y^2 \end{cases}$$



donc si, et seulement si :

$$\begin{cases} x = 1 - z^2 \\ x = 1 + y^2 - 2z^2 + z^4. \end{cases}$$

Il suffit donc d'étudier l'intersection des surfaces d'équations $x = 1 - z^2$ et $x = 1 + y^2 - 2z^2 + z^4$. On a :

$$(1-z^2) - (1 + y^2 - 2z^2 + z^4) = z^2 - y^2 + \mathrm{o}\big(\|(y,z)\|^2\big).$$

La signature de la forme quadratique $\Phi(y,z) = z^2 - y^2$ est $(1,1)$. Cela montre que la courbe de Viviani a un point double en $(1,0,0)$.

---

3 Exemples

3.1 Nappes réglées

Définition 24. Une *nappe réglée* est une nappe paramétrée $\boldsymbol{f} : I \times \mathbb{R} \to \mathbb{R}^3$ de la forme :

$$\boldsymbol{f}(s,t) := \gamma(s) + t \cdot \vec{v}(s)$$

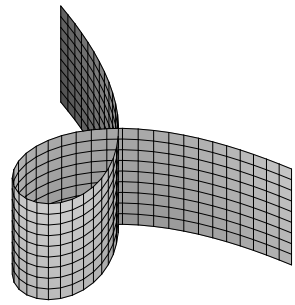
avec $I \subset \mathbb{R}$ un intervalle ouvert d'intérieur non vide, $\gamma : I \to \mathbb{R}^3$ une courbe paramétrée et $\vec{v} : I \to \mathbb{R}^3 \setminus \{\vec{0}\}$ une application continue.

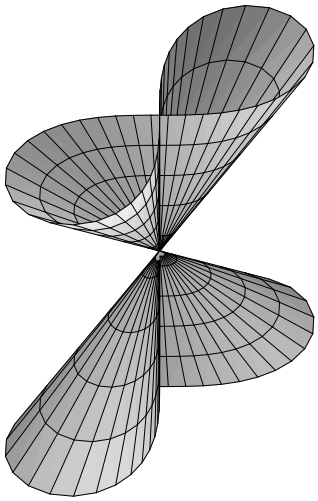
Si l'on note D_s la droite passant par le point $\gamma(s)$ et dirigée par le vecteur $\vec{v}(s)$, alors le support de la nappe réglée définie par $\boldsymbol{f}(s,t) = \gamma(s) + t \cdot \vec{v}(s)$ est la réunion des droites D_s. Ces droites s'appellent des *génératrices* du support de la nappe.
La courbe paramétrée $\gamma : I \to \mathbb{R}^3$ s'appelle une *directrice* de la nappe réglée.

Exemples.

1. Nappe cylindrique. Lorsque l'application $\vec{v}$ est constante, toutes les génératrices sont parallèles. Dans ce cas, on dit que la nappe est une *nappe cylindrique*. C'est par exemple le cas pour la nappe réglée :

$$\boldsymbol{f} : (s,t) \in \,]-2,2[\, \times \mathbb{R} \mapsto \left(\frac{1-s^2}{1+s^2}, \frac{s-s^3}{1+s^2}, t \right).$$

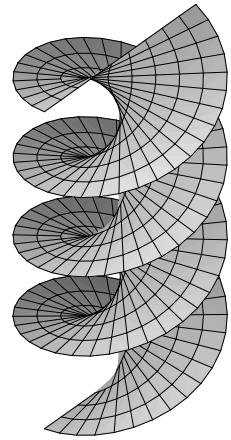




2. Nappe conique. Lorsque la courbe paramétrée γ est constante, $\gamma(s) = M_0 \in \mathbb{R}^3$ pour tout $s \in I$, toutes les génératrices passent par le point M_0. On dit que la nappe est une *nappe conique de sommet* M_0. C'est par exemple le cas pour la nappe réglée :

$$\boldsymbol{f} : (s,t) \in \mathbb{R} \times \mathbb{R} \mapsto \left(t\frac{\sin s}{1+\cos^2 s}, t\frac{\sin s \cos s}{1+\cos^2 s}, t \right).$$

3. Nappe conoïde. Lorsque la courbe paramétrée γ est non constante et prend ses valeurs dans une droite affine D, et lorsque l'application $\vec{v}$ prend ses valeurs dans un plan vectoriel qui ne contient pas la direction de D, on dit que la nappe est *conoïde*. C'est par exemple le cas de la nappe réglée $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}(s,t) = (t\cos s, t\sin s, s)$ qui est une nappe *hélicoïdale*.



Un point $\boldsymbol{m} := \big((s,t), \boldsymbol{f}(s,t)\big)$ est un point régulier d'une nappe réglée si, et seulement si, la différentielle de $\boldsymbol{f}$ en (s,t) est de rang 2. Or :

$$\frac{\partial \boldsymbol{f}}{\partial s}(s,t) = \vec{\gamma}'(s) + t\vec{v}'(s) \quad \text{et} \quad \frac{\partial \boldsymbol{f}}{\partial t}(s,t) = \vec{v}(s).$$

Le point $\boldsymbol{m}$ est donc un point régulier de la nappe si, et seulement si, les vecteurs $\vec{\gamma}'(s) + t\vec{v}'(s)$ et $\vec{v}(s)$ ne sont pas colinéaires. L'espace tangent à la nappe au point $\boldsymbol{m}$ est alors le plan vectoriel engendré par ces deux vecteurs.

Exercice 15.

Montrer que si $\boldsymbol{m} := \big((s,t), \boldsymbol{f}(s,t)\big)$ est un point régulier d'une nappe réglée, la génératrice passant par $\boldsymbol{f}(s,t)$ et dirigée par $\vec{v}(s)$ appartient au plan tangent à la nappe au point $\boldsymbol{m}$.

Solution. Puisque la génératrice est la droite $\boldsymbol{f}(s,t) + \mathrm{Vect}(\vec{v}(s))$ et puisque le plan tangent à la nappe au point $\boldsymbol{m}$ est le plan affine $M + \mathrm{Im}\, D_{(s,t)}\boldsymbol{f}$, il suffit de montrer que $\vec{v}(s)$ appartient à $\mathrm{Im}\, D_{(s,t)}\boldsymbol{f}$. Cela est immédiat étant donné que $D_{(s,t)}\boldsymbol{f} = \mathrm{Vect}\big(\vec{\gamma}'(s) + t\vec{v}'(s), \vec{v}(s)\big)$.

Exercice 16.

Soient P_1 et P_2 deux plans vectoriels de $\mathbb{R}^3$ non confondus, d'équations respectives :

$$a_1x + b_1y + c_1z = 0 \quad \text{et} \quad a_2x + b_2y + c_2z = 0.$$

Soit $F : \mathbb{R}^2 \to \mathbb{R}$ une fonction continue. Montrer que la surface implicite d'équation $F(a_1x + b_1y + c_1z, a_2x + b_2y + c_2z) = 0$ est un *cylindre* au sens où le support est une réunion de droites parallèles de direction $P_1 \cap P_2$.

Solution. Soit $\Gamma \subset \mathbb{R}^2$ l'ensemble des couples (u_1, u_2) tels que $F(u_1, u_2) = 0$. Alors, le point $M := (x,y,z)$ appartient au support Σ de la surface d'équation :

$$F(a_1x + b_1y + c_1z, a_2x + b_2y + c_2z) = 0$$

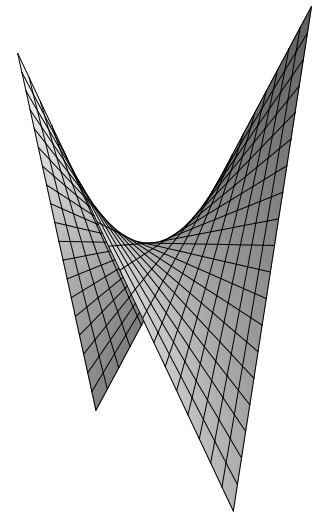
si, et seulement si, il existe $(u_1, u_2) \in \Gamma$ tel que :

$$a_1x + b_1y + c_1z = u_1 \quad \text{et} \quad a_2x + b_2y + c_2z = u_2.$$

Les plans d'équations respectives $a_1x + b_1y + c_1z = u_1$ et $a_2x + b_2y + c_2z = u_2$ ne sont pas parallèles. Ils se coupent donc le long d'une droite $D_{(u_1,u_2)}$. Le support Σ est la réunion des droites $D_{\boldsymbol{u}}$ avec $\boldsymbol{u} \in \Gamma$.

Exercice 17.
Soit F un polynôme homogène de degré k en (x, y, z), c'est-à-dire tel que $F(tx, ty, tz) = t^k F(x, y, z)$ pour tout $t \in \mathbb{R}$. Montrer que la surface implicite d'équation $F(x, y, z) = 0$ est un *cône* de sommet O au sens où le support est une réunion de droites passant par O.
Solution. Soit Σ le support de la surface d'équation $F(x, y, z) = 0$. On a $F(x, y, z) = 0$ si, et seulement si, pour tout $t \in \mathbb{R}$, $F(tx, ty, tz) = 0$. Par conséquent, un point M de coordonnées (x, y, z) appartient à Σ si, et seulement si, la droite (OM) est incluse dans Σ.

Nous avons vu dans l'exercice 8 de la page 141 que l'intersection d'un paraboloïde hyperbolique avec ses plans tangents est la réunion de deux droites sécantes. Par tout point du paraboloïde hyperbolique, il passe donc deux droites sécantes : le paraboloïde hyperbolique est une surface *doublement réglée*.



3.2 Nappes de révolution

Pour introduire la notion de nappe de révolution, on suppose que l'on travaille dans un espace euclidien muni d'un repère orthonormé, ou encore que l'espace vectoriel $\mathbb{R}^3$ est muni du produit scalaire usuel.

Définition 25. Une nappe paramétrée $\boldsymbol{f} : I \times \mathbb{R} \to \mathbb{R}^3$ est une *nappe de révolution* s'il existe un repère orthonormé $(O; \vec{e}_1, \vec{e}_2, \vec{e}_3)$ et des fonctions continues $\rho : I \to \mathbb{R}$ et $\zeta : I \to \mathbb{R}$ avec $I \subset \mathbb{R}$ un intervalle ouvert d'intérieur non vide, telles que :

$$\boldsymbol{f}(t, \alpha) = O + \rho(t) \cos \alpha \cdot \vec{e}_1 + \rho(t) \sin \alpha \cdot \vec{e}_2 + \zeta(t) \cdot \vec{e}_3.$$

Pour chaque $\alpha \in \mathbb{R}$, la courbe paramétrée $\gamma_\alpha : t \in I \mapsto \boldsymbol{f}(t, \alpha)$ s'appelle un *méridien*. Pour chaque $t \in I$, la courbe $\alpha \in \mathbb{R} \mapsto \boldsymbol{f}(t, \alpha)$ s'appelle un *parallèle*. Le support de la nappe de révolution est la réunion des supports Γ_α des méridiens. Notons que Γ_α s'obtient à partir de $\Gamma_{\alpha'}$ par la rotation d'axe $\Delta := O + \mathrm{Vect}(\vec{e}_3)$, d'angle $\alpha' - \alpha$. Le support d'une nappe de révolution est donc invariant par rotation autour d'une droite Δ, appelée *axe de révolution* de la nappe. Le support de chaque parallèle est un cercle contenu dans un plan P perpendiculaire à Δ, centré en $P \cap \Delta$. Un tel cercle s'appelle un cercle d'axe Δ.

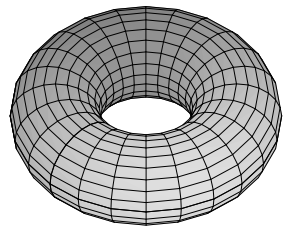
Exemples. Les nappes suivantes sont des nappes de révolution d'axe (Oz). Les trois premières ont déjà été rencontrées dans ce module.

1. La nappe paramétrée $\boldsymbol{f} : (t, \alpha) \mapsto (\cos t \cos \alpha, \cos t \sin \alpha, \sin t)$. Son support est la sphère de centre O et de rayon 1.

2. La nappe cylindrique $\boldsymbol{f} : (t, \alpha) \mapsto (\cos\alpha, \sin\alpha, t)$. Son support est un *cylindre de révolution*
3. La nappe conique $\boldsymbol{f} : (t, \alpha) \mapsto (t\cos\alpha, t\sin\alpha, t)$.
   Son support est un *cône de révolution* de sommet O.
4. La nappe :

$\boldsymbol{f} : (t, \alpha) \mapsto \big((2 + \cos t)\cos\alpha, (2 + \cos t)\sin\alpha, \sin t\big)$.

   Son support est un *tore* (voir figure ci-contre).



Pour $\alpha \in \mathbb{R}$, posons :

$$\vec{u}_\alpha := \cos\alpha \cdot \vec{e}_1 + \sin\alpha \cdot \vec{e}_2 \quad \text{et} \quad \vec{v}_\alpha := \frac{\partial \vec{u}_\alpha}{\partial \alpha} = -\sin\alpha \cdot \vec{e}_1 + \cos\alpha \cdot \vec{e}_2.$$

La famille $(\vec{u}_\alpha, \vec{v}_\alpha, \vec{e}_3)$ est une base orthonormée de $\mathbb{R}^3$.
Observons que :

$$\boldsymbol{f}(t, \alpha) = O + \rho(t) \cdot \vec{u}_\alpha + \zeta(t) \cdot \vec{e}_3.$$

Par conséquent :

$$\frac{\partial \boldsymbol{f}}{\partial t}(t, \alpha) = \rho'(t) \cdot \vec{u}_\alpha + \zeta'(t) \cdot \vec{e}_3 = \vec{\gamma}'_\alpha(t).$$

C'est un vecteur tangent à γ_α au temps t. De plus :

$$\frac{\partial \boldsymbol{f}}{\partial \alpha}(t, \alpha) = \rho(t) \cdot \vec{v}_\alpha.$$

Les vecteurs $\dfrac{\partial \boldsymbol{f}}{\partial t}(t, \alpha)$ et $\dfrac{\partial \boldsymbol{f}}{\partial \alpha}(t, \alpha)$ sont orthogonaux. Le point $\boldsymbol{m} := \big((t, \alpha), \boldsymbol{f}(t, \alpha)\big)$ est un point régulier de la nappe si, et seulement si, ces deux vecteurs sont non nuls, donc si, et seulement si, $\rho(t) \neq 0$ et le point t est un point régulier de la courbe γ_α.
Notons également qu'en tout point régulier $\boldsymbol{m} := \big((t, \alpha), \boldsymbol{f}(t, \alpha)\big)$ d'une nappe de révolution d'axe Δ , le plan tangent est perpendiculaire au plan passant par $\boldsymbol{f}(t, \alpha)$ et contenant Δ .

Exercice 18.

Soit Δ la droite passant par (a_1, b_1, c_1) et dirigée par le vecteur (a_2, b_2, c_2). Soit $F : \mathbb{R}^2 \to \mathbb{R}$ une fonction continue. Montrer que la surface implicite d'équation :

$$F\big((x - a_1)^2 + (y - b_1)^2 + (z - c_1)^2, a_2x + b_2y + c_2z\big) = 0$$

est une *surface de révolution* d'axe Δ au sens que son support est une réunion de cercles d'axe Δ et de points de Δ.

Solution. Soit $\Gamma \subset \mathbb{R}^2$ l'ensemble des couples (u_1, u_2) tels que $F(u_1, u_2) = 0$. Alors, le point $M := (x, y, z)$ appartient au support Σ de la surface d'équation :

$$F\big((x - a_1)^2 + (y - b_1)^2 + (z - c_1)^2, a_2x + b_2y + c_2z\big) = 0$$

si, et seulement si, il existe $(u_1, u_2) \in \Gamma$ tel que :

$$(x - a_1)^2 + (y - b_1)^2 + (z - c_1)^2 = u_1 \quad \text{et} \quad a_2x + b_2y + c_2z = u_2.$$

Le plan d'équation $a_2x + b_2y + c_2z = u_2$ est perpendiculaire à la droite Δ.

La surface d'équation $(x-a_1)^2+(y-b_1)^2+(z-c_1)^2=u_1$ est soit une sphère centrée en un point de Δ, soit un point de Δ. L'ensemble des points (x,y,z) tels que :

$$(x-a_1)^2+(y-b_1)^2+(z-c_1)^2=u_1 \quad \text{et} \quad a_2x+b_2y+c_2z=u_2$$

est donc soit un cercle d'axe Δ, soit un singleton ne contenant qu'un point de Δ. Le support Σ est la réunion de ces ensembles quand (u_1,u_2) parcourt Γ.

Exercice 19.

Soient Δ et D deux droites non coplanaires et non perpendiculaires. Étant donné $\alpha\in\mathbb{R}$, on note D_α la droite obtenue à partir de D par rotation d'angle α autour de Δ. Montrer que la réunion des droites D_α est un hyperboloïde à une nappe.

Solution. Quitte à changer de repère, on peut supposer que Δ est l'axe (Oz) et que la perpendiculaire commune à Δ et D coupe Δ en O et D en $(a,0,0)$ avec $a>0$. La droite D est alors dirigée par un vecteur $\vec{v}:=(0,\cos\varphi,\sin\varphi)$ avec $\varphi\neq 0 \mod \pi/2$. Une équation paramétrique de la droite D est alors de la forme :

$$\begin{pmatrix} a\\ 0\\ 0\end{pmatrix}+t\begin{pmatrix} 0\\ \cos\varphi\\ \sin\varphi\end{pmatrix} \quad \text{avec} \quad t\in\mathbb{R}.$$

La droite D_α est la droite d'équation paramétrique :

$$\begin{pmatrix} \cos\alpha & -\sin\alpha & 0\\ \sin\alpha & \cos\alpha & 0\\ 0 & 0 & 1\end{pmatrix}\cdot\begin{pmatrix} a\\ t\cos\varphi\\ t\sin\varphi\end{pmatrix}, \quad t\in\mathbb{R}.$$

La réunion des droites D_α est donc le support de la nappe paramétrée :

$$(t,\alpha)\in\mathbb{R}^2\mapsto\begin{pmatrix} a\cos\alpha-t\cos\varphi\sin\alpha\\ a\sin\alpha+t\cos\varphi\cos\alpha\\ t\sin\varphi\end{pmatrix}.$$

D'une part, si l'on pose $X:=a\cos\alpha-t\cos\varphi\sin\alpha$, $Y:=a\sin\alpha+t\cos\varphi\cos\alpha$ et $Z:=t\sin\varphi$, on a :

$$\begin{aligned}\frac{X^2}{a^2}+\frac{Y^2}{a^2}-\frac{Z^2}{a^2\tan^2\varphi}= &\quad \cos^2\alpha-\frac{2t\cos\varphi\cos\alpha\sin\alpha}{a}+\frac{t^2\cos^2\varphi\sin^2\alpha}{a^2}\\ &+\sin^2\alpha+\frac{2t\cos\varphi\cos\alpha\sin\alpha}{a}+\frac{t^2\cos^2\varphi\cos^2\alpha}{a^2}\\ &-\frac{t^2\sin^2\varphi}{a^2\tan^2\varphi}=1.\end{aligned}$$

Le support de la nappe est donc contenu dans l'hyperboloïde à une nappe d'équation :

$$\frac{X^2}{a^2}+\frac{Y^2}{a^2}-\frac{Z^2}{a^2\tan^2\varphi}=1.$$

D'autre part, l'intersection de l'hyperboloïde à une nappe avec un plan $Z=\text{cste}$ perpendiculaire à l'axe (Oz) est un cercle. Comme un tel plan coupe la droite D, les points du cercle appartiennent à l'une des droites D_α. L'hyperboloïde est donc égal à la réunion des droites D_α.

3.3 Quadriques

On suppose de nouveau que $\mathbb{R}^3$ est muni du produit scalaire usuel.

Définition 26. Une surface (F, Σ) de $\mathbb{R}^3$ est une *quadrique* si F est de la forme :

$$F(x, y, z) := \Phi(x, y, z) + L(x, y, z) + b$$

avec Φ une forme quadratique non nulle sur $\mathbb{R}^3$, L une forme linéaire sur $\mathbb{R}^3$ et b un réel.

Remarque. En d'autres termes, F est un polynôme de degré 2 en les trois variables x, y et z.

Nous souhaitons déterminer l'allure du support Σ. Nous nous autoriserons pour cela à changer de repère, et à multiplier F par des réels non nuls. La surface (F, Σ) ne sera plus la même, mais le support reste inchangé.

Notons φ la forme polaire de Φ et observons que si $(O; \vec{\imath}, \vec{\jmath}, \vec{k})$ et $(O'; \vec{\imath}', \vec{\jmath}', \vec{k}')$ sont deux repères de $\mathbb{R}^3$, alors

$$\begin{aligned} F(M) &= \Phi(\overrightarrow{OM}) + L(\overrightarrow{OM}) + b \\ &= \Phi(\overrightarrow{O'M}) + 2\varphi(\overrightarrow{OO'}, \overrightarrow{O'M}) + \Phi(\overrightarrow{OO'}) + L(\overrightarrow{O'M}) + L(\overrightarrow{OO'}) + b \\ &= \Phi(\overrightarrow{O'M}) + L'(\overrightarrow{O'M}) + b' \end{aligned}$$

avec

$$L'(\overrightarrow{O'M}) = 2\varphi(\overrightarrow{OO'}, \overrightarrow{O'M}) + L(\overrightarrow{O'M}) \quad \text{et} \quad b' = b + \Phi(\overrightarrow{OO'}) + L(\overrightarrow{OO'}).$$

La forme de l'expression ne dépend donc pas du choix de repère. La forme quadratique Φ ne dépend pas non plus du choix du centre du repère.

Observons que d'après les résultats du module « Espaces préhilbertiens » du livre [L2], il existe un endomorphisme u de $\mathbb{R}^3$ qui est symétrique et vérifie :

$$\varphi(\vec{v}, \vec{w}) = \big(u(\vec{v}) \mid \vec{w} \big).$$

De plus, d'après le théorème spectral de ce même module, il existe une base orthonormée $\mathcal{B} := (\vec{e}_1, \vec{e}_2, \vec{e}_3)$ formée de vecteurs propres de u. Si l'on note λ_1, λ_2 et λ_3 les valeurs propres associées, on a :

$$\Phi(x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3) = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \lambda_3 x_3^2.$$

Dans le repère $(O; \vec{e}_1, \vec{e}_2, \vec{e}_3)$, l'équation de la quadrique prend la forme :

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \lambda_3 x_3^2 + a_1 x_1 + a_2 x_2 + a_3 x_3 + b = 0.$$

Définition 27. On dit que $O' \subset \mathbb{R}^3$ est un *centre* de la quadrique (Σ, F) s'il existe un réel b' tel que :

$$F(M) = \Phi(\overrightarrow{O'M}) + b'.$$

On a alors le résultat suivant.

Proposition 9. Il y a un unique centre O' si, et seulement si, la forme quadratique Φ est non dégénérée. Sinon, soit il n'y a aucun centre, soit il y a une infinité de centres. Dans ce dernier cas, l'ensemble des centres est soit une droite, soit un plan.

Démonstration. Le point O' est donc un centre si, et seulement si, l'application linéaire :

$$L' := 2\varphi(\overrightarrow{OO'}, \cdot) + L$$

est l'application nulle. Notons $Q \in M_3(\mathbb{R})$ la matrice de la forme quadratique Φ dans la base $(\vec{e}_1, \vec{e}_2, \vec{e}_3)$, $A \in M_{1,3}$ la matrice de la forme linéaire L dans cette même base, et $\boldsymbol{x} \in \mathbb{R}^3$ le vecteur des coordonnées de $\overrightarrow{OO'}$. Alors, O' est un centre si, et seulement si, $2{}^t\boldsymbol{x} \cdot Q + A = 0$, donc si, et seulement si :

$${}^tQ \cdot \boldsymbol{x} = -\frac{1}{2}{}^tA.$$

Il s'agit d'un système affine de trois équations à trois inconnues. L'ensemble des solutions est soit vide, soit un sous-espace affine de $\mathbb{R}^3$. Il y a une unique solution si, et seulement si, la matrice Q est inversible, et dans ce cas, on a :

$$\boldsymbol{x} = -\frac{1}{2}{}^t(AQ^{-1}).$$

La matrice Q est inversible si, et seulement si, la forme quadratique Φ est non dégénérée. ∎

Exercice 20.

Montrer que si l'on a une quadrique à centre d'équation $F(x, y, z) = 0$, alors le centre est l'unique point M qui vérifie :

$$\frac{\partial F}{\partial x}(M) = \frac{\partial F}{\partial y}(M) = \frac{\partial F}{\partial z}(M) = 0.$$

Solution. Soit M un point de $\mathbb{R}^3$. Alors :

$$\begin{aligned} F(M + \vec{v}) &= \Phi(\overrightarrow{OM}) + L(\overrightarrow{OM}) + 2\varphi(\vec{OM}, \vec{v}) + L(\vec{v}) + \mathrm{o}\big(\|\vec{v}\|\big) \\ &= F(M) + 2\varphi(\overrightarrow{OM}, \vec{v}) + L(\vec{v}) + \mathrm{o}\big(\|\vec{v}\|\big). \end{aligned}$$

On a donc $2\varphi(\vec{OM}, \vec{v}) + L(\vec{v}) = D_M F(\vec{v}) = \big(\overrightarrow{\mathrm{grad}}\, F(M) \mid \vec{v}\big)$ avec :

$$\overrightarrow{\mathrm{grad}}\, F(M) = \begin{pmatrix} \dfrac{\partial F}{\partial x}(M) \\ \dfrac{\partial F}{\partial y}(M) \\ \dfrac{\partial F}{\partial z}(M) \end{pmatrix}.$$

Comme signalé dans la démonstration précédente, la différentielle $D_M F = 2\varphi(\overrightarrow{OM}, \cdot) + L$ s'annule si, et seulement si, M est le centre de la quadrique.

3.3.1 Quadriques à centre

Nous commencerons par étudier le cas où Φ est non dégénérée. Dans ce cas, la quadrique a un unique centre O'. Dans le repère $(O'; \vec{e}_1, \vec{e}_2, \vec{e}_3)$, une équation de la quadrique prend la forme :

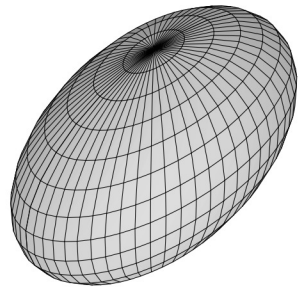
$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \lambda_3 x_3^2 + b' = 0.$$

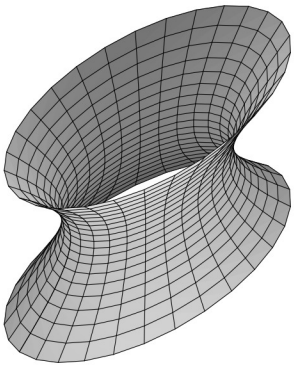
Les trois réels λ_i sont non nuls, et deux d'entre eux ont donc le même signe. Quitte à permuter les coordonnées, on peut donc supposer que l'équation de la quadrique est de la forme :

$$\frac{x_1^2}{\alpha_1^2} + \frac{x_2^2}{\alpha_2^2} + \varepsilon \frac{x_3^2}{\alpha_3^2} + \varepsilon' = 0 \quad \text{avec} \quad \begin{cases} \varepsilon \in \{-1, 1\} \\ \varepsilon' \in \{-1, 0, 1\}. \end{cases}$$

On peut alors étudier l'allure de la quadrique suivant les valeurs de ε et ε'. Notons que dans tous les cas, l'intersection de Σ avec un plan perpendiculaire à l'axe (Oz) est soit vide, soit une ellipse d'équation $\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} = \text{cste}$. Ce sont donc toutes des ellipses homothétiques les unes des autres. Pour déterminer l'allure de Σ, il suffit donc de déterminer l'intersection C de Σ avec le plan d'équation $x_3 = 0$: Σ est la réunion d'ellipses passant par ces points et d'équations $\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} = \text{cste}$.

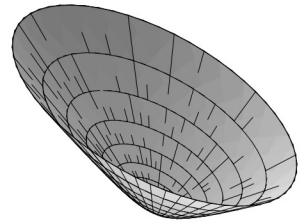
1. Si $\varepsilon' = 0$ et $\varepsilon = +1$, alors Σ ne contient qu'un seul point : O'.
2. Si $\varepsilon' = 0$ et $\varepsilon = -1$, alors le point M appartient à Σ si, et seulement si, $O' + t\overrightarrow{O'M}$ appartient à Σ pour tout $t \in \mathbb{R}$. La courbe C est la réunion de deux droites passant par O'. On dit que Σ est un *cône elliptique* de sommet O'.
3. Si $\varepsilon' = 1$ et $\varepsilon = +1$ alors $\Sigma = \varnothing$.
4. Si $\varepsilon' = -1$ et $\varepsilon = 1$, alors la courbe C est l'ellipse d'équation $\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_3^2}{\alpha_3^2} = 1$. On dit que Σ est un *ellipsoïde*.

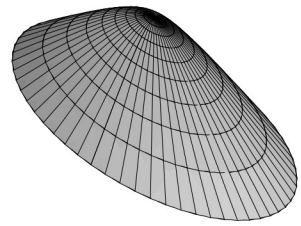




5. Si $\varepsilon' = -1$ et $\varepsilon = -1$, alors la courbe C est l'hyperbole d'axe non transverse $O' + \mathbb{R} \cdot \vec{e}_3$ d'équation $\dfrac{x_1^2}{\alpha_1^2} - \dfrac{x_3^2}{\alpha_3^2} = 1$.
   On dit que Σ est un *hyperboloïde à une nappe*.

6. Si $\varepsilon' = -1$ et $\varepsilon = 1$, alors la courbe C est l'hyperbole d'axe transverse $O' + \mathbb{R} \cdot \vec{e}_3$ d'équation $\dfrac{x_1^2}{\alpha_1^2} - \dfrac{x_3^2}{\alpha_3^2} = -1$. On dit que Σ est un *hyperboloïde à deux nappes*.





En résumé, on a montré que, quitte à faire un changement de repère orthonormé, on peut se ramener à une équation plus simple dite *équation réduite*. On a obtenu la classification suivante.

Équation réduite	$\boldsymbol{\Sigma}$
$\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} + \dfrac{x_3^2}{\alpha_3^2} = 0$	$\{O\}$
$\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} - \dfrac{x_3^2}{\alpha_3^2} = 0$	cône elliptique
$\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} + \dfrac{x_3^2}{\alpha_3^2} + 1 = 0$	$\varnothing$
$\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} + \dfrac{x_3^2}{\alpha_3^2} - 1 = 0$	ellipsoïde
$\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} - \dfrac{x_3^2}{\alpha_3^2} - 1 = 0$	hyperboloïde à une nappe
$\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} - \dfrac{x_3^2}{\alpha_3^2} + 1 = 0$	hyperboloïde à deux nappes

Exemple. La surface d'équation $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} - \dfrac{z^2}{c^2} = 1$ est un *hyperboloïde à une nappe*. C'est le support des nappes réglées :

$$\boldsymbol{f}_1 : (s,t) \in \mathbb{R}^2 \mapsto \gamma(s) + t \cdot \vec{v}_1(s) \quad \text{et} \quad \boldsymbol{f}_2 : (s,t) \in \mathbb{R}^2 \mapsto \gamma(s) + t \cdot \vec{v}_2(s)$$

avec

- $\gamma(s) := (a \cos s, b \sin s, 0)$,
- $\vec{v}_1(s) := (a \sin s, -b \cos s, c)$ et $\vec{v}_2(s) := (a \sin s, -b \cos s, -c)$.

En effet, les supports des nappes sont contenus dans l'hyperboloïde puisque :

$$\frac{(a\cos s + at\sin s)^2}{a^2} + \frac{(b\sin s - bt\cos s)^2}{b^2} - \frac{(\pm ct)^2}{c^2} = 1.$$

De plus, l'intersection de l'hyperboloïde avec le plan d'équation :

$$z = c\tan\varphi \qquad \text{avec} \qquad \varphi \in \left]-\pi/2, \pi/2\right[,$$

est une ellipse d'équation :

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 + \tan^2\varphi \left(= \frac{1}{\cos^2\varphi}\right).$$

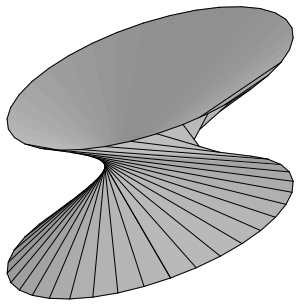
Un point (x_0, y_0, z_0) de cette intersection est de la forme :

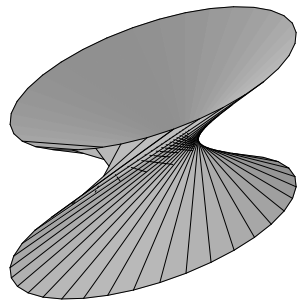
$$x_0 = a\frac{\cos\alpha}{\cos\varphi}, \quad y_0 = b\frac{\sin\alpha}{\cos\varphi} \quad \text{et} \quad z_0 = c\tan\varphi \quad \text{avec} \quad \alpha \in \mathbb{R}.$$

Une bonne utilisation des formules trigonométriques donne alors :

$$(x_0, y_0, z_0) = \boldsymbol{f}_1(\alpha + \varphi, \tan\varphi) = \boldsymbol{f}_2(\alpha - \varphi, -\tan\varphi).$$

On voit donc que par tout point de l'hyperboloïde passent deux droites sécantes. En particulier, l'intersection de l'hyperboloïde avec un plan tangent quelconque contient toujours deux droites sécantes.





Nous allons maintenant montrer que cette intersection est la réunion de ces deux droites. En effet, l'équation du plan tangent en un point $M_0 := (x_0, y_0, z_0)$ est :

$$\frac{2x_0}{a^2}(x - x_0) + \frac{2y_0}{b^2}(y - y_0) - \frac{2z}{c^2}(z - z_0) = 0.$$

Le point (x, y, z) appartient à l'intersection de l'hyperboloïde et du plan tangent si, et seulement si :

$$\begin{cases} \dfrac{2x_0}{a^2}(x - x_0) + \dfrac{2y_0}{b^2}(y - y_0) - \dfrac{2z}{c^2}(z - z_0) = 0 \\ \dfrac{x^2 - x_0^2}{a^2} + \dfrac{y^2 - y_0^2}{b^2} - \dfrac{z^2 - z_0^2}{c^2} = 0. \end{cases}$$

En soustrayant, on obtient :

$$\begin{cases} \dfrac{2x_0}{a^2}(x - x_0) + \dfrac{2y_0}{b^2}(y - y_0) - \dfrac{2z}{c^2}(z - z_0) = 0 \\ \dfrac{(x - x_0)^2}{a^2} + \dfrac{(y - y_0)^2}{b^2} - \dfrac{(z - z_0)^2}{c^2} = 0. \end{cases}$$

Il s'agit de l'intersection d'un plan passant par le point M_0 avec un cône elliptique de sommet M_0. Cette intersection contient au plus (et donc exactement) deux droites sécantes en M_0.

Muni du lemme de Morse dans $\mathbb{R}^3$, l'étude des quadriques à centre nous permet de comprendre l'allure des surfaces de niveau d'une fonction $F : U \subset \mathbb{R}^3 \to \mathbb{R}$ de classe $\mathcal{C}^3$ au voisinage d'un point critique non dégénéré.

Exercice 21.

Soient $U \subset \mathbb{R}^3$ un ouvert et $F : U \to \mathbb{R}$ une fonction de classe $\mathcal{C}^3$. On suppose que $M_0 \in U$ est un point critique non dégénéré de F, i.e. :

$$F(M) = F(M_0) + \Phi(\overrightarrow{M_0M}) + \mathrm{o}\big(\|\overrightarrow{M_0M}\|^2\big)$$

avec Φ une forme quadratique non dégénérée sur $\mathbb{R}^3$. Étudier l'allure des surfaces de niveaux $F = \mathrm{cste}$ au voisinage de M_0.

Solution. La forme quadratique Φ étant non dégénérée, le lemme de Morse dans $\mathbb{R}^3$ affirme qu'il existe un voisinage $V \subset U$ de M_0, un voisinage $W \subset \mathbb{R}^3$ de $\mathbf{0}$ et un difféomorphisme $\psi : V \to W$ de classe $\mathcal{C}^1$ envoyant M_0 sur $\mathbf{0}$ tels que :

$$F_{|V} = \Phi \circ \psi.$$

Au voisinage de M_0, les surfaces de niveaux $F = \mathrm{cste}$ sont donc les images par ψ^{-1} des surfaces de niveaux de Φ au voisinage de $\mathbf{0}$.

- Si la signature de Φ est $(3,0)$, la surface d'équation $F(M) = F(M_0)$ ne contient que le point M_0, les surfaces d'équation $F(M) = \mathrm{cste} < F(M_0)$ sont vides (au voisinage de M_0) et les surfaces d'équation $F(M) = \mathrm{cste} > F(M_0)$ ressemblent à des ellipsoïdes (si la constante est suffisamment proche de $F(M_0)$).
- Si la signature de Φ est $(0,3)$, la surface d'équation $F(M) = F(M_0)$ ne contient que le point M_0, les surfaces d'équation $F(M) = \mathrm{cste} > F(M_0)$ sont vides (au voisinage de M_0) et les surfaces d'équation $F(M) = \mathrm{cste} > F(M_0)$ ressemblent à des ellipsoïdes (si la constante est suffisamment proche de $F(M_0)$).
- Si la signature de Φ est $(2,1)$, la surface d'équation $F(M) = F(M_0)$ ressemble à un cône au voisinage de son sommet. Les surfaces d'équations $F(M) = \mathrm{cste} > F(M_0)$ ressemblent à un morceau d'hyperboloïde à une nappe (si la constante est suffisamment proche de $F(M_0)$) et les surfaces d'équations $F(M) = \mathrm{cste} < F(M_0)$ ressemblent à deux morceaux d'hyperboloïde à deux nappes (si la constante est suffisamment proche de $F(M_0)$).
- Si la signature de Φ est $(1,2)$, la surface d'équation $F(M) = F(M_0)$ ressemble à un cône au voisinage de son sommet. Les surfaces d'équations $F(M) = \mathrm{cste} < F(M_0)$ ressemblent à un morceau d'hyperboloïde à une nappe (si la constante est suffisamment proche de $F(M_0)$) et les surfaces d'équations $F(M) = \mathrm{cste} > F(M_0)$ ressemblent à deux morceaux d'hyperboloïde à deux nappes (si la constante est suffisamment proche de $F(M_0)$).

3.3.2 Cas où Φ est de rang 2

Quitte à permuter les coordonnées, on peut supposer que λ_1 et λ_2 sont non nuls et que $\lambda_3 = 0$. L'équation est alors de la forme :

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + a_1 x_1 + a_2 x_2 + a_3 x_3 + b = 0.$$

On commence par poser :

$$O' := O - \frac{a_1}{2\lambda_1}\vec{e}_1 - \frac{a_2}{2\lambda_2}\vec{e}_2.$$

Dans le repère $(O'; \vec{e}_1, \vec{e}_2, \vec{e}_3)$, l'équation prend alors la forme :

$$\lambda_1 (x_1')^2 + \lambda_2 (x_2')^2 + a_3 x_3 + b' = 0 \quad \text{avec} \quad b' := b + \frac{a_1^2}{4\lambda_1} + \frac{a_2^2}{4\lambda_2}\cdot$$

Si $a_3 = 0$, on peut donc supposer que l'équation de la quadrique est de la forme :

$$\frac{x_1^2}{\alpha_1^2} + \varepsilon \frac{x_2^2}{\alpha_2^2} + \varepsilon' = 0 \quad \text{avec} \quad \begin{cases} \varepsilon \in \{-1, 1\} \\ \varepsilon' \in \{-1, 0, 1\}. \end{cases}$$

Notons que le point M appartient à Σ si, et seulement si, la droite $M + \mathrm{Vect}(\vec{e}_3)$ est contenue dans Σ.

1. Si $\varepsilon = 1$ et $\varepsilon' = 0$, alors $x_1' = x_2' = 0$ et Σ est la droite $O' + \mathrm{Vect}(\vec{e}_3)$.
2. Si $\varepsilon = -1$ et $\varepsilon' = 0$, alors $\alpha_2 x_1 + \alpha_1 x_2 = 0$ ou $\alpha_2 x_1 - \alpha_1 x_2 = 0$ et Σ est la réunion de deux plans sécants.
3. Si $\varepsilon = 1$ et $\varepsilon' = 1$, alors $\Sigma = \varnothing$.
4. Si $\varepsilon = 1$ et $\varepsilon' = -1$, l'intersection de Σ avec les plans d'équations $x_3 = \mathrm{cste}$ est une ellipse. On dit que Σ est un *cylindre elliptique*.
5. Si $\varepsilon = -1$ et $\varepsilon' = \pm 1$, l'intersection de Σ avec les plans d'équations $x_3 = \mathrm{cste}$ est une hyperbole. On dit que Σ est un *cylindre hyperbolique*.

Si $a_3 \neq 0$, on pose :

$$O'' := O' - \frac{b'}{a_3}\vec{e}_3.$$

Dans le repère $(O''; \vec{e}_1, \vec{e}_2, \vec{e}_3)$, l'équation prend alors la forme :

$$\lambda_1 (x_1'')^2 + \lambda_2 (x_2'')^2 + a_3 x_3'' = 0.$$

Quitte à permuter les coordonnées, on peut donc supposer que l'équation de la quadrique est de la forme :

$$\frac{x_1^2}{\alpha_1^2} + \varepsilon \frac{x_2^2}{\alpha_2^2} - x_3 = 0 \quad \text{avec} \quad \varepsilon \in \{-1, 1\}.$$

On voit donc que Σ est le graphe de la forme quadratique $(x_1, x_2) \mapsto \dfrac{x_1^2}{\alpha_1^2} + \varepsilon \dfrac{x_2^2}{\alpha_2^2}\cdot$

1. Si $\varepsilon = +1$, on dit que Σ est un *paraboloïde elliptique*.
2. Si $\varepsilon = -1$, on dit que Σ est un *paraboloïde hyperbolique*.

En résumé, on a montré que, quitte à faire un changement de repère orthonormé, on a les équations réduites suivantes.

Équation réduite	$\boldsymbol{\Sigma}$
$\dfrac{x_1^2}{\alpha_1^2}+\dfrac{x_2^2}{\alpha_2^2}=0$	droite (Oz)
$\dfrac{x_1^2}{\alpha_1^2}-\dfrac{x_2^2}{\alpha_2^2}=0$	réunion de deux plans sécants
$\dfrac{x_1^2}{\alpha_1^2}+\dfrac{x_2^2}{\alpha_2^2}+1=0$	$\varnothing$
$\dfrac{x_1^2}{\alpha_1^2}+\dfrac{x_2^2}{\alpha_2^2}-1=0$	cylindre elliptique
$\dfrac{x_1^2}{\alpha_1^2}-\dfrac{x_2^2}{\alpha_2^2}\pm 1=0$	cylindre hyperbolique
$\dfrac{x_1^2}{\alpha_1^2}+\dfrac{x_2^2}{\alpha_2^2}-x_3=0$	paraboloïde elliptique
$\dfrac{x_1^2}{\alpha_1^2}-\dfrac{x_2^2}{\alpha_2^2}-x_3=0$	paraboloïde hyperbolique

3.3.3 Cas où Φ est de rang 1

Quitte à permuter les coordonnées, on peut supposer que $\lambda_1=\lambda_2=0$ que $\lambda_3\neq 0$. L'équation est alors de la forme :

$$\lambda_3 x_3^2+a_1x_1+a_2x_2+a_3x_3+b=0.$$

On pose :

$$O' := O-\frac{a_3}{2\lambda_3}\vec{e}_3.$$

Dans le repère $(O';\vec{e}_1,\vec{e}_2,\vec{e}_3)$, l'équation prend alors la forme :

$$\lambda_3(x_3')^2+a_1x_1'+a_2x_2'+b'=0 \quad \text{avec} \quad b' := b+\frac{a_3^2}{4\lambda_3}\cdot$$

Si $a_1=a_2=0$, on s'est donc ramené à une équation de la forme :

$$\frac{x_3^2}{\alpha_3^2}+\varepsilon=0 \quad \text{avec} \quad \varepsilon\in\{-1,0,1\}.$$

1. Si $\varepsilon=0$, alors Σ est le plan d'équation $x_3=0$.

2. Si $\varepsilon = 1$, alors $\Sigma = \varnothing$.
3. Si $\varepsilon = -1$, alors Σ est la réunion de plans parallèles d'équations $x_3 = -\alpha_3$ et $x_3 = \alpha_3$.

Si $(a_1, a_2) \neq (0, 0)$, on pose :

$$\vec{e}_1' := \frac{a_1\vec{e}_1 + a_2\vec{e}_2}{\sqrt{a_1^2 + a_2^2}}, \quad \vec{e}_1' := \frac{-a_2\vec{e}_1 + a_1\vec{e}_2}{\sqrt{a_1^2 + a_2^2}} \quad \text{et} \quad \vec{e}_3' := e_3.$$

Dans le repère orthonormé $(O'; \vec{e}_1', \vec{e}_2', \vec{e}_3')$, l'équation prend la forme :

$$\lambda_3(x_3')^2 + \sqrt{a_1^2 + a_2^2}x_1' = 0.$$

Autrement dit, on vient de montrer que l'on peut se ramener à une équation de la forme $x_1 = k \cdot x_3^2$. Un point M appartient à Σ si, et seulement si, la droite $M + \mathrm{Vect}(\vec{e}_2)$ appartient à Σ. L'intersection de Σ avec chaque plan $x_2 = \mathrm{cste}$ est une parabole. On dit que Σ est un *cylindre parabolique*.
En résumé, on a montré que, quitte à faire un changement de repère orthonormé, on a les équations réduites suivantes.

Équation réduite	**Σ**
$x_3^2 = 0$	plan
$x_3^2 + \alpha_3^2 = 0$ avec $\alpha_3 \neq 0$	$\varnothing$
$x_3^2 - \alpha_3^2 = 0$ avec $\alpha_3 \neq 0$	réunion de deux plans parallèles et distincts
$\dfrac{x_3^2}{\alpha_3^2} \pm x_1 = 0$	cylindre parabolique

Exercice type corrigé

II.1.0 Soit $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ la nappe cartésienne définie par :

$$\boldsymbol{f}(x, y) = \big(x, y, x^3 + y^3 - 3xy\big).$$

Déterminer les plans tangents à la nappe et la position de la nappe par rapport à ses plans tangents.
Solution. Il s'agit d'une nappe cartésienne. Donc, la nappe est simple et tous les points de la nappe sont des points réguliers. Le plan tangent à la nappe au point (x_0, y_0, z_0) avec $z_0 := x_0^3 + y_0^3 - 3x_0y_0$ est le plan d'équation paramétrique :

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} + u \begin{pmatrix} 1 \\ 0 \\ 3x_0^2 - 3y_0 \end{pmatrix} + v \begin{pmatrix} 0 \\ 1 \\ 3y_0^2 - 3x_0 \end{pmatrix} \quad \text{avec} \quad (u, v) \in \mathbb{R}^2.$$

Autrement dit :

$$\begin{cases} x - x_0 = u \\ y - y_0 = v \\ z - z_0 = u \cdot (3x_0^2 - 3y_0) + v \cdot (3y_0^2 - 3x_0). \end{cases}$$

Le plan tangent a donc pour équation cartésienne :

$$z - z_0 = (x - x_0) \cdot (3x_0^2 - 3y_0) + (y - y_0) \cdot (3y_0^2 - 3x_0).$$

Si l'on pose $F(x, y) = x^3 + y^3 - 3xy$, l'équation du plan tangent est :

$$z - z_0 = (x - x_0) \cdot \frac{\partial F}{\partial x}(x_0, y_0) + (y - y_0) \cdot \frac{\partial F}{\partial y}(x_0, y_0).$$

Pour étudier la position de la nappe par rapport à ses plans tangents, il suffit d'étudier le signe de :

$$F(x, y) - F(x_0, y_0) - (x - x_0) \cdot \frac{\partial F}{\partial x}(x_0, y_0) - (y - y_0) \cdot \frac{\partial F}{\partial y}(x_0, y_0).$$

Or, au voisinage de (x_0, y_0), on a :

$$\begin{aligned} F(x, y) = F(x_0, y_0) &+ (x - x_0) \cdot \frac{\partial F}{\partial x}(x_0, y_0) + (y - y_0) \cdot \frac{\partial F}{\partial y}(x_0, y_0) \\ &+ \frac{1}{2}\big(r(x - x_0)^2 + 2s(x - x_0)(y - y_0) + t(y - y_0)^2\big) + \mathrm{o}\big(\|(x - x_0, y - y_0)\|^2\big) \end{aligned}$$

avec :

$$r := \frac{\partial^2 F}{\partial x^2}(x_0, y_0) = 6x_0, \quad s := \frac{\partial^2 F}{\partial x \partial y}(x_0, y_0) = -3 \quad \text{et} \quad t := \frac{\partial^2 F}{\partial y^2}(x_0, y_0) = 6y_0.$$

- Si $rt - s^2 = 36x_0y_0 - 9 > 0$ et si $r + t = 6(x_0 + y_0) > 0$, alors au voisinage de (x_0, y_0), la nappe se trouve au dessus de son plan tangent. Cela se produit lorsque $x_0y_0 > 1/4$ avec $x_0 > 0$ (et donc $y_0 > 0$).

- Si $x_0y_0 > 1/4$ avec $x_0 < 0$ (et donc $y_0 < 0$), alors au voisinage de (x_0, y_0), la nappe se trouve au dessus de son plan tangent.
- Si $x_0y_0 < 1/4$, la nappe traverse le plan tangent.

Il ne reste plus qu'à déterminer la position par rapport au plan tangent lorsque $x_0y_0 = 1/4$.
Écrivons :

$$F(x,y) = F(x_0,y_0) + (x-x_0)\cdot\frac{\partial F}{\partial x}(x_0,y_0) + (y-y_0)\cdot\frac{\partial F}{\partial y}(x_0,y_0) + G(x,y)$$

avec :

$$\begin{aligned} G(x,y) &= (x^3+y^3-3xy) - (x_0^3+y_0^3-3x_0y_0) \\ &\quad - (x-x_0)\cdot(3x_0^2-3y_0) - (y-y_0)\cdot(3y_0^2-3x_0). \end{aligned}$$

Posons $u := x - x_0$ et $v := y - y_0$, de sorte que :

$$\begin{aligned} G(x,y) &= x_0^3 + 3x_0^2u + 3x_0u^2 + u^3 + y_0^3 + 3y_0^2v + 3y_0v^2 + v^3 \\ &\quad - 3x_0y_0 - 3x_0v - 3y_0u - 3uv \\ &\quad - x_0^3 - y_0^3 + 3x_0y_0 - 3x_0^2u + 3y_0u - 3y_0^2v + 3x_0v \\ &= u^3 + v^3 + 3(x_0u^2 + y_0v^2 - uv). \end{aligned}$$

Étant donné que $x_0y_0 = 1/4$, on a :

$$x_0u^2 + y_0v^2 - uv = x_0(u - 2y_0v)^2.$$

Le long de la droite d'équation $u = 2y_0v$, on a donc :

$$G(x,y) = u^3 + v^3 = (1 + 8y_0^3)v^3.$$

Tant que $y_0 \neq -1/2$, et donc $x_0 \neq -1/2$, le signe de $G(x,y)$ change donc avec le signe de $y - y_0$, ce qui montre que la nappe traverse son plan tangent en (x_0, y_0).
Enfin, si $x_0 = y_0 = -1/2$, on a :

$$G(x,y) = u^3 + v^3 - \frac{3}{2}(u+v)^2.$$

Le long de la courbe $v = -u - u^3$, on a donc :

$$G(x,y) = u^3 - u^3(1+u^2)^3 - \frac{3}{2}u^6 = -3u^5 + \mathrm{o}(u^5).$$

La fonction $G(x,y)$ change donc de signe au voisinage de $(-1/2, -1/2)$ et la surface traverse son plan tangent en $(-1/2, -1/2)$.

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 la section 1

II.1.1 Déterminer les points multiples de la nappe paramétrée $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}(u,v) = (uv, v^2, v^3)$.

II.1.2 Déterminer le support, les points multiples et les points réguliers de la nappe paramétrée $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$ définie par :

$$\boldsymbol{f}(x,y) := \left(\frac{2x}{1+x^2+y^2}, \frac{2y}{1+x^2+y^2}, \frac{x^2+y^2-1}{1+x^2+y^2}\right).$$

II.1.3 Les nappes suivantes ont-elles le même support ? Représentent-elles les même nappes géométriques ?

1. $\boldsymbol{f}_1 : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}_1(x,y) := \big(x, y, x^2+y^2\big)$ et
2. $\boldsymbol{f}_2 : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}_2(r,\theta) := \big(r\cos\theta, r\sin\theta, r^2\big)$.

II.1.4 La nappe paramétrée $\boldsymbol{f} :]-1,1[\times \mathbb{R} \to \mathbb{R}^3$ définie par :

$$\boldsymbol{f}(t,u) := \big((2+t\cos u)\cos(2u), (2+t\cos u)\sin(2u), t\sin u\big)$$

est-elle orientable ?

II.1.5 Déterminer les points singuliers des nappes suivantes :

1. $\boldsymbol{f}_1 : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}_1(r,t) := (r\cos t, r\sin t, rt)$.
2. $\boldsymbol{f}_2 : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}_2(u,v) = (u^2, v^3, uv)$.
3. La nappe astroïdale $\boldsymbol{f}_3 : \mathbb{R}^2 \to \mathbb{R}^3$ définie par :

$$\boldsymbol{f}_3(u,v) := (\cos^3 u\cos^3 v, \sin^3 u\cos^3 v, \sin^3 v).$$

4. La nappe d'Enneper $\boldsymbol{f}_4 : \mathbb{R}^2 \to \mathbb{R}^3$ définie par :

$$\boldsymbol{f}_4(u,v) := \left(u - \frac{u^3}{3} + uv^2, v - \frac{v^3}{3} + vu^2, u^2 - v^2\right).$$

II.1.6 Donner une équation cartésienne des plans tangents en tout point régulier :

1. $\boldsymbol{f}_1 : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}_1(u,v) := (u+v, u^2+v, v^3)$.
2. $\boldsymbol{f}_2 : \mathbb{R}^2 \to \mathbb{R}^3$ définie par $\boldsymbol{f}_2(x,y) := \big(x, y, (x^2+y^2)(x^2+y^2-8)\big)$.

Exercices sur la section 2

II.1.7 Déterminer les points singuliers des surfaces implicites suivantes :

1. Le berlingot d'équation : $x^2(z+1)^2 + y^2(z-1)^2 - (z^2-1)^2 = 0$.
2. La surface de Cayley d'équation : $2xyz + (x^2+y^2+z^2) - 1 = 0$.
3. La surface de Cassini d'équation : $\big((x+1)^2+y^2\big)\big((x-1)^2+y^2\big) - z^4 = 0$.
4. La surface de Scherk d'équation : $e^z \cos y - \cos x = 0$.

II.1.8 On considère l'application $F : \mathbb{R}^3 \to \mathbb{R}$ définie par :

$$F(x,y,z) = x^2 + y^2 + 2xz + 2yz + 4y.$$

1. Déterminer le gradient de F au point $M_0 := \big((\sqrt{3}+3)/3, (\sqrt{3}-3)/3, (\sqrt{3}-3)/3\big)$.
2. En déduire l'équation du plan tangent à la surface d'équation $F(M) = F(M_0)$ au point M_0.
3. Étudier la position par rapport au plan tangent au voisinage de M_0.

Exercices sur la section 3

II.1.9 À quelle condition sur $a \in \mathbb{R}$, la quadrique $x(a-y) + y(z-a) + z(x-a) - a = 0$ est-elle un cône ?

II.1.10 Montrer que la surface d'équation $x^3 + y^3 + z^3 - 3xyz - 1 = 0$ est de révolution.

II.1.11 Identifier le support de la surface d'équation : $(x^2+y^2+z^2+3)^2 - 16(x^2+y^2) = 0$.

II.1.12 Montrer que la surface d'équation $z(x^2+y^2) - (x^2-y^2) = 0$ est un conoïde.

II.1.13 Soit $\Phi(x,y,z) = Ax^2 + A'y^2 + A''z^2 + 2Bxy + 2B'yz + 2B''zx$ une forme quadratique non dégénérée sur $\mathbb{R}^3$. Soit $L(x,y,z) = 2Dx + 2Ey + 2Fz$ une forme linéaire sur $\mathbb{R}^3$. On souhaite déterminer le type de la quadrique d'équation $\Phi(x,y,z) + L(x,y,z) + G = 0$ le plus simplement possible. Pour cela, on va déterminer le centre $M_0 := (x_0, y_0, z_0)$ de la quadrique et la constante G' telle que l'équation de la quadrique dans le repère $(M_0; \vec{\imath}, \vec{\jmath}, \vec{k})$ soit $\Phi(x,y,z) + G' = 0$.

1. Montrer que :

$$(X,Y,Z,T) \mapsto T^2 \cdot \left(\Phi\left(\frac{X}{T}, \frac{Y}{T}, \frac{Z}{T}\right) + L\left(\frac{X}{T}, \frac{Y}{T}, \frac{Z}{T}\right) + G\right)$$

définit une forme quadratique sur $\mathbb{R}^4$ dont la matrice est :

$$M := \begin{pmatrix} A & B & B'' & D \\ B & A' & B' & E \\ B'' & B' & A'' & F \\ D & E & F & G \end{pmatrix}.$$

2. Montrer que $M \cdot \begin{pmatrix} x_0 \\ y_0 \\ z_0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ G' \end{pmatrix}$.
3. Quel est le type de la quadrique d'équation $2yz + 4zx - 4xy - 2x - 2y + 1 = 0$.

II.1.14 Donner une équation réduite et identifier le type des quadriques suivantes.

1. $3x^2 + 3y^2 + 3z^2 + 2xy - 6z - 1 = 0$.
2. $15x^2 - 2y^2 - 19z^2 + 2\sqrt{6}xy - 2\sqrt{3}xz - 34\sqrt{2}yz - 1 = 0$.

II.1.15 Identifier le support des nappes paramétrées suivantes :

1. $(u, v) \in \mathbb{R} \times]-\pi/2, \pi/2[\mapsto \left(\dfrac{u}{2\cos v}, u\tan v, u^2\right)$.
2. $(\theta, \varphi) \in \mathbb{R}^2 \mapsto (\alpha_1 \sin\theta\cos\varphi, \alpha_2 \sin\theta\sin\varphi, \alpha_3\cos\theta)$.
3. $(\theta, \varphi) \in \mathbb{R}^2 \mapsto (\alpha_1 \cosh\theta\cos\varphi, \alpha_2 \cosh\theta\sin\varphi, \alpha_3\sinh\theta)$.
4. $(\theta, \varphi) \in \mathbb{R}^2 \mapsto (\alpha_1 \sinh\theta\cos\varphi, \alpha_2 \sinh\theta\sin\varphi, \alpha_3\cosh\theta)$.

Analyse

Partie III

Intégration III.1

La première partie de ce module est consacrée à une présentation de la notion *d'intégrale multiple*, c'est-à-dire d'intégrale d'une fonction de plusieurs variables, *dans le cadre de l'intégrale de Riemann*. Nous avons choisi de donner *le minimum de bases théoriques* nécessaires à une présentation *correcte et rigoureuse* des résultats fondamentaux nécessaires pour les applications élémentaires. Ce minimum n'est malheureusement pas si élémentaire... et certains résultats peuvent être admis en première lecture par le lecteur principalement intéressé par les applications. Signalons qu'aujourd'hui les logiciels de calcul formel proposent des calculs d'intégrales multiples mais qu'il ne faut pas être trop optimiste sur leur efficacité et que, pour les utiliser judicieusement, il vaut mieux avoir bien compris ce que l'on calcule...
Cette première partie complète l'étude de l'intégrale de Riemann des fonctions d'une variable faite dans le module *Intégration* du livre [L2]. Le lecteur dispose ainsi d'une théorie de l'intégration largement suffisante pour nombre d'applications.

> ***Remarque importante***. Dans *tous les autres modules de ce livre* nous n'utiliserons *que l'intégrale de Riemann*. Dans la plupart des cas on n'utilise en fait que l'intégration des fonctions *d'une variable*, le module *Intégration* du livre [L2] est alors suffisant.

On peut souhaiter aller plus loin qu'une théorie de l'intégration « pour débutants » et présenter, *au niveau de la Licence* (essentiellement en troisième année), une théorie « moderne » comme celle de Lebesgue, beaucoup plus puissante et d'usage plus souple. Cela pose à ce niveau des problèmes pédagogiques bien connus et assez difficiles à surmonter. La présentation classique (inspirée de celle de Lebesgue lui-même) consiste à construire tout d'abord la mesure de Lebesgue, puis à introduire la notion de fonction mesurable, ce qui permet (enfin...) de définir les fonctions intégrables et l'intégrale. Tout cela est assez délicat et abstrait et remporte, pour le moins, peu de succès auprès des étudiants... Même si l'on admet les résultats les plus difficiles, la mise en place des définitions n'est pas aisée. Divers expédients sont possibles pour « évacuer » les difficultés : on peut, par exemple, « faire comme si toutes les fonctions étaient mesurables ». Cela ne risque pas de causer de difficulté dans la pratique mais n'est pas vraiment satisfaisant pour un ouvrage qui se voudrait « de référence ». Nous avons donc adopté un point de vue tout autre en donnant la priorité à la notion *de fonction intégrable* sur celle *d'ensemble mesurable* (renouant avec

l'approche de N. Bourbaki un peu oubliée aujourd'hui : « chez Bourbaki, les théorèmes relatifs à la mesure des ensembles, qui ne sont évidemment pas oubliés, découlent en général directement des propriétés des intégrales »[1]).

La définition que nous utiliserons (*cf.* la partie 2.1) est due à Henstock et Kurzweil et remonte au milieu du XXe siècle. Elle présente de multiples avantages. La théorie de Henstock-Kurzweil est très proche de celle de Riemann. L'intégrale de Henstock-Kurzweil porte également le nom *d'intégrale de Riemann complète* ou celui *d'intégrale de jauge*. Au lieu de considérer toutes les limites de sommes de Riemann, il s'agit de n'en sélectionner que certaines, judicieusement choisies. La *simplicité* de la définition de l'intégrale au sens de Henstock-Kurzweil nous permet de l'introduire *sans avoir à développer auparavant la théorie de la mesure*. Par ailleurs, la théorie de Henstock-Kurzweil est *plus complète* que celle de Lebesgue : elle permet d'intégrer toutes les fonctions ayant des primitives, toutes les fonctions intégrables au sens de Riemann ou au sens de Riemann généralisé et toutes les fonctions intégrables au sens de Lebesgue. Inversement, il existe des fonctions qui sont intégrables au sens de Henstock-Kurzweil mais qui ne le sont pas au sens de Lebesgue. Nous verrons par exemple que la dérivée d'une fonction dérivable est toujours intégrable au sens de Henstock-Kurzweil, ce qui conduit à un énoncé très général du théorème fondamental de l'analyse qui n'est pas vrai en théorie de Lebesgue. La théorie de Henstock-Kurzweil permet aussi d'intégrer des fonctions « oscillantes » (apparaissant fréquemment en physique) qui ne sont pas intégrables au sens de Lebesgue. Au-delà, les succès de *l'intégrale de Feynmann* en physique théorique (mécanique quantique et théorie quantique des champs) montrent qu'il peut être utile, contrairement aux idées reçues, de « découpler » la notion de mesure de celle d'intégrale et de revenir à l'esprit des sommes de Riemann (*cf.* aussi, dans un cadre élémentaire, la remarque 3 de la page 288).

Les très importants théorèmes de convergence monotone et dominée de Lebesgue, qui permettent d'étudier l'intégrale de la limite d'une suite de fonctions, sont valides, *mutatis mutandis*, pour l'intégrale de Henstock-Kurzweil. Nous les démontrerons dans la partie 2.7 où nous étudierons également les fonctions définies par une intégrale.

Dans la partie 3, nous montrerons comment l'on peut *déduire* la théorie classique de la *mesure de Lebesgue* sur $\mathbb{R}$ de la théorie de l'intégration de Henstock-Kurzweil. Cette théorie de la mesure permet, en suivant les idées originales de Lebesgue, de construire la théorie « classique » de l'intégrale de Lebesgue dont nous donnons une « traduction » très simple en termes d'intégrale de Henstock-Kurzweil.

Pour finir, dans la partie 4, nous verrons comment la théorie de Henstock-Kurzweil se généralise dans le cas de l'intégrale de fonctions de plusieurs variables définies sur des ouverts de $\mathbb{R}^n$. Nous y montrerons une version du théorème de Fubini et du théorème de changement de variables pour les intégrales multiples, améliorant les résultats obtenus dans la partie 1.

Nous tenons à remercier Jean-Pierre Demailly qui a rédigé un texte [De] dont nous nous sommes *très largement* inspirés dans notre présentation de l'intégrale de Henstock-Kurzweil et de ses relations avec l'intégrale de Lebesgue.

[1] R. Godement, *Analyse mathématique IV*, Springer-Verlag, 1998.

Un peu d'Histoire

Les ancêtres de la théorie des intégrales multiples sont les calculs d'aires et de volumes dans l'antiquité (*cf. Un peu d'histoire* dans l'introduction du module *Intégration* du livre [L2]). La formulation moderne apparaît lors de la mise en place du calcul infinitésimal avec, tout d'abord, une notion *d'intégrale indéfinie* : on note $\iint f(x,y)\,\mathrm{d}x\mathrm{d}y$ une solution de $\dfrac{\partial^2 z}{\partial x \partial y} = f(x,y)$. Les intégrales doubles sur des domaines bornés limités par des arcs de courbes (algébriques ou analytiques) apparaissent chez Euler qui donne, en 1769, la formule de changement de variables pour les intégrales doubles sous la forme $\iint Z\,\mathrm{d}x\mathrm{d}y = \iint Z(VR - ST)\,\mathrm{d}t\mathrm{d}u$. Un peu plus tard, en 1773, Lagrange utilise des changements de variables en dimension 3 et l'on trouve des calculs analogues chez Legendre et Laplace. En 1841, indépendamment, Catalan et Jacobi donnent la formule de changement de variables dans les intégrales multiples en dimension quelconque.

1 Initiation aux intégrales multiples

Dans cette partie, nous allons aborder la notion *d'intégrale multiple*, c'est-à-dire d'intégrale d'une fonction de plusieurs variables, en nous plaçant d'abord dans le cadre de l'intégrale de Riemann. Nous mettrons l'accent sur le cas des fonctions *continues* sur des parties dites « cubables », cas le plus utile pour les applications élémentaires usuelles. Dans la partie 4 nous étendrons la théorie dans le cadre de l'intégration au sens de Henstock-Kurzweil.
Nous nous limiterons au cas des fonctions à valeurs réelles. On peut en déduire sans difficulté les définitions et résultats similaires pour le cas des fonctions à valeurs dans un espace vectoriel réel ou complexe *de dimension finie* en procédant comme dans la section 4 du module *Intégration* du livre [L2].
Dans cette initiation aux intégrales multiples, nous avons choisi de donner *le minimum de bases théoriques* nécessaires à une présentation *correcte et rigoureuse* des résultats fondamentaux pour les applications, en particulier en géométrie, physique, mécanique…
Dans un premier temps, le lecteur pourra se limiter aux cas de deux ou trois variables (intégrales doubles ou triples) et à des domaines limités par des inégalités définies par des fonctions *simples*. C'est suffisant pour un très grand nombre d'applications. Nous donnons aussi quelques indications permettant de « passer sur les preuves » d'un certain nombre de résultats un peu techniques.

Convention importante. Dans toute cette partie, on munit $\mathbb{R}^n$ ($n \in \mathbb{N}^*$) de la norme :

$$\|\mathbf{x} := (x_1, \ldots, x_n)\| := \max\{|x_1|, \ldots, |x_n|\}.$$

L'intégration à plusieurs variables est étroitement liée aux notions *d'aire* et de *volume* et à leur généralisation. On utilise donc implicitement la structure euclidienne canonique de $\mathbb{R}^n$. Notre présentation, basée sur la notion de *pavé*, est *a priori* liée à un *choix de coordonnées*. Nous verrons qu'il est finalement possible de s'en affranchir à l'aide du théorème de changement de variables (*cf.* la proposition 60 de la page 220). Tant que l'on ne dispose pas de ce résultat les propriétés d'invariance de l'aire et du volume par isométrie ne sont *pas du tout* évidentes (*cf.* la partie 1.4.3).

Un peu d'Histoire

Après les travaux de Riemann et Darboux sur l'intégrale simple, la notion d'intégrale de Riemann à plusieurs variables a été développée. Au départ, l'intégrale était définie sur un pavé, puis elle a été étendue à d'autres domaines (bornés) suivant une idée de Dirichlet que nous expliquerons plus loin (*cf.* la partie 1.2.2). Divers auteurs, en particulier Stolz, Pringsheim, Jordan (1892), Cantor, Peano (1887), se sont alors posé le problème de la nature des domaines sur lesquels on pouvait intégrer. Cela a conduit à la notion de parties *mesurables au sens de Peano-Jordan* que nous présentons ici sous le nom de parties *cubables* (ou *quarrables* en dimension $n = 2$). En relation avec ces progrès sur la mesure des ensembles, diverses formules « du type Fubini » permettant de calculer une intégrale multiple par itération de calculs d'intégrales simples ont été établies. Citons Stolz pour les intégrales doubles (1886).

1.1 Intégration sur un pavé

1.1.1 Pavés et fonctions en escalier

Les notions introduites pour des fonctions définies sur des segments dans le module *Intégration* du livre [L2] se généralisent facilement au cas de fonctions de plusieurs variables définies sur des pavés fermés bornés de $\mathbb{R}^n$.

Définition 1.

(i) Un *pavé* de $\mathbb{R}^n$ est un ensemble de la forme $\mathbf{P} := I_1 \times I_2 \times \cdots \times I_n$, où chaque I_j (arête) est un intervalle borné (a_j, b_j) de $\mathbb{R}$ (fermé, ouvert ou semi-ouvert, avec $a_j < b_j$). Le volume d'un pavé $\mathbf{P}$ est par définition $\operatorname{vol} \mathbf{P} := \prod_{j=1}^{n} \operatorname{long} I_j$.

Le diamètre de $\mathbf{P}$ est $\delta(\mathbf{P}) := \max_{j \in [\![1,n]\!]} (b_j - a_j)$.

Si $\mathbf{x} := (x_1, \ldots, x_n) \in \mathbb{R}^n$ et $\rho > 0$, le *pavé fermé de centre* $\mathbf{x}$ *et de rayon* ρ est :

$$\mathbf{P}(\mathbf{x}, \rho) := \big\{(y_1, \ldots, y_n) \in \mathbb{R}^n \;\big|\; \forall j \in [\![1, n]\!],\ |y_j - x_j| \leqslant \rho\big\}.$$

(ii) Une *subdivision* d'un pavé fermé $\mathbf{P}$ de $\mathbb{R}^n$ est une collection finie $\big\{\mathbf{P}_k\big\}_{1 \leqslant k \leqslant N}$ où les $\mathbf{P}_k$ sont des pavés fermés, *d'intérieurs deux à deux disjoints*, recouvrant $\mathbf{P}$.

(iii) Une *subdivision pointée* d'un pavé fermé $\mathbf{P}$ de $\mathbb{R}^n$ est une collection finie $\mathcal{D} := \big\{(\mathbf{P}_k, \xi_k)\big\}_{1 \leqslant k \leqslant N}$ où les $\mathbf{P}_k$ sont des pavés fermés, *d'intérieurs disjoints*, recouvrant $\mathbf{P}$ et où les ξ_k sont des points tels que $\xi_k \in \mathbf{P}_k$ pour tout $k \in [\![1, N]\!]$.

(iv) On appelle *pas* de la subdivision $\mathbf{u} := \big\{\mathbf{P}_k\big\}_{1 \leqslant k \leqslant N}$ le réel $\delta(\mathbf{u}) := \max_{1 \leqslant k \leqslant N} \delta(\mathbf{P}_k)$.

Un pavé $\mathbf{P} := I_1 \times I_2 \times \cdots \times I_n$ est fermé (respectivement ouvert) si, et seulement si, *tous* les I_j $(j = 1, \ldots, n)$ sont fermés (respectivement ouverts).

On dit qu'une subdivision d'un pavé fermé $\mathbf{P} := I_1 \times I_2 \times \cdots \times I_n$ est une *subdivision produit* si elle est obtenue en prenant des subdivisions de chacune des arêtes de $\mathbf{P}$ et en considérant tous les produits constitués d'un segment de chacune de ces subdivisions.

Un peu d'Histoire

La théorie de l'intégration que nous présentons dans la partie 1 était donc fixée, pour l'essentiel, au début du XXe siècle.[a]

Mais, vers la même époque, le sujet a été « révolutionné » par Borel et Lebesgue qui ont considérablement généralisé les notions de mesure et d'intégration, d'abord en dimension un puis en dimension quelconque.

Vers 1898, Borel a étendu, en dimension un, la notion de mesurabilité de Peano-Jordan en utilisant des unions *dénombrables* de segments (le passage du *fini* au *dénombrable* est un apport crucial). Lebesgue a amélioré ces résultats dans sa thèse *Intégrale, longueur, aire* soutenue à la Sorbonne en 1905 puis les a utilisés pour mettre en place une nouvelle théorie de l'intégration beaucoup plus « souple » et surtout d'utilisation bien plus facile. Nous en parlerons dans la partie 3. Plus tard, Lebesgue a étendu sa théorie aux intégrales multiples, nous donnerons une introduction à ce sujet (du point de vue de Henstock-Kurzweil) dans la section 4.

Dans ce cours, nous ne présentons pas la formule de Stokes (*cf.* par exemple [MPA1]). Nous n'évoquons donc pas non plus son histoire.

Récemment, l'étude de certaines intégrales multiples est (re)devenue un sujet de recherche très actif en mathématiques (théorie de la transcendance) et en physique théorique (intégrales de Feynmann en théorie quantique des champs) avec la théorie des *périodes*[b] On appelle *période* tout nombre complexe dont les parties réelles et imaginaires sont des *intégrales* (éventuellement généralisées mais *absolument convergentes*) de fonctions *rationnelles* à coefficients dans $\mathbb{Q}$ sur des domaines de $\mathbb{R}^n$ définis par des inégalités polynomiales à coefficients dans $\mathbb{Q}$. Par exemple :

$$\pi = \int_{\{x^2+y^2\leqslant 1\}} \mathrm{d}x\,\mathrm{d}y \quad \text{et} \quad \sqrt{2} = \int_{\{2x^2\leqslant 1\}} \mathrm{d}x$$

sont des périodes. Les nombres algébriques et beaucoup de nombres transcendants « intéressants » sont des périodes. Les « règles du calcul intégral » (linéarité, formules de changements algébriques de variables, formule de Stokes) fournissent des relations entre périodes (*cf.* les exercices 15 de la page 235 et III.1.20 de la page 365).

On conjecture que e n'est pas une période, ce qui a conduit à une notion plus générale, celle de *périodes exponentielle*. : on autorise l'intégrande à être multiplié par l'exponentielle d'une fonction algébrique. Par exemple : e, $\sqrt{\pi} = \displaystyle\int_{-\infty}^{+\infty} \mathrm{e}^{-x^2}\,\mathrm{d}x$ ou les valeurs de la fonction Γ aux points rationnels sont des périodes exponentielles.

[a] *cf.* par exemple : James Pierpoint *On Multiple Integrals*, 1905.

[b] Le lecteur intéressé pourra consulter : Maxim Kontsevitch and Don Zagier, Mathematics unlimited – 2001 and beyond, Springer-Verlag, 2001.

Définition 2. On dit qu'une subdivision $\mathbf{u}' := \{\mathbf{P}'_m\}_{m\in[\![1,N']\!]}$ du pavé $\mathbf{P}$ est *plus fine* qu'une subdivision $\mathbf{u} := \{\mathbf{P}_k\}_{k\in[\![1,N]\!]}$ de ce pavé si tout pavé $\mathbf{P}_k$ de $\mathbf{u}$ se décompose en une réunion de pavés de $\mathbf{u}'$. On dit aussi que la subdivision $\mathbf{u}'$ est emboîtée dans la subdivision $\mathbf{u}$.

On a une définition similaire pour les subdivisions pointées en utilisant les subdivisions non pointées sous-jacentes. Si $\mathbf{u}$ et $\mathbf{u}'$ sont deux subdivisions du pavé $\mathbf{P}$, il existe une subdivision de ce pavé plus fine que $\mathbf{u}$ et $\mathbf{u}'$. On peut par exemple prendre la subdivision formée de tous les pavés $\mathbf{P}_k \cap \mathbf{P}'_m$, $k \in [\![1, N]\!]$, $m \in [\![1, N']\!]$.
Notons qu'une subdivision $\mathbf{u}$ d'un pavé fermé $\mathbf{P}$ de $\mathbb{R}^n$ n'est pas nécessairement une subdivision produit. Cependant, en prenant les projections de tous les sommets de tous les pavés $\mathbf{P}_k$ de $\mathbf{u}$ sur chacun des axes, on définit des subdivisions de chaque arête d'où l'on déduit une subdivision produit $\{\mathbf{Q}_\ell\}$ plus fine que $\mathbf{u}$. Elle induit une subdivision produit sur chaque pavé $\mathbf{P}_k$.

Proposition 1. Soit $\mathbf{u} := \{\mathbf{P}_k\}_{k\in[\![1,N]\!]}$ une subdivision du pavé $\mathbf{P}$. Alors :

$$\operatorname{vol}\mathbf{P} = \sum_{k=1}^{N} \operatorname{vol}\mathbf{P}_k.$$

Démonstration. Dans le cas d'une subdivision produit, cela découle de la distributivité de la multiplication par rapport à l'addition. Dans le cas général, on s'y ramène en considérant une subdivision *produit* $\{\mathbf{Q}_\ell\}$ plus fine que $\{\mathbf{P}_k\}$ et en utilisant l'associativité de la somme. ■

Exercice 1.
Soient $\mathbf{P}$ et $\mathbf{P}'$ deux pavés fermés tels que $\mathbf{P}' \subset \mathbf{P}$.

1. Montrer qu'il existe une subdivision $\mathbf{u} := \{\mathbf{P}_k\}_{k\in[\![1,N]\!]}$ de $\mathbf{P}$ telle que $\mathbf{P}'$ soit l'un des $\mathbf{P}_k$.
2. Soit $\mathbf{v} := \{\mathbf{Q}_m\}_{k\in[\![1,M]\!]}$ une subdivision de $\mathbf{P}$. Montrer qu'il existe une subdivision $\mathbf{v}' := \{\mathbf{Q}'_r\}_{r\in[\![1,M']\!]}$ de $\mathbf{P}$ plus fine que $\mathbf{v}$ telle que $\mathbf{P}'$ soit une réunion de pavés $\mathbf{Q}'_r$.

Solution.

1. On considère les projections des sommets de $\mathbf{P}'$ sur les axes. On en déduit des subdivisions des arêtes de $\mathbf{P}$ (en 1, 2 ou 3 intervalles). On vérifie que la subdivision produit répond à la question.
2. Une subdivision plus fine que $\mathbf{v}$ et que la subdivision $\mathbf{u}$ obtenue à la question 1 répond à la question.

Définition 3. Soit $\mathbf{P}$ un pavé fermé. Une fonction $f : \mathbf{P} \to \mathbb{R}$ est *en escalier* si elle est *bornée* et s'il existe une subdivision $\{\mathbf{P}_k\}_{1\leqslant k\leqslant N}$ de $\mathbf{P}$ telle que f soit constante sur chacun des pavés *ouverts* $\operatorname{Int}\mathbf{P}_k$, $k \in [\![1, N]\!]$.
Une telle subdivision est dite *adaptée* à la fonction en escalier f.

Pour $n = 1$, on retrouve la définition donnée dans le module *Intégration* du livre [L2].
Si une subdivision est plus fine qu'une subdivision *adaptée* à f elle est aussi *adaptée* à f.
Si l'on a deux pavés fermés $\mathbf{Q} \subset \mathbf{P}$, une subdivision $\{\mathbf{P}_k\}_{1 \leqslant k \leqslant N}$ de $\mathbf{P}$ induit une subdivision $\{\mathbf{Q}_k := \mathbf{P}_k \cap \mathbf{Q}\}_{1 \leqslant k \leqslant N}$ de $\mathbf{Q}$. La restriction à $\mathbf{Q}$ d'une fonction en escalier sur $\mathbf{P}$ est une fonction en escalier sur $\mathbf{Q}$.
On note $\mathcal{E}(\mathbf{P})$ l'ensemble des fonctions en escalier sur $\mathbf{P}$.

Proposition 2. L'ensemble $\mathcal{E}(\mathbf{P})$ est un sous-espace vectoriel de $\mathcal{F}(\mathbf{P})$.

Démonstration. Soient $f, g \in \mathcal{E}(\mathbf{P})$ et $\mathbf{u}$, $\mathbf{v}$ des subdivisions de $\mathbf{P}$ adaptées respectivement à f et g. Soit $\mathbf{w}$ une subdivision plus fine que $\mathbf{u}$ et $\mathbf{v}$, elle est adaptée à f et g. Pour tous λ, $\mu \in \mathbb{R}$, la fonction $\lambda f + \mu g$ est constante sur chacun des intérieurs des pavés de $\mathbf{w}$, donc $\lambda f + \mu g$ est une fonction en escalier et la subdivision $\mathbf{w}$ est adaptée à cette fonction. ∎

Proposition et définition 3. Soient $f \in \mathcal{E}(\mathbf{P})$ et $\mathbf{u} := \{\mathbf{P}_k\}_{1 \leqslant k \leqslant N}$ une subdivision adaptée à f. Si, pour tout $k \in [\![1, N]\!]$, on note $c_k \in \mathbb{R}$ la valeur de f sur $\operatorname{Int} \mathbf{P}_k$, alors la quantité :

$$\sum_{k \in [\![1,N]\!]} c_k \operatorname{vol} \mathbf{P}_k$$

ne dépend pas de la subdivision adaptée $\mathbf{u}$ choisie.
On l'appelle *intégrale* de f sur $\mathbf{P}$, on la note $\displaystyle\int_{\mathbf{P}} f$.
L'application $f \mapsto \displaystyle\int_{\mathbf{P}} f$ est une forme linéaire positive sur $\mathcal{E}(\mathbf{P})$.

Démonstration. Il suffit de prouver le résultat pour deux subdivisions $\mathbf{u}$ et $\mathbf{v}$ telles que $\mathbf{u}$ soit plus fine que $\mathbf{v}$. On vérifie alors que l'on peut se ramener, par associativité de l'addition de $\mathbb{R}$, au cas où $\mathbf{v}$ est réduit au seul pavé $\mathbf{P}$ et où f est constante et égale à c sur $\operatorname{Int} \mathbf{P}$. Soit $\mathbf{u} := \{\mathbf{P}_k\}_{1 \leqslant k \leqslant N}$, on a, pour tout $k \in [\![1, N]\!]$, $c_k = c$, donc, en utilisant la proposition 1 :

$$\sum_{k \in [\![1,N]\!]} c_k \operatorname{vol} \mathbf{P}_k = \sum_{k \in [\![1,N]\!]} c \operatorname{vol} \mathbf{P}_k = c \sum_{k \in [\![1,N]\!]} \operatorname{vol} \mathbf{P}_k = c \operatorname{vol} \mathbf{P}.$$

On prouve la linéarité de l'intégrale en raisonnant comme dans la preuve de la proposition précédente. La positivité est évidente. ∎

On vérifie facilement que si l'on a deux pavés fermés $\mathbf{Q} \subset \mathbf{P}$ et si $f \geqslant 0$ est une fonction en escalier sur $\mathbf{P}$, on a $0 \leqslant \displaystyle\int_{\mathbf{Q}} f_{|\mathbf{Q}} \leqslant \int_{\mathbf{P}} f$.

1.1.2 Intégrale de Riemann sur un pavé

Dans cette partie $\mathbf{P}$ est un pavé *fermé* de $\mathbb{R}^n$. Nous généralisons ce que nous avons fait dans le cas $n = 1$, les preuves sont essentiellement les mêmes *mutatis mutandis* et nous les laissons dans la plupart des cas au lecteur.

Définition de l'intégrabilité au sens de Riemann sur un pavé

Définition 4. On dit que $f : \mathbf{P} \to \mathbb{R}$ est *intégrable au sens de Riemann* sur $\mathbf{P}$ si, pour tout $\varepsilon > 0$, il existe des fonctions en escalier φ_ε et ψ_ε sur $\mathbf{P}$ telles que $\psi_\varepsilon \leqslant f \leqslant \varphi_\varepsilon$ et $\displaystyle\int_{\mathbf{P}} \varphi_\varepsilon - \int_{\mathbf{P}} \psi_\varepsilon = \int_{\mathbf{P}} \big(\varphi_\varepsilon - \psi_\varepsilon\big) \leqslant \varepsilon$.

On écrira aussi que f est $\mathscr{R}$-intégrable sur $\mathbf{P}$. On notera $\mathscr{R}\big(\mathbf{P}\big)$ l'ensemble des fonctions intégrables au sens de Riemann sur $\mathbf{P}$.
Toute fonction en escalier étant bornée, il résulte immédiatement de la définition que toute fonction intégrable au sens de Riemann est *bornée*.

Proposition 4. Soient $\mathbf{Q} \subset \mathbf{P}$ deux pavés fermés et $f : \mathbf{P} \to \mathbb{R}$ une fonction $\mathscr{R}$-intégrable sur $\mathbf{P}$. Alors la restriction de f à $\mathbf{Q}$ est $\mathscr{R}$-intégrable sur $\mathbf{Q}$.

Démonstration. Soit $\varepsilon > 0$. Il existe des fonctions en escalier φ et ψ sur $\mathbf{P}$ telles que $\psi \leqslant f \leqslant \varphi$ et $\displaystyle\int_{\mathbf{P}} \big(\varphi - \psi\big) \leqslant \varepsilon$.

On a $\psi_{|\mathbf{Q}} \leqslant f_{|\mathbf{Q}} \leqslant \varphi_{|\mathbf{Q}}$ et $\displaystyle\int_{\mathbf{Q}} (\varphi_{|\mathbf{Q}} - \psi_{|\mathbf{Q}}) \leqslant \int_{\mathbf{P}} \big(\varphi - \psi\big) \leqslant \varepsilon$. Le résultat s'en déduit ($\varphi_{|\mathbf{Q}}$ et $\psi_{|\mathbf{Q}}$ sont des fonctions en escalier). ∎

Intégrales supérieure et inférieure

On peut, comme en dimension 1, définir les notions d'intégrale supérieure et d'intégrale inférieure (au sens de Darboux) de *toute fonction bornée* sur $\mathbf{P}$, ce qui permet de reformuler la notion d'intégrabilité au sens de Riemann.
Soit $f : \mathbf{P} \to \mathbb{R}$ une fonction *bornée*. Il existe $M, m \in \mathbb{R}$ tels que $m \leqslant f \leqslant M$. L'ensemble $\mathcal{E}_f^+ := \big\{\varphi \in \mathcal{E}\big(\mathbf{P}\big) \mid f \leqslant \varphi\big\}$ n'est pas vide (il contient la fonction constante M) et l'ensemble $\mathcal{E}_f^- := \big\{\psi \in \mathcal{E}\big(\mathbf{P}\big) \mid \psi \leqslant f\big\}$ n'est pas vide (il contient la fonction constante m). Si $\varphi \in \mathcal{E}_f^+$, $m \leqslant \varphi$, donc l'ensemble $\left\{\displaystyle\int_{\mathbf{P}} \varphi \mid \varphi \in \mathcal{E}_f^+\right\} \subset \mathbb{R}$ est *minoré* par $\displaystyle\int_{\mathbf{P}} m = m \operatorname{vol} \mathbf{P}$. De même si $\psi \in \mathcal{E}_f^-$, $\psi \leqslant M$, donc l'ensemble $\left\{\displaystyle\int_{\mathbf{P}} \psi \mid \psi \in \mathcal{E}_f^-\right\} \subset \mathbb{R}$ est *majoré* par $\displaystyle\int_{\mathbf{P}} M = M \operatorname{vol} \mathbf{P}$.
Cela justifie la définition suivante.

Proposition et définition 5. Soit $f : \mathbf{P} \to \mathbb{R}$ une fonction *bornée*.

(i) L'ensemble $\left\{ \int_a^b \varphi \mid \varphi \in \mathcal{E}_f^+ \right\} \subset \mathbb{R}$ est borné inférieurement. Sa *borne inférieure* est appelée *intégrale supérieure* de f sur $\mathbf{P}$ et notée $\displaystyle\overset{*}{\int_{\mathbf{P}}} f$.

(ii) L'ensemble $\left\{ \int_a^b \psi \mid \psi \in \mathcal{E}_f^- \right\} \subset \mathbb{R}$ est borné supérieurement. Sa *borne supérieure* est appelée *intégrale inférieure* de f sur $\mathbf{P}$ et notée $\displaystyle\underset{*}{\int_{\mathbf{P}}} f$.

Si f est une fonction en escalier, l'ensemble $\mathcal{E}_f^- := \{\psi \in \mathcal{E}(\mathbf{P}) \mid \psi \leqslant f\}$ (respectivement l'ensemble $\mathcal{E}_f^+ := \{\varphi \in \mathcal{E}(\mathbf{P}) \mid f \leqslant \varphi\}$) admet f comme plus grand (respectivement plus petit) élément. L'intégrale d'une fonction en escalier est donc égale à son intégrale supérieure et à son intégrale inférieure.
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 6. Soient $f, g : \mathbf{P} \to \mathbb{R}$ deux fonctions *bornées*.

(i) Pour tout $\lambda \geqslant 0$ on a :

$$\overset{*}{\int_{\mathbf{P}}} \lambda f = \lambda \overset{*}{\int_{\mathbf{P}}} f \quad \text{et} \quad \underset{*}{\int_{\mathbf{P}}} \lambda f = \lambda \underset{*}{\int_{\mathbf{P}}} f$$

(ii) Pour tout $\lambda \leqslant 0$ on a :

$$\overset{*}{\int_{\mathbf{P}}} \lambda f = \lambda \underset{*}{\int_{\mathbf{P}}} f \quad \text{et} \quad \underset{*}{\int_{\mathbf{P}}} \lambda f = \lambda \overset{*}{\int_{\mathbf{P}}} f$$

(iii) On a :

$$\overset{*}{\int_{\mathbf{P}}} (f+g) \leqslant \overset{*}{\int_{\mathbf{P}}} f + \overset{*}{\int_{\mathbf{P}}} g \quad \text{et} \quad \underset{*}{\int_{\mathbf{P}}} f + \underset{*}{\int_{\mathbf{P}}} g \leqslant \underset{*}{\int_{\mathbf{P}}} (f+g). \tag{1}$$

Intégrale d'une fonction intégrable au sens de Riemann

On vérifie facilement le résultat suivant.

Proposition 7. Soit $f : \mathbf{P} \to \mathbb{R}$. Les deux conditions suivantes sont équivalentes :

(i) f est intégrable au sens de Riemann sur $\mathbf{P}$;

(ii) f est bornée et ses intégrales supérieure et inférieure sont égales.

Cette proposition permet la définition suivante.

Définition 5. Soit $f : \mathbf{P} \to \mathbb{R}$ une fonction intégrable au sens de Riemann sur $\mathbf{P}$. On *définit l'intégrale de f sur* $\mathbf{P}$ par :

$$\int_{\mathbf{P}} f = {}^{*}\!\!\int_{\mathbf{P}} f = {}_{*}\!\!\int_{\mathbf{P}} f.$$

Si f est une fonction en escalier cette définition coïncide évidemment avec celle donnée précédemment.

Proposition 8. Toute fonction *continue* sur un pavé fermé est intégrable au sens de Riemann sur ce pavé.

Démonstration. Soit $f : \mathbf{P} \to \mathbb{R}$ 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 $\displaystyle\int_{\mathbf{P}} \varphi - \int_{\mathbf{P}} \psi \leqslant \varepsilon$.

D'après le théorème de Heine, la fonction f est *uniformément continue* sur le pavé fermé $\mathbf{P} = I_1 \times \cdots \times I_n$. Donc pour tout $\varepsilon > 0$ il existe $\delta > 0$ tel que si $x, y \in \mathbf{P}$ sont tels que $\|y - x\| \leqslant \delta$, l'on ait $\big|f(y) - f(x)\big| \leqslant \frac{\varepsilon}{\operatorname{vol} \mathbf{P}}$. Il existe $m \in \mathbb{N}^*$ tel que $\displaystyle\max_{j \in [\![1,n]\!]} \frac{\operatorname{long} I_j}{m} \leqslant \delta$.

On définit une subdivision $\mathbf{u}_m := \{\mathbf{P_k}\}_{k \in [\![1,m^n]\!]}$ comme produit des subdivisions des I_j en m parties égales. Cette subdivision est de pas inférieur à δ.

Notons :

$$M_k := \sup_{x \in \mathbf{P}_k} f(x) \quad \text{et} \quad m_k := \inf_{x \in \mathbf{P}_k} f(x).$$

Il existe $\mathbf{x}_k, \mathbf{y}_k \in \mathbf{P}_k$ tels que $f(\mathbf{x}_k) = M_k$ et $f(\mathbf{y}_k) = m_k$, donc $0 \leqslant M_k - m_k \leqslant \frac{\varepsilon}{\operatorname{vol} \mathbf{P}}$.

La fonction $\psi : \mathbf{P} \to \mathbb{R}$ définie par :

$$\begin{cases} \forall k \in [\![1, m^n]\!],\ \forall \mathbf{x} \in \operatorname{Int} \mathbf{P}_k, \quad \psi(\mathbf{x}) := m_k \\ \forall \mathbf{x} \in \mathbf{P} \setminus \displaystyle\bigcup_{k \in [\![1,m^n]\!]} \operatorname{Int} \mathbf{P}_k, \quad \psi(\mathbf{x}) := f(\mathbf{x}) \end{cases}$$

est un élément de $\mathcal{E}_f^-$ et la fonction $\varphi : \mathbf{P} \to \mathbb{R}$ définie par :

$$\begin{cases} \forall k \in [\![1, m^n]\!],\ \forall \mathbf{x} \in \operatorname{Int} \mathbf{P}_k, \quad \varphi(\mathbf{x}) := M_k \\ \forall \mathbf{x} \in \mathbf{P} \setminus \displaystyle\bigcup_{k \in [\![1,m^n]\!]} \operatorname{Int} \mathbf{P}_k, \quad \varphi(\mathbf{x}) := f(\mathbf{x}) \end{cases}$$

est un élément de $\mathcal{E}_f^+$.

On a :

$$\int_{\mathbf{P}} \varphi - \int_{\mathbf{P}} \psi = \sum_{k \in [\![1,m^n]\!]} (M_k - m_k) \operatorname{vol} \mathbf{P}_k \leqslant \operatorname{vol} \mathbf{P} \frac{\varepsilon}{\operatorname{vol} \mathbf{P}} = \varepsilon.$$

Ainsi f est $\mathscr{R}$-intégrable. ∎

Propriétés des fonctions intégrables et de l'intégrale

Certains des résultats énoncés ci-dessous (les propositions 9, 10, 13, 14) ont été prouvés dans le cas de l'intégrale simple ($n = 1$) dans le module *Intégration* du livre [L1]. Les preuves sont similaires, *mutatis mutandis*, et nous ne les détaillerons pas.

Proposition 9. L'ensemble $\mathscr{R}(\mathbf{P})$ est un *sous-espace vectoriel* de $\mathcal{F}(\mathbf{P})$ et l'application $f \in \mathscr{R}(\mathbf{P}) \mapsto \displaystyle\int_{\mathbf{P}} f \in \mathbb{R}$ est une *forme linéaire* sur $\mathscr{R}(\mathbf{P})$.

Proposition 10. Soient $f, g : \mathbf{P} \to \mathbb{R}$ intégrables au sens de Riemann sur $\mathbf{P}$.

(i) Si $f \geqslant 0$, alors $\displaystyle\int_{\mathbf{P}} f \geqslant 0$. La forme linéaire $f \mapsto \displaystyle\int_a^b f$ sur $\mathscr{R}(\mathbf{P})$ est *positive*.

(ii) Si $f \geqslant g$, alors $\displaystyle\int_{\mathbf{P}} f \geqslant \int_{\mathbf{P}} g$.

Démonstration. Si $f \geqslant 0$, la fonction nulle appartient à $\mathcal{E}_f^-$, donc $\displaystyle\int_{\mathbf{P}} f \geqslant 0$. Si $g \geqslant f$, $g - f \geqslant 0$ et il suffit d'appliquer le résultat précédent. ■

Proposition 11. Soient $\mathbf{P}' \subset \mathbf{P} \subset \mathbb{R}^n$ deux pavés fermés et $f : \mathbf{P}' \to \mathbb{R}$, on note $\tilde{f} : \mathbf{P} \to \mathbb{R}$ le prolongement par 0 de f à $\mathbf{P}$ ($\tilde{f}(\mathbf{x}) := f(\mathbf{x})$ si $\mathbf{x} \in \mathbf{P}'$ et $\tilde{f}(\mathbf{x}) := 0$ sinon). Les deux conditions suivantes sont équivalentes.

(i) La fonction f est intégrable au sens de Riemann sur $\mathbf{P}'$.

(ii) La fonction $\tilde{f}$ est intégrable au sens de Riemann sur $\mathbf{P}$.

Si elles sont satisfaites, alors $\displaystyle\int_{\mathbf{P}'} f = \int_{\mathbf{P}} \tilde{f}$.

Démonstration. On a (ii) $\Longrightarrow$ (i) d'après la proposition 4 de la page 180.

Montrons (i) $\Longrightarrow$ (ii). Supposons f intégrable au sens de Riemann sur $\mathbf{P}'$. D'après le résultat de la question 2 de l'exercice 1 de la page 178, il existe une subdivision $\mathbf{u} := \{\mathbf{P}_k\}_{k \in [\![1,N]\!]}$ de $\mathbf{P}$ telle qu'il existe $j \in [\![1, N]\!]$ tel que $\mathbf{P}_j = \mathbf{P}'$.

Soit $\varepsilon > 0$. Il existe deux fonctions en escalier φ, $\psi : \mathbf{P}' \to \mathbb{R}$ telles que $\psi \leqslant f \leqslant \varphi$ sur $\mathbf{P}'$ et $\displaystyle\int_{\mathbf{P}'} \varphi - \int_{\mathbf{P}'} \psi \leqslant \varepsilon$. On désigne par $\tilde{\varphi}$ et $\tilde{\psi}$ les prolongements respectifs de φ et ψ par 0 sur $\mathbf{P}$. On vérifie que ce sont des fonctions en escalier et que l'on a $\tilde{\psi} \leqslant \tilde{f} \leqslant \tilde{\varphi}$ sur $\mathbf{P}'$, $\displaystyle\int_{\mathbf{P}'} \varphi = \int_{\mathbf{P}} \tilde{\varphi}$, $\displaystyle\int_{\mathbf{P}'} \psi = \int_{\mathbf{P}} \tilde{\psi}$. Par suite, $\displaystyle\int_{\mathbf{P}} \tilde{\varphi} - \int_{\mathbf{P}} \tilde{\psi} \leqslant \varepsilon$.

Ainsi $\tilde{f}$ est intégrable sur $\mathbf{P}$ et l'on a $\displaystyle\int_{\mathbf{P}'} \psi \leqslant \int_{\mathbf{P}} \tilde{f} \leqslant \int_{\mathbf{P}'} \varphi$, donc $\displaystyle\int_{\mathbf{P}'} f = \int_{\mathbf{P}} \tilde{f}$. ■

Corollaire 12. Soient $\mathbf{u} := \{\mathbf{P}_k\}_{k\in[\![1,N]\!]}$ une subdivision de $\mathbf{P}$ et $f : \mathbf{P} \to \mathbb{R}$. Alors f est $\mathscr{R}$-intégrable sur $\mathbf{P}$ si, et seulement si, pour tout $k \in [\![1, N]\!]$, la restriction $f_{|\mathbf{P}_k}$ est $\mathscr{R}$-intégrable sur $\mathbf{P}_k$. S'il en est ainsi $\displaystyle\int_{\mathbf{P}} f = \sum_{k\in[\![1,N]\!]} \int_{\mathbf{P}_k} f_{|\mathbf{P}_k}$.

Démonstration. La condition est nécessaire d'après la proposition 4 de la page 180. Montrons qu'elle est suffisante. Pour tout $k \in [\![1, N]\!]$, on note g_k la restriction de f à $\mathbf{P}_k$ et l'on suppose que g_k est $\mathscr{R}$-intégrable sur $\mathbf{P}_k$. On note $\tilde{g}_k : \mathbf{P} \to \mathbb{R}$ la fonction obtenue en prolongeant g_k par 0. D'après la proposition ci-dessus, $\tilde{g}_k$ est $\mathscr{R}$-intégrable sur $\mathbf{P}$ et $\displaystyle\int_{\mathbf{P}} \tilde{g}_k = \int_{\mathbf{P}_k} g_k$.
Notons $g := \displaystyle\sum_{k\in[\![1,N]\!]} \tilde{g}_k$, c'est une fonction $\mathscr{R}$-intégrable sur $\mathbf{P}$ et l'on a :

$$\int_{\mathbf{P}} g = \sum_{k\in[\![1,N]\!]} \int_{\mathbf{P}} \tilde{g}_k = \sum_{k\in[\![1,N]\!]} \int_{\mathbf{P}_k} g_k.$$

Pour tout $k \in [\![1, N]\!]$ la restriction de g à $\operatorname{Int} \mathbf{P}_k$ est égale à celle de g_k donc à celle de f, il en résulte que $f - g$ est une fonction en escalier identiquement nulle sur $\displaystyle\bigcup_{k\in[\![1,N]\!]} \operatorname{Int} \mathbf{P}_k$.

Cette fonction est $\mathscr{R}$-intégrable sur $\mathbf{P}$ et $\displaystyle\int_{\mathbf{P}} (f-g) = 0$. On en déduit que $f = (f-g)+g$ est $\mathscr{R}$-intégrable sur $\mathbf{P}$ et que :

$$\int_{\mathbf{P}} f = \int_{\mathbf{P}} (f-g) + \int_{\mathbf{P}} g = \int_{\mathbf{P}} g = \sum_{k\in[\![1,N]\!]} \int_{\mathbf{P}_k} g_k. \qquad \blacksquare$$

Proposition 13. Soit $f : \mathbf{P} \to \mathbb{R}$ intégrable au sens de Riemann sur $\mathbf{P}$. Alors f^+, f^- et $|f|$ sont intégrables au sens de Riemann sur $\mathbf{P}$. On a les inégalités :

$$\Big|\int_{\mathbf{P}} f\Big| \leqslant \int_{\mathbf{P}} |f| \leqslant \operatorname{vol} \mathbf{P} \sup_{x\in\mathbf{P}} \big|f(x)\big|. \tag{2}$$

Proposition 14. Soient f, $g : \mathbf{P} \to \mathbb{R}$ intégrables au sens de Riemann sur $\mathbf{P}$. Alors fg est intégrable au sens de Riemann sur $\mathbf{P}$.

1.1.3 Sommes de Darboux et sommes de Riemann

Dans cette partie $\mathbf{P}$ est un pavé *fermé* de $\mathbb{R}^n$.

Sommes de Darboux

Définition 6 (Sommes de Darboux). Soit $f : \mathbf{P} \to \mathbb{R}$ une fonction *bornée*. À toute subdivision $\mathbf{u} := \{\mathbf{P}_k\}_{k\in[\![1,N]\!]}$ de $\mathbf{P}$ on associe les sommes :

$$d_{\mathbf{u}}(f) := \sum_{k=1}^{N} m_k \operatorname{vol} \mathbf{P}_k \quad \text{et} \quad D_{\mathbf{u}}(f) := \sum_{k=1}^{N} M_k \operatorname{vol} \mathbf{P}_k,$$

où m_k et M_k désignent respectivement la borne inférieure et la borne supérieure de f sur le pavé $\mathbf{P}_k$. Ces deux nombres réels sont dits *sommes de Darboux inférieure et supérieure* de f pour la subdivision $\mathbf{u}$.

Soit $\mathbf{u} := \{\mathbf{P}_k\}_{k\in[\![1,N]\!]}$ une subdivision de $\mathbf{P}$. La fonction $f^d_{\mathbf{u}} : \mathbf{P} \to \mathbb{R}$ définie par :

$$\begin{cases} \forall k \in [\![1,N]\!],\ \forall \mathbf{x} \in \operatorname{Int} \mathbf{P}_k, & f^d_{\mathbf{u}}(\mathbf{x}) := m_k \\ \forall \mathbf{x} \in \mathbf{P} \setminus \displaystyle\bigcup_{k\in[\![1,m^n]\!]} \operatorname{Int} \mathbf{P_k} & f^d_{\mathbf{u}}(\mathbf{x}) := f(\mathbf{x}) \end{cases}$$

est un élément de $\mathcal{E}^-_f$ et la fonction $f^D_{\mathbf{u}} : \mathbf{P} \to \mathbb{R}$ définie par :

$$\begin{cases} \forall k \in [\![1,N]\!],\ \forall \mathbf{x} \in \operatorname{Int} \mathbf{P}_k, & f^D_{\mathbf{u}}(\mathbf{x}) = M_k \\ \forall \mathbf{x} \in \mathbf{P} \setminus \displaystyle\bigcup_{k\in[\![1,m^n]\!]} \operatorname{Int} \mathbf{P}_k, & f^D_{\mathbf{u}}(\mathbf{x}) := f(\mathbf{x}) \end{cases}$$

est un élément de $\mathcal{E}^+_f$.

On a :

$$\int_{\mathbf{P}} f^d_{\mathbf{u}} = d_{\mathbf{u}}(f) \qquad \text{et} \qquad \int_{\mathbf{P}} f^D_{\mathbf{u}} = D_{\mathbf{u}}(f).$$

On en déduit :

$$d_{\mathbf{u}}(f) \leqslant {}_{*}\!\!\int_{\mathbf{P}} f \leqslant \overset{*}{\int_{\mathbf{P}}} f \leqslant D_{\mathbf{u}}(f). \tag{3}$$

Le lecteur vérifiera que pour toute subdivision $\mathbf{v}$ plus fine que la subdivision $\mathbf{u}$, l'on a :

$$d_{\mathbf{u}}(f) \leqslant d_{\mathbf{v}}(f) \leqslant D_{\mathbf{v}}(f) \leqslant D_{\mathbf{u}}(f). \tag{4}$$

Nous allons voir que (tout comme en dimension un), pour toute fonction *bornée* f sur $\mathbf{P}$, l'intégrale supérieure $\overset{*}{\int_{\mathbf{P}}} f$ (respectivement inférieure ${}_{*}\!\!\int_{\mathbf{P}} f$) est en un certain sens la « limite » des sommes de Darboux $D_{\mathbf{u}}(f)$ (respectivement $d_{\mathbf{u}}(f)$) quand le pas $\delta(\mathbf{u})$ de la subdivision $\mathbf{u}$ tend vers 0.

Le point clé de la démonstration est, comme en dimension un, le lemme suivant. La démonstration et la preuve du lemme, sont similaires *mutatis mutandis* à celles faites en dimensin un, elles sont laissées au lecteur.

Lemme 15. Soient $f : \mathbf{P} \to \mathbb{R}$ une fonction *bornée* et $g \in \mathcal{E}(\mathbf{P})$ 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 $\mathbf{P}$ de pas inférieur à δ, l'on ait :

$$\int_{\mathbf{P}} g - \varepsilon \leqslant d_{\mathbf{u}}(f) \leqslant {}_{*}\!\!\int_{\mathbf{P}} f.$$

Théorème 16 (Théorème de Darboux). Soit $f : \mathbf{P} \to \mathbb{R}$ une fonction *bornée*. Pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour toute subdivision $\mathbf{u}$ de $\mathbf{P}$ de pas inférieur à δ, l'on ait :

$$0 \leqslant \underset{*}{\int_{\mathbf{P}}} f - d_{\mathbf{u}}(f) \leqslant \varepsilon \quad \text{et} \quad 0 \leqslant D_{\mathbf{u}}(f) - \overset{*}{\int_{\mathbf{P}}} f \leqslant \varepsilon.$$

Notons $d(f) := \sup\limits_{\mathbf{u}} d_{\mathbf{u}}(f)$ et $D(f) := \inf\limits_{\mathbf{u}} D_{\mathbf{u}}(f)$ (les $\inf$ et $\sup$ étant pris sur toutes les subdivisions de $\mathbf{P}$).

Corollaire 17.
Soit $f : \mathbf{P} \to \mathbb{R}$ une fonction *bornée*. On a $d(f) = \underset{*}{\int_{\mathbf{P}}} f$ et $D(f) = \overset{*}{\int_{\mathbf{P}}} f$.

Corollaire 18. (Critère de Darboux) Soit $f : \mathbf{P} \to \mathbb{R}$ 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 $\mathbf{u} := \{\mathbf{P}_k\}_{k \in [\![1,N]\!]}$ telle que :

$$D_{\mathbf{u}}(f) - d_{\mathbf{u}}(f) = \sum_{k=1}^{N} (M_k - m_k) \operatorname{vol}(\mathbf{P}_k) \leqslant \varepsilon.$$

Nous « traduirons » plus loin le théorème de Darboux et son corollaire en montrant que l'intégrale d'une fonction positive f $\mathscr{R}$-intégrable sur un pavé représente le « volume sous le graphe de f ».

Sommes de Riemann

Définition 7. Soient $\mathbf{P} \subset \mathbb{R}^n$ un pavé fermé et $f : \mathbf{P} \to \mathbb{R}$ une fonction. La *somme de Riemann* de f pour une subdivision pointée $\mathcal{D} := \{(\mathbf{P}_k, \xi_k)\}_{k \in [\![1,N]\!]}$ de $\mathbf{P}$ est :

$$S_{\mathcal{D}}(f) = \sum_{k \in [\![1,N]\!]} f(\xi_k) \operatorname{vol} \mathbf{P}_k.$$

On prouve le résultat suivant comme dans le cas $n = 1$ (*cf.* la partie *Sommes de Riemann* du module *Intégration* du livre [L2]). Nous l'utiliserons plus loin pour comparer la $\mathscr{R}$-intégrabilité et la $\mathscr{HK}$-intégrabilité à plusieurs variables (*cf.* la partie 4).

Proposition 19. Soient $\mathbf{P} \subset \mathbb{R}^n$ un pavé fermé et $f : \mathbf{P} \to \mathbb{R}$ une fonction.
Les assertions suivantes sont équivalentes :

(i) f est $\mathscr{R}$-intégrable sur $\mathbf{P}$;

(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 $\mathbf{P}$ de pas inférieur à δ, l'on ait $\left|S_{\mathcal{D}}(f) - I\right| \leqslant \varepsilon$.

(iii) f est bornée et $d(f) = D(f)$;

(iv) f est bornée et $\displaystyle\underset{*}{\int}_{\mathbf{P}} f = \overset{*}{\int}_{\mathbf{P}}$.

Si ces conditions sont satisfaites, on a :

$$I = \int_{\mathbf{P}} f = \underset{*}{\int}_{\mathbf{P}} f = \overset{*}{\int}_{\mathbf{P}} f = d(f) = D(f).$$

1.1.4 Caractérisation des fonctions $\mathscr{R}$-intégrables sur un pavé

Sous-ensembles négligeables

Dans le module *Intégration* du livre [L2] nous avons introduit la notion de *sous-ensemble négligeable* de la droite réelle. La principale motivation était de l'utiliser pour caractériser les fonctions $\mathscr{R}$-intégrables sur un intervalle. Nous allons étendre la définition au cas de plusieurs variables, avec une variante, la notion de sous-ensemble $\mathscr{R}$-négligeable. Nous utiliserons ces notions pour caractériser d'une part les fonctions $\mathscr{R}$-intégrables sur un pavé et d'autre part les parties dites cubables sur lesquelles on peut intégrer une fonction continue. Nous reviendrons sur la question en étudiant la mesure de Lebesgue sur $\mathbb{R}$ dans la partie 3.

Définition 8. Une partie A de $\mathbb{R}^n$ est dite *négligeable* si pour tout $\varepsilon > 0$, il existe une *suite* $(\mathbf{P}_n)_{n\in\mathbb{N}}$ de pavés fermés telle que :

(i) $A \subset \bigcup\limits_{n\in\mathbb{N}} \mathbf{P}_n$;

(ii) $\sum\limits_{n\in\mathbb{N}} \operatorname{vol} \mathbf{P}_n \leqslant \varepsilon$.

Définition 9. Une partie A de $\mathbb{R}^n$ est dite *négligeable au sens de Riemann* (ou $\mathscr{R}$-négligeable) si pour tout $\varepsilon > 0$, il existe une famille *finie* $(\mathbf{P}_i)_{i\in I}$ de pavés fermés telle que :

(i) $A \subset \bigcup\limits_{i\in I} \mathbf{P}_i$;

(ii) $\sum\limits_{i\in I} \operatorname{vol} \mathbf{P}_i \leqslant \varepsilon$.

On vérifie facilement que l'on obtient des définitions équivalentes en remplaçant les pavés fermés par des pavés ouverts. Une partie $\mathscr{R}$-négligeable est évidemment négligeable. Tout sous-ensemble d'une partie négligeable (respectivement $\mathscr{R}$-négligeable) est négligeable

(respectivement $\mathscr{R}$-négligeable). Toute réunion *finie* de parties $\mathscr{R}$-négligeables est $\mathscr{R}$-négligeable.
On dit qu'une propriété portant sur les points d'une partie $C \subset \mathbb{R}^n$ est vraie *presque partout* sur C si elle vraie sur le complémentaire dans C d'un ensemble négligeable de $\mathbb{R}^n$ contenu dans C.

On prouve le résultat suivant en reprenant la méthode utilisée en dimension $n = 1$.

Proposition 20. Toute réunion *dénombrable* de parties négligeables est négligeable.

Exercice 2.
Soient $n \in \mathbb{N}^*$ et $p, q \in \mathbb{N}^*$ tels que $n = p + q$. Soit $A \subset \mathbb{R}^p$ une partie bornée.
Montrer que $A \times \{0_{\mathbb{R}^q}\}$ est une partie $\mathscr{R}$-négligeable de $\mathbb{R}^n$.
Solution. Soit $\mathbf{P}$ un pavé de $\mathbb{R}^p$ contenant A. Pour tout $r > 0$, on note $B_{r/2}$ la boule fermée de $\mathbb{R}^q$ de centre 0 et de rayon $r/2$ pour la norme $\|.\|_\infty$ de $\mathbb{R}^q$. La boule $B_{r/2}$ est un pavé de $\mathbb{R}^q$ et $\mathbf{P}' := \mathbf{P} \times B_{r/2}$ est un pavé de $\mathbb{R}^n$ de volume $\operatorname{vol} \mathbf{P}' = \operatorname{vol} \mathbf{P} \times r^q$.
Si $\varepsilon > 0$, il existe $r > 0$ tel que $\operatorname{vol} \mathbf{P} \times r^q < \varepsilon$.

En utilisant la propriété de Borel-Lebesgue pour les compacts, on obtient le résultat suivant.

Proposition 21. Soit K un *compact* de $\mathbb{R}^n$.
Les deux conditions suivantes sont équivalentes :

(i) K est négligeable ;

(ii) K est $\mathscr{R}$-négligeable.

La proposition suivante est importante. Elle permet d'obtenir des critères pour montrer qu'une partie est $\mathscr{R}$-négligeable.

Proposition 22. Soient $K \subset \mathbb{R}^p$ un compact et $f : K \to \mathbb{R}^q$ une application *continue*. Le graphe de f est une partie $\mathscr{R}$-négligeable de $\mathbb{R}^{p+q}$.

Démonstration. On vérifie qu'il suffit de prouver le résultat pour $q = 1$. Supposons $q \geqslant 2$ et notons $f = (f_1, \ldots, f_{q-1}, f_q)$. Le graphe de $(f_1, \ldots, f_{q-1})$ est un compact $K' \subset \mathbb{R}^{p+q-1}$. Le graphe de f s'interprète comme le graphe de l'application $K' \to \mathbb{R}$ définie par $\big(\mathbf{x}, f_1(\mathbf{x}), \ldots, f_{q-1}(\mathbf{x})\big) \mapsto f_q(\mathbf{x})$.
Soit maintenant $q = 1$. Il existe un pavé fermé $\mathbf{P} := I^p$ de $\mathbb{R}^p$ contenant K.
On note $\Gamma \subset \mathbb{R}^p \times \mathbb{R} = \mathbb{R}^{p+1}$ le graphe de f. On reprend la méthode de la preuve de la proposition 8 de la page 182.
D'après le théorème de Heine, la fonction f est *uniformément continue* sur K. Donc pour tout $\varepsilon > 0$ il existe $\delta > 0$ tel que si $\mathbf{x}$, $\mathbf{y} \in K$ vérifient $\|\mathbf{y} - \mathbf{x}\| \leqslant \delta$, l'on ait $\big|f(\mathbf{y}) - f(\mathbf{x})\big| \leqslant \dfrac{\varepsilon}{2 \operatorname{vol} \mathbf{P}}$. Il existe $m \in \mathbb{N}^*$ tel que $\dfrac{\operatorname{long} I}{m} \leqslant \delta$. On définit une subdivision $\mathbf{u} := \{\mathbf{P}_k\}_{k \in [\![1, m^p]\!]}$ comme produit des subdivisions des p facteurs I en m parties

égales. Cette subdivision est de pas inférieur à δ. On a :

$$\forall k \in [\![1, m^p]\!]\,,\ \operatorname{vol} \mathbf{P}_k = \frac{\operatorname{vol} \mathbf{P}}{m^p}.$$

Soit $J \subset [\![1, m^p]\!]$ le sous-ensemble formé des indices k tels que $\mathbf{P}_k \cap K \neq \varnothing$. La famille $(\mathbf{P}_k)_{k \in J}$ est un recouvrement de K.
Pour tout $k \in J$, on choisit un point $\xi_k \in \mathbf{P}_k \cap K$ et l'on note :

$$\mathbf{Q}_k := \mathbf{P}_k \times \left[f(\xi_k) - \frac{\varepsilon}{2 \operatorname{vol} \mathbf{P}}, f(\xi_k) + \frac{\varepsilon}{2 \operatorname{vol} \mathbf{P}} \right].$$

On a $\operatorname{vol} \mathbf{Q}_k = \dfrac{\varepsilon}{\operatorname{vol} \mathbf{P}} \operatorname{vol} \mathbf{P}_k = \dfrac{\varepsilon}{m^p}$.
La famille $(\mathbf{Q}_k)_{k \in J}$ est un recouvrement de Γ. En effet, si $(\mathbf{x}, z) \in \Gamma$, $z = f(\mathbf{x})$ et $\mathbf{x} \in K$. Il existe donc $k \in J$ tel que $\mathbf{x} \in \mathbf{P}_k$ et l'on a $|z - f(\xi_k)| \leqslant \frac{\varepsilon}{2 \operatorname{vol} \mathbf{P}}$, donc $(\mathbf{x}, z) \in \mathbf{Q}_k$. On a :

$$\sum_{k \in J} \operatorname{vol} \mathbf{Q}_k \leqslant (\operatorname{card}\ J) \operatorname{vol} \mathbf{Q}_k \leqslant m^p \frac{\varepsilon}{m^p} = \varepsilon. \qquad \blacksquare$$

Exemples.

1. Toute partie bornée A d'un hyperplan de $\mathbb{R}^n$ est $\mathscr{R}$-négligeable. En effet, quitte à permuter les coordonnées, tout hyperplan de $\mathbb{R}^n$ peut être interprété comme le graphe d'une fonction affine $f : \mathbb{R}^{n-1} \to \mathbb{R}$ et l'on peut appliquer la proposition précédente à la projection sur $\mathbb{R}^{n-1}$ de l'adhérence $\bar{A}$ de A.
2. Tout hyperplan $H \subset \mathbb{R}^n$ est réunion *dénombrable* de parties bornées $A_k \subset H$, $k \in \mathbb{N}$, donc, d'après 1, H est négligeable.
3. Toute sous-variété affine $V \subset \mathbb{R}^n$, $V \neq \mathbb{R}^n$, est contenue dans un hyperplan, donc V est négligeable.
4. La frontière d'un pavé de $\mathbb{R}^n$ est réunion finie de parties bornées d'hyperplans, elle est donc $\mathscr{R}$-négligeable.

On montre souvent qu'un ensemble de $\mathbb{R}^n$ est négligeable en prouvant qu'il est réunion finie de graphes de fonctions continues $f : K \subset \mathbb{R}^p \to \mathbb{R}^q$, où K est compact et $p + q = n$. Nous allons terminer ce paragraphe en donnant quelques résultats dans cette direction. Nous reviendrons plus loin sur cette méthode dans la partie 1.2.2 et dans divers exemples.

Exercice 3.

Soit $n \in \mathbb{N}^*$. Montrer que la sphère unité S de $\mathbb{R}^n$ muni de sa structure euclidienne canonique est $\mathscr{R}$-négligeable.
Solution. Si $n = 1$, $S = \{\pm 1\} \subset \mathbb{R}$, donc S est $\mathscr{R}$-négligeable.
Supposons $n \geqslant 2$. Alors $S = S^+ \cup S^-$, où :

$$S^+ := \{\mathbf{x} := (x_1, \ldots, x_n) \in S \mid x_n \geqslant 0\} \quad \text{et} \quad S^- := \{\mathbf{x} := (x_1, \ldots, x_n) \in S \mid x_n \leqslant 0\}$$

sont respectivement la demi-sphère « supérieure » et la demi-sphère « inférieure » fermées.

Notons $B := \{(x_1, \dots, x_{n-1}) \in \mathbb{R}^{n-1} \mid x_1^1 + \cdots + x_{n-1}^2 \leqslant 1\}$. Les demi-sphères S^+ et S^- sont les graphes des applications continues :

$$\begin{array}{rccl} f^+ : & B & \longrightarrow & \mathbb{R} \\ & (x_1, \dots, x_{n-1}) & \longmapsto & \sqrt{1 - x_1^2 - \cdots - x_{n-1}^2} \end{array}$$

et :

$$\begin{array}{rccl} f^- : & B & \longrightarrow & \mathbb{R} \\ & (x_1, \dots, x_{n-1}) & \longmapsto & -\sqrt{1 - x_1^2 - \cdots - x_{n-1}^2}. \end{array}$$

La boule fermée B étant compacte, S^+ et S^- sont $\mathscr{R}$-négligeables, donc leur réunion S est $\mathscr{R}$-négligeable.

Nous allons généraliser le cas de l'exercice ci-dessus, d'une part au cas d'une variété lisse compacte et d'autre part à la frontière d'une partie convexe bornée.
Pour énoncer et prouver les deux résultats suivants nous utiliserons des définitions et résultats de la partie *Théorème d'inversion locale — Théorème des fonctions implicites* du module *Fonctions de plusieurs variables* à laquelle nous renvoyons le lecteur.

Proposition 23. Soit $\mathbf{X}$ une sous-variété lisse de classe $\mathcal{C}^1$ de $\mathbb{R}^n$, de dimension *strictement inférieure* à n.

(i) Pour tout point $\mathbf{a} \in \mathbf{X}$ il existe un voisinage ouvert $U_{\mathbf{a}}$ de $\mathbf{a}$ dans $\mathbb{R}^n$ tel que $\mathbf{X} \cap U_{\mathbf{a}}$ soit $\mathscr{R}$-négligeable.

(ii) Si $K \subset \mathbf{X}$ est un compact, il est $\mathscr{R}$-négligeable.

(iii) Si $\mathbf{X}$ est compacte, elle est $\mathscr{R}$-négligeable.

Démonstration.

(i) Soit $\mathbf{a} = (a_1, \dots, a_n) \in \mathbf{X}$. On suppose que X est de dimension k au voisinage de $\mathbf{a}$. On peut supposer $k > 0$, sinon le résultat est évident.

Nous savons que, quitte à *réordonner* les coordonnées de $\mathbb{R}^n$, il existe :

- un voisinage ouvert $V_{\mathbf{a}}$ de $\mathbf{a}$ dans $\mathbb{R}^n$,
- un voisinage ouvert W de $(a_1, \dots, a_k)$ dans $\mathbb{R}^k$,
- une application, de classe $\mathcal{C}^1$: $\mathbf{f} : W \to \mathbb{R}^{n-k}$,

tels que :

$$\mathbf{x} = (x_1, \dots, x_n) \in \mathbf{X} \cap V_{\mathbf{a}} \Longleftrightarrow (x_1, \dots, x_k) \in W \text{ et } (x_{k+1}, \dots, x_n) = \mathbf{f}(x_1, \dots, x_k).$$

Soit W' un voisinage ouvert de $(a_1, \dots, a_k)$ dans $\mathbb{R}^k$ dont l'adhérence est contenue dans W. D'après la proposition 22, le graphe Γ de la restriction à $\overline{W'}$ de $\mathbf{f}$ est un sous-ensemble de $\mathbf{X}$ *$\mathscr{R}$-négligeable* dans $\mathbb{R}^n$.

Notons Γ' le graphe de la restriction à W' de $\mathbf{f}$ et $U_{\mathbf{a}} := V_{\mathbf{a}} \cap (W' \times \mathbb{R}^{n-k})$. On a $\Gamma' = \mathbf{X} \cap U_{\mathbf{a}}$, $U_{\mathbf{a}}$ est un voisinage ouvert de $\mathbf{a}$ dans $\mathbb{R}^n$ et $\Gamma' \subset \Gamma$ est $\mathscr{R}$-négligeable. Ainsi $U_{\mathbf{a}}$ répond à la question.

(ii) D'après (i) et la compacité de K, il existe une famille *finie* de points $(\mathbf{a}_i)_{i\in I}$ de $\mathbf{X}$ telle que les ouverts $(U_{\mathbf{a}_i})_{i\in I}$ recouvrent K. On a alors $K \subset \bigcup_{i\in I}(K \cap U_{\mathbf{a}_i})$. Les $K \cap U_{\mathbf{a}_i}$ sont $\mathscr{R}$-négligeables et I est fini, donc K est $\mathscr{R}$-négligeable.

(iii) C'est un cas particulier de (ii), avec $K := \mathbf{X}$. ■

Proposition 24. Soit $n \geqslant 2$.
L'image d'une courbe paramétrée M-lipchitzienne $\gamma : [0,1] \to \mathbb{R}^n$ est $\mathscr{R}$-négligeable.

En particulier l'image d'une courbe paramétrée de classe $\mathcal{C}^1$ est $\mathscr{R}$-négligeable.

Démonstration. On note $\Gamma = \gamma([0,1])$ l'image de γ. On partage $[0,1]$ en $m > 0$ segments égaux $J_1, \ldots, J_m$ de longueur $1/m$: $J_k := \left[\frac{k-1}{m}, \frac{k}{m}\right]$. Pour tout $k \in [\![1,m]\!]$ et tous $x, y \in J_k$, $\|\gamma(y) - \gamma(x)\| \leqslant M|y-x| \leqslant M/m$. Notons $\xi_k := k/m \in J_k$, les m pavés fermés $\mathbf{Q}_k$ de centres $\gamma(\xi_1), \ldots, \gamma(\xi_m)$ et de rayon M/m recouvrent Γ : $\gamma(J_k) \subset \mathbf{Q}_k$ donc $\Gamma = \bigcup_{k\in[\![1,m]\!]} \gamma(J_k) \subset \bigcup_{k\in[\![1,m]\!]} \mathbf{Q}_k$. Le volume de chaque $\mathbf{Q}_k$ est $\frac{(2M)^n}{m^n}$, le volume de $\bigcup_{k\in[\![1,m]\!]} \mathbf{Q}_k$ est donc majoré par $m\frac{(2M)^n}{m^n} = \frac{(2M)^n}{m^{n-1}} \leqslant \frac{(2M)^n}{m} \underset{m\to+\infty}{\longrightarrow} 0$, d'où le résultat. ■

On peut démontrer, plus généralement, que l'image dans $\mathbb{R}^n$ d'un cube de dimension strictement inférieure à n par une application lipchitzienne est $\mathscr{R}$-négligeable (*cf.* l'exercice III.1.2).
Nous admettons provisoirement le résultat suivant. La preuve fait l'objet de l'exercice 32 de la page 351.

Proposition 25.
Soit $C \subset \mathbb{R}^n$ un convexe compact. Alors la frontière de C est $\mathscr{R}$-négligeable.

Oscillation d'une fonction

Si $A \subset \mathbb{R}$ est une partie bornée, on note $\delta(A) := \sup_{x,y\in A} |y-x|$ son *diamètre*. Si $A \subset \mathbb{R}$ n'est pas bornée, on pose $\delta(A) := +\infty$.

Définition 10. Soient (E,d) un espace métrique et $f : E \to \mathbb{R}$. Pour tout $\mathbf{x} \in E$, on appelle *oscillation* de f en $\mathbf{x}$ et l'on note $\omega(f,\mathbf{x})$:

$$\omega(f,\mathbf{x}) := \inf_{r>0} \delta\big(f(B(\mathbf{x},r))\big),$$

$B(\mathbf{x},r)$ étant la boule ouverte de centre $\mathbf{x}$ et de rayon r dans E.

Proposition 26.
La fonction $f : (E,d) \to \mathbb{R}$ est *continue* en $x \in E$ si, et seulement si, $\omega(f,\mathbf{x}) = 0$.

Démonstration.

– Si f est continue en $\mathbf{x}$, pour tout $\varepsilon > 0$, il existe $r > 0$ tel que :

$$f\big(B(\mathbf{x}, r)\big) \subset \,]f(\mathbf{x}) - \varepsilon/2, f(\mathbf{x}) + \varepsilon/2[\,,$$

on a alors $\delta\big(f(B(\mathbf{x}, r)\big) \leqslant \varepsilon$. On en déduit que $\omega(f, \mathbf{x}) = 0$.

– Si $\omega(f, \mathbf{x}) = 0$, pour pour tout $\varepsilon > 0$, il existe $r > 0$ tel que $\delta\big(f(B(\mathbf{x}, r)\big) < \varepsilon/2$. On a alors $f\big(B(\mathbf{x}, r)\big) \subset \,]f(\mathbf{x}) - \varepsilon/2, f(\mathbf{x}) + \varepsilon/2[$. ■

Proposition 27. Soit $f : (E, d) \to \mathbb{R}$.
Pour tout $\varepsilon > 0$, $A_\varepsilon := \{\mathbf{x} \in E \mid \omega(f, \mathbf{x}) \geqslant \varepsilon\}$ est un *fermé*.

Démonstration. Soit $\mathbf{x} \in \bar{A}_\varepsilon$. Pour tout $r > 0$, $B(\mathbf{x}, r) \cap A_\varepsilon \neq \varnothing$. Soit $\mathbf{y} \in B(\mathbf{x}, r) \cap A_\varepsilon$; alors $\omega(f, \mathbf{y}) \geqslant \varepsilon$ et, puisque $B(\mathbf{x}, r)$ est un voisinage de $\mathbf{y}$, il existe $r' > 0$ tel que $B(\mathbf{y}, r') \subset B(\mathbf{x}, r)$. On a $\delta\big(f(B(\mathbf{y}, r')\big) \geqslant \varepsilon$, donc, *a fortiori*, $\delta\big(f(B(\mathbf{x}, r)\big) \geqslant \varepsilon$ et, r étant arbitraire, $\omega(f, \mathbf{x}) \geqslant \varepsilon$. Ainsi $\mathbf{x} \in A_\varepsilon$. ■

Lemme 28. Soient $\mathbf{P}$ un pavé fermé de $\mathbb{R}^n$ et $f : \mathbf{P} \to \mathbb{R}$ une fonction $\mathscr{R}$-intégrable. Alors, pour tout $\alpha > 0$:

$$A_\alpha := \{\mathbf{x} \in \mathbf{P} \mid \omega(f, \mathbf{x}) \geqslant \alpha\}$$

est négligeable.

Démonstration. Soit $\varepsilon > 0$. Il existe des fonctions en escalier φ, $\psi : \mathbf{P} \to \mathbb{R}$ telles que $\psi \leqslant f \leqslant \varphi$ et $\displaystyle\int_{\mathbf{P}} (\varphi - \psi) \leqslant \alpha\varepsilon/4$. Soit $\mathbf{u} := \{\mathbf{P}_k\}_{k \in [\![1, N]\!]}$ une subdivision de $\mathbf{P}$ adaptée à la fois à φ et ψ. Pour tout $k \in [\![1, N]\!]$, la fonction $\varphi - \psi$ est constante sur $\operatorname{Int} \mathbf{P}_k$, on note $c_k \in \mathbb{R}_+$ sa valeur.
Pour $k \in [\![1, N]\!]$, on a deux cas possibles, *s'excluant mutuellement* :

a) La fonction $\varphi - \psi \geqslant 0$ restreinte à $\operatorname{Int} \mathbf{P}_k$ (qui est constante) vérifie $\varphi - \psi \leqslant \alpha/2$.

b) La fonction $\varphi - \psi \geqslant 0$ restreinte à $\operatorname{Int} \mathbf{P}_k$ vérifie $\varphi - \psi > \alpha/2$.

Soit $I \subset [\![1, N]\!]$ (respectivement $J \subset [\![1, N]\!]$) l'ensemble des indices correspondant au cas a) (respectivement b)). On a $I \cap J = \varnothing$ et $I \cup J = [\![1, N]\!]$.
Montrons que si $k \in I$, $A_\alpha \cap \operatorname{Int} \mathbf{P}_k = \varnothing$.
Pour tous $\mathbf{x}, \mathbf{y} \in \operatorname{Int} \mathbf{P}_k$:

$$\psi(\mathbf{x}) \leqslant f(\mathbf{x}) \leqslant \varphi(\mathbf{x}) \quad \text{et} \quad \psi(\mathbf{y}) \leqslant f(\mathbf{y}) \leqslant \varphi(\mathbf{y}).$$

Puisque $\psi(\mathbf{x}) = \psi(\mathbf{y})$ et $\varphi(\mathbf{x}) = \varphi(\mathbf{y})$, on en déduit :

$$\big|f(\mathbf{y}) - f(\mathbf{x})\big| \leqslant \varphi(\mathbf{x}) - \psi(\mathbf{x}) = c_k \leqslant \alpha/2.$$

Donc, pour tout $\mathbf{z} \in \operatorname{Int} \mathbf{P}_k$, $\omega(f, \mathbf{z}) \leqslant \alpha/2$ ($\operatorname{Int} \mathbf{P}_k$ contient une boule ouverte de centre $\mathbf{z}$) et $\mathbf{z} \notin A_\alpha$. Ainsi :

$$\forall k \in I, \quad A_\alpha \cap \operatorname{Int} \mathbf{P}_k = \varnothing. \tag{5}$$

Notons $B := \mathbf{P} \setminus \bigcup_{k\in[\![1,N]\!]} \operatorname{Int} \mathbf{P}_k = \bigcup_{k\in[\![1,N]\!]} \partial\mathbf{P}_k$. On a, d'après la définition de I, J et (5) :

$$A_\alpha \subset \mathbf{P} \setminus \bigcup_{k\in I} \operatorname{Int} \mathbf{P}_k = B \cup \bigcup_{k\in J} \operatorname{Int} \mathbf{P}_k \subset B \cup \bigcup_{k\in J} \mathbf{P}_k.$$

On a, pour tout $k \in J$, $c_k > \alpha/2$, donc :

$$\frac{\alpha}{2} \sum_{k\in J} \operatorname{vol} \mathbf{P}_k \leqslant \sum_{k\in J} c_k \operatorname{vol} \mathbf{P}_k \leqslant \int_{\mathbf{P}} (\varphi - \psi) \leqslant \frac{\alpha\varepsilon}{4}$$

et $\sum_{k\in J} \operatorname{vol} \mathbf{P}_k \leqslant \varepsilon/2$ (si $J = \varnothing$, $\sum_{k\in J} \operatorname{vol} \mathbf{P}_k = 0 < \varepsilon/2$).

Par ailleurs, l'ensemble B est $\mathscr{R}$-négligeable (c'est une réunion finie de frontières de pavés), il peut donc être recouvert par une famille *finie* de pavés fermés dont la somme des volumes est inférieure à $\varepsilon/2$, donc l'ensemble $A_\alpha \subset B \cup \bigcup_{k\in J} \mathbf{P}_k$ peut être recouvert par une famille finie de pavés dont la somme des volumes est inférieure à $\varepsilon/2 + \varepsilon/2 = \varepsilon$. Par suite A_α est $\mathscr{R}$ négligeable et il est, *a fortiori*, négligeable. ■

Théorème de Lebesgue-Vitali

Le théorème suivant généralise le théorème de Lebesgue en dimension un énoncé dans le module intégration du livre [L2].

Théorème 29 (Lebesgue-Vitali). Soient P un pavé fermé de $\mathbb{R}^n$ et $f : P \to \mathbb{R}$. Les deux conditions suivantes sont équivalentes.

(i) La fonction f est $\mathscr{R}$-intégrable sur P.

(ii) La fonction f est *bornée* et l'ensemble de ses points de discontinuité est négligeable.

Autrement dit une fonction est intégrable au sens de Riemann sur $\mathbf{P}$ si, et seulement si, elle est bornée et presque partout continue sur $\mathbf{P}$.

Démonstration.

Montrons (i) $\Rightarrow$ (ii).

Soit $f : \mathbf{P} \to \mathbb{R}$ une fonction $\mathscr{R}$-intégrable.

Une fonction intégrable au sens de Riemann étant bornée, il reste à montrer que l'ensemble des points de discontinuité de f est négligeable. Notons S cet ensemble, on a, d'après la proposition 26, $S = \{x \in \mathbf{P} \mid \omega(f,x) > 0\}$. On a :

$$S = \{x \in \mathbf{P} \mid \omega(f,x) > 0\} = \bigcup_{n\in\mathbb{N}^*} \{x \in \mathbf{P} \mid \omega(f,x) \geqslant 1/n\} = \bigcup_{n\in\mathbb{N}^*} A_{1/n}.$$

D'après le lemme 28, les ensembles $A_{1/n}$, $n \in \mathbb{N}$, sont négligeables, donc S est négligeable.

Montrons (ii) $\Rightarrow$ (i). La preuve est un peu difficile et elle peut être omise en première lecture.

On suppose que la fonction f est *bornée* et que l'ensemble $S \subset \mathbf{P}$ de ses points de discontinuité est *négligeable*.

Soit $M > 0$ tel que $\big|f(x)\big| \leqslant M$ pour tout $x \in \mathbf{P}$. Soit $\varepsilon > 0$. Nous allons montrer qu'il existe deux fonctions en escalier φ et ψ sur $\mathbf{P}$ telles que :

$$\psi \leqslant f \leqslant \varphi \quad \text{et} \quad \int_{\mathbf{P}} (\varphi - \psi) \leqslant \varepsilon. \tag{6}$$

La $\mathscr{R}$-intégrabilité de f sur $\mathbf{P}$ en résultera.

Nous procéderons en deux temps. Nous admettrons d'abord le résultat $(\star)$ énoncé ci-dessous et en déduirons (6), puis nous prouverons $(\star)$.

$(\star)$ Il existe une subdivision $\{\mathbf{P}_i\}_{i \in [\![1,p]\!]}$ de $\mathbf{P}$ et une *partition* de $[\![1,p]\!]$: $[\![1,p]\!] = J' \cup J''$ telles que :

a) $\sum\limits_{i \in J'} \operatorname{vol} \mathbf{P_i} \leqslant \varepsilon/(4M)$;

b) pour tout $i \in J''$ et pour tous $y, z \in \operatorname{Int} P_i$, on a $\big|f(z) - f(y)\big| < \frac{\varepsilon}{4 \operatorname{vol} \mathbf{P}}$.

L'idée sous-jacente à $(\star)$ est que $\sum\limits_{i \in J'} \operatorname{vol} \mathbf{P_i}$ est « petit » tandis que la fonction f « oscille peu » sur les $\operatorname{Int} P_i$ quand $i \in J''$.

Montrons $(\star) \Rightarrow$ (6).

Pour tout $i \in J''$ on choisit un point $\xi_i \in \operatorname{Int} \mathbf{P}_i$. On définit deux fonctions en escalier φ et ψ sur $\mathbf{P}$ par :

$$\varphi : x \in \mathbf{P} \mapsto \begin{cases} f(\xi_i) + \dfrac{\varepsilon}{4 \operatorname{vol} \mathbf{P}} & \text{s'il existe } i \in J'' \text{ tel que } x \in \operatorname{Int} \mathbf{P_i} \\ M & \text{si } x \in \bigcup\limits_{i \in J'} \operatorname{Int} \mathbf{P_i} \\ f(x) & \text{si } x \in \mathbf{P} \setminus \bigcup\limits_{i \in [\![1,p]\!]} \operatorname{Int} \mathbf{P_i} \end{cases}$$

et :

$$\psi : x \in \mathbf{P} \mapsto \begin{cases} f(\xi_i) - \dfrac{\varepsilon}{4 \operatorname{vol} \mathbf{P}} & \text{s'il existe } i \in J'' \text{ tel que } x \in \operatorname{Int} \mathbf{P_i} \\ -M & \text{si } x \in \bigcup\limits_{i \in J'} \operatorname{Int} \mathbf{P_i} \\ f(x) & \text{si } x \in \mathbf{P} \setminus \bigcup\limits_{i \in [\![1,p]\!]} \operatorname{Int} \mathbf{P_i} \end{cases}$$

En utilisant a) et b), on vérifie que $\psi \leqslant f \leqslant \varphi$ et que :

$$\int_{\mathbf{P}} (\varphi - \psi) = \frac{\varepsilon}{2 \operatorname{vol} \mathbf{P}} \sum_{i \in J''} \operatorname{vol} P_i \; + 2M \sum_{i \in J'} \operatorname{vol} P_i \leqslant \frac{\varepsilon}{2 \operatorname{vol} \mathbf{P}} \operatorname{vol} \mathbf{P} + 2M \frac{\varepsilon}{4M} \leqslant \varepsilon.$$

Montrons l'assertion $(\star)$.

On note $C := \big\{x \in \mathbf{P} \;\big|\; \omega(f,x) \geqslant \frac{\varepsilon}{4 \operatorname{vol} \mathbf{P}}\big\}$, c'est un *fermé* de $\mathbf{P}$ (*cf.* la proposition 27), donc un *compact*.

Si $x \in C$, f est discontinue en x, donc $C \subset S$ et C est négligeable. Comme il est compact il est $\mathscr{R}$-négligeable, il existe donc une famille *finie* $\{\mathbf{U}_j\}_{j \in [\![1,q]\!]}$ de pavés ouverts de $\mathbb{R}^n$ recouvrant C telle que $\sum\limits_{j \in [\![1,q]\!]} \operatorname{vol} \mathbf{U}_j \leqslant \varepsilon/(4M)$. On pose $\mathbf{Q}_j := \overline{\mathbf{U}_j} \cap \mathbf{P}$. Puisque $C \subset \mathbf{P}$,

la famille $\{\mathbf{Q}_j\}_{j\in[\![1,q]\!]}$ est un recouvrement de C par des pavés fermés inclus dans $\mathbf{P}$ et l'on a $\sum\limits_{j\in[\![1,q]\!]} \operatorname{vol}\mathbf{Q}_j \leqslant \varepsilon/(4M)$.

Soit $D := \mathbf{P}\cap\Big(\mathbb{R}^n\setminus \bigcup\limits_{j\in[\![1,q]\!]}\mathbf{U}_j\Big)$, c'est un fermé contenu dans $\mathbf{P}$, donc un compact, et $D\cap C=\varnothing$.

Donc, pour tout $x\in D$, on a $\omega(f,x)<\frac{\varepsilon}{4\operatorname{vol}\mathbf{P}}$. D'après la définition de l'oscillation il existe un pavé ouvert $U(x)$ de centre x dont l'image par f a un diamètre strictement inférieur à $\frac{\varepsilon}{4\operatorname{vol}\mathbf{P}}$. En utilisant la compacité de D, on en déduit l'existence d'un nombre fini de points $\{x_k\in D\}_{k\in[\![1,s]\!]}$ tels que les pavés ouverts $U(x_k)$ recouvrent D et que, pour tout $k\in[\![1,s]\!]$, on ait $\delta\left(f\big(U(x_k)\big)\right)<\frac{\varepsilon}{4\operatorname{vol}\mathbf{P}}$. Notons $\mathbf{R}_k := \overline{U(x_k)}\cap\mathbf{P}$, c'est un pavé fermé contenu dans $\mathbf{P}$ et $\delta\left(f\left(\operatorname{Int}\mathbf{R}_k\right)\right)<\frac{\varepsilon}{4\operatorname{vol}\mathbf{P}}$. Puisque $D\subset\mathbf{P}$, les pavés fermés $\{\mathbf{R}_k\}_{k\in[\![1,s]\!]}$ recouvrent D.

On a $\mathbf{P}=D\cup\bigcup\limits_{j\in[\![1,q]\!]}\mathbf{Q}_j$, donc la réunion des deux familles $\{\mathbf{Q}_j\}_{j\in[\![1,q]\!]}$ et $\{\mathbf{R}_k\}_{k\in[\![1,s]\!]}$ est un recouvrement de $\mathbf{P}$ par des sous-pavés fermés. On projette les sommets de tous les $\{\mathbf{Q}_j\}_{j\in[\![1,q]\!]}$ et de tous les $\{\mathbf{R}_k\}_{k\in[\![1,s]\!]}$ sur les arêtes de $\mathbf{P}$. En ordonnant ces projections sur chaque axe de coordonnées on obtient des *subdivisions* de chaque arête. On considère la *subdivision produit* correspondante de $\mathbf{P}$ notée $\{\mathbf{P}_i\}_{i\in[\![1,p]\!]}$.

Soient $i\in[\![1,p]\!]$ et $\mathbf{x}\in\operatorname{Int}\mathbf{P}_i$, il existe un $\mathbf{Q}_j$ *ou* un $\mathbf{R}_k$ contenant $\mathbf{x}$. Si $\mathbf{x}\in\mathbf{Q}_j$, alors, par construction, $\operatorname{Int}\mathbf{P}_i\subset\operatorname{Int}\mathbf{Q}_j$ et $\mathbf{P}_i\subset\mathbf{Q}_j$. On raisonnerait de même si $\mathbf{x}\in\mathbf{R}_k$. On en déduit que chaque $\mathbf{P}_i$ est contenu dans un $\mathbf{Q}_j$ *ou* un $\mathbf{R}_k$.

Il existe une partition (unique) $[\![1,p]\!]=J'\cup J''$ telle que $i\in J''$, si, et seulement s'il n'existe aucun indice $j\in[\![1,q]\!]$ tel que $\mathbf{P}_i\subset\mathbf{Q}_j$.

- Si $i\in J'$, il existe un indice $j\in[\![1,q]\!]$ tel que $\mathbf{P_i}\subset\mathbf{Q}_j$. La réunion $\bigcup\limits_{i\in J'}\operatorname{Int}\mathbf{P_i}$ étant *disjointe*, on en déduit $\sum\limits_{i\in J'}\operatorname{vol}\mathbf{P_i}\leqslant\sum\limits_{j\in[\![1,q]\!]}\operatorname{vol}\mathbf{Q}_j\leqslant\varepsilon/(4M)$, c'est-à-dire a).
- Si $i\in J''$, il existe un indice $k\in[\![1,q]\!]$ tel que $\mathbf{P}_i\subset\mathbf{R}_k$ (et, par suite, $\operatorname{Int}\mathbf{P}_i\subset\operatorname{Int}\mathbf{R}_k$). On a $\delta\left(f\left(\operatorname{Int}\mathbf{R}_k\right)\right)<\frac{\varepsilon}{4\operatorname{vol}\mathbf{P}}$, donc $\delta\left(f\left(\operatorname{Int}\mathbf{P}_i\right)\right)<\frac{\varepsilon}{4\operatorname{vol}\mathbf{P}}$, c'est-à-dire b).

  Cela termine la preuve de l'assertion $(\star)$, donc celle de (ii) $\Rightarrow$ (i). ∎

1.2 Intégration sur un ensemble cubable

Pour les applications en géométrie, physique, mécanique…, il ne suffit pas de savoir intégrer sur un pavé, il faut étendre la théorie à des fonctions définies sur un domaine « convenable » (boule, polyèdre, intérieur d'ellipsoïde…). Nous nous intéresserons essentiellement au cas des fonctions *continues* sur certains ensembles compacts.

Si l'on veut pouvoir intégrer les fonctions continues sur un compact K, il faut déjà savoir intégrer la fonction 1 sur K, ce qui revient à savoir intégrer la fonction caractéristique 1_K sur $\mathbb{R}^n$. Cela conduit à la notion d'ensemble *cubable*. Nous définirons d'abord ces ensembles et étudierons leur propriétés. Pour appliquer ces résultats il faut évidemment savoir montrer qu'un ensemble K est cubable, nous donnerons un critère « assez facile à vérifier » portant sur la frontière ∂K (celle-ci doit être négligeable). Nous montrerons ensuite que les fonctions continues sur un compact K cubable, prolongées par 0 sur l'extérieur de K, sont $\mathscr{R}$-intégrables sur tout pavé fermé contenant K.

1.2.1 Parties cubables

Notre but est de mettre en évidence une famille de parties bornées de $\mathbb{R}^n$, dites *cubables*, pour lesquelles on dispose d'une notion « raisonnable » de *volume*.

Nous allons généraliser ce que nous avons fait pour la notion de *longueur* dans le cas de la dimension $n = 1$ (*cf.* la partie *Sous-ensembles élémentaires et longueurs* à la page 634 du module *Intégration* du livre [L2]) en utilisant des idées similaires avec quelques modifications techniques.

Comme en dimension 1 on va mettre en œuvre deux approches « équivalentes » :

- en utilisant la théorie de l'intégrale de Riemann sur un pavé appliquée aux fonctions caractéristiques ;
- en « approchant » une partie $A \subset \mathbb{R}^n$ par des parties « simples » dont on connaît le volume (dites *pavables*) contenues dans A ou contenant A dans l'esprit de ce que faisait Archimède en dimension $n = 2$ et $n = 3$.

Parties pavables

La famille des parties pavables est engendrée par celle des pavés fermés par réunion finie. Pour les éléments de cette famille il y a une façon simple de définir le volume.

Définition 11. Une partie $X \subset \mathbb{R}^n$ est dite pavable si elle est la réunion d'une famille *finie* de pavés fermés dont les intérieurs sont deux à deux disjoints. Une telle famille est dite *adaptée* à la partie pavable X.

Une partie pavable est une réunion finie de compacts, donc compacte. Si $n = 1$ les parties pavables sont les sous-ensembles élémentaires *fermés* introduits dans le module *Intégration* du livre [L2].

On vérifie qu'une partie est pavable si, et seulement si, elle est une réunion finie de pavés fermés en montrant que l'on peut redécouper les pavés d'une réunion finie pour obtenir une famille adaptée. On utilise pour ce faire deux remarques :

- l'intersection de deux pavés fermés est un pavé fermé ;
- si $\mathbf{P}$ et $\mathbf{P}'$ sont deux pavés fermés tels que $\mathbf{P}' \subset \mathbf{P}$, la partie $\mathbf{P} \setminus \operatorname{Int} \mathbf{P}'$ est pavable (*cf.* l'exercice 1 de la page 178).

Nous laissons les détails au lecteur (qui pourra procéder par récurrence sur le nombre de pavés d'une réunion finie).

La proposition suivante permet de définir le *volume* d'une partie pavable.

Proposition et définition 30. Soit $X \subset \mathbb{R}^n$ une partie pavable. Pour toute famille $(\mathbf{P}_i)_{i \in I}$ adaptée à X le nombre réel positif $\sum_{i \in I} \operatorname{vol} \mathbf{P}_i$ est indépendant du choix de la famille adaptée, on l'appelle *volume* de X et on le note $\operatorname{vol} X$. On convient de $\operatorname{vol} \varnothing := 0$.

Démonstration. On fait les deux remarques suivantes.

- Si l'on remplace un pavé $\mathbf{P}_i$ d'une famille adaptée à X par une subdivision on obtient

encore une famille adaptée à X et l'on ne change pas $\sum_{i\in I} \operatorname{vol} \mathbf{P}_i$. Par suite si l'on a deux familles adaptées $\mathbf{u} := (\mathbf{P}_i)_{i\in I}$ et $\mathbf{v} := (\mathbf{Q}_j)_{j\in J}$ de X telles que la seconde soit obtenue en redécoupant des pavés de la première, on a $\sum_{i\in I} \operatorname{vol} \mathbf{P}_i = \sum_{j\in J} \operatorname{vol} \mathbf{Q}_j$

– Soient deux familles $\mathbf{u} := (\mathbf{P}_i)_{i\in I}$ et $\mathbf{v} := (\mathbf{Q}_j)_{j\in J}$ adaptées à X. On considère la famille $\mathbf{w} = (\mathbf{P}_i \cap \mathbf{Q}_j)_{(i,j)\in I\times J}$, elle est adaptée à X et elle est obtenue en subdivisant d'une part les pavés de $\mathbf{u}$ et d'autre part les pavés de $\mathbf{v}$.

Le résultat s'en déduit. On a en effet, en utilisant la proposition 1 de la page 178 et l'associativité de l'addition :

$$\sum_{i\in I} \operatorname{vol} \mathbf{P}_i = \sum_{(i,j)\in I\times J} \operatorname{vol}(\mathbf{P}_i \cap \mathbf{Q}_j) \quad \text{et} \quad \sum_{j\in J} \operatorname{vol} \mathbf{Q}_j = \sum_{(i,j)\in I\times J} \operatorname{vol}(\mathbf{P}_i \cap \mathbf{Q}_j),$$

donc $\sum_{i\in I} \operatorname{vol} \mathbf{P_i} = \sum_{j\in J} \operatorname{vol} \mathbf{Q_j}$. ∎

Proposition 31. Si $X, Y \subset \mathbb{R}^n$ sont des parties pavables, $X \cup Y$, $X \cap Y$, $X \setminus \operatorname{Int} Y$ sont des parties pavables et l'on a :

$$\operatorname{vol}(X \cup Y) + \operatorname{vol}(X \cap Y) = \operatorname{vol} X + \operatorname{vol} Y. \tag{7}$$

Démonstration. Soient X et Y des parties pavables.

Elles sont réunions finies de pavés fermés : $X = \bigcup_{k\in[\![1,p]\!]} \mathbf{P}_k$, $Y = \bigcup_{h\in[\![1,q]\!]} \mathbf{Q}_h$ donc $X \cup Y$ l'est aussi et est pavable. On a $X \cap Y = \bigcup_{k\in[\![1,p]\!]} (\mathbf{P}_k \cap Y)$ et $\mathbf{P}_k \cap Y = \bigcup_{h\in[\![1,q]\!]} \mathbf{P}_k \cap \mathbf{Q}_h$.

Ainsi, pour tout $k \in [\![1,p]\!]$, $\mathbf{P}_k \cap Y$ est réunion finie de pavés donc pavable. Par suite, $X \cap Y$ est réunion finie de pavables donc pavable.

Les preuves du fait que $X \setminus \operatorname{Int} Y$ est pavable et de l'égalité (7) font l'objet de l'exercice III.1.1. ∎

Proposition 32. Soient $\mathbf{P}$ un pavé fermé et $A \subset \mathbf{P}$ une partie pavable.

Alors la fonction 1_A, restreinte à $\mathbf{P}$, est en escalier sur $\mathbf{P}$ et $\displaystyle\int_{\mathbf{P}} 1_A = \operatorname{vol} A$.

Démonstration. La partie $\mathbf{P} \setminus \operatorname{Int} A$ est pavable. On considère une subdivision adaptée $(\mathbf{R}_l)_{l\in L}$ de $\mathbf{P} \setminus \operatorname{Int} A$ et une subdivision adaptée $(\mathbf{Q}_j)_{j\in J}$ de A. On vérifie que :

$(\mathbf{P}_k)_{k\in I} := (\mathbf{Q}_j)_{j\in J} \cup (\mathbf{R}_l)_{l\in L}$ est une subdivision du pavé $\mathbf{P}$. Pour tout $l \in L$, la fonction 1_A est identiquement nulle sur $\operatorname{Int} \mathbf{R}_l$, et pour tout $j \in J$, la fonction 1_A est égale à 1 sur $\operatorname{Int} \mathbf{Q}_j$. Ainsi 1_A est une fonction en escalier et $\displaystyle\int_{\mathbf{P}} 1_A = \operatorname{vol} A$. ∎

Parties cubables

Soit $A \subset \mathbb{R}^n$ une partie *bornée*. Si $\mathbf{P}$ et $\mathbf{P}'$ sont des pavés fermés contenant A, on vérifie que la fonction caractéristique 1_A est $\mathscr{R}$-intégrable sur $\mathbf{P}$ si, et seulement si, elle l'est sur $\mathbf{P}'$ et que l'on a alors $\displaystyle\int_{\mathbf{P}} 1_A = \int_{\mathbf{P}'} 1_A$. Puisque l'intersection de deux pavés fermés est un pavé fermé, il suffit de le faire dans le cas où $\mathbf{P}' \subset \mathbf{P}$ (en utilisant $\mathbf{P} \cap \mathbf{P}' \subset \mathbf{P}$ et $\mathbf{P} \cap \mathbf{P}' \subset \mathbf{P}'$). Dans ce cas on utilise le résultat de la proposition 11 de la page 183.
On note $\displaystyle\int 1_A := \int_{\mathbf{P}} 1_A$.
Cela justifie la définition suivante.

Définition 12. Une partie *bornée* $A \subset \mathbb{R}^n$ est dite *cubable* si sa fonction caractéristique 1_A est $\mathscr{R}$-intégrable sur un pavé contenant A.
On dit alors que $\displaystyle\int 1_A$ est le *volume* de A ; on le note $\operatorname{vol} A$.

Les parties cubables sont aussi dites *de Peano Jordan* ou *mesurables au sens de Peano-Jordan*. Pour $n = 2$ on dit aussi *quarrable* au lieu de cubable.
Un pavé est évidemment cubable et son volume au sens ci-dessus est son volume usuel. Plus généralement, d'après la proposition 32 une partie pavable est cubable et la définition du volume donnée dans la définition ci-dessus est compatible avec celle donnée plus haut pour les parties pavables.
En utilisant les propriétés des fonctions caractéristiques on prouve la proposition suivante.

Proposition 33. Si $X, Y \subset \mathbb{R}^n$ sont des parties cubables, $X \cup Y$, $X \cap Y$ et $X \setminus \operatorname{Int} Y$ sont des parties cubables et l'on a :

$$\operatorname{vol}(X \cup Y) + \operatorname{vol}(X \cap Y) = \operatorname{vol} X + \operatorname{vol} Y.$$

Les réunions finies et intersections finies de parties cubables sont cubables.

Nous allons donner une caractérisation des parties *cubables* en utilisant les parties *pavables*. C'est une généralisation du corollaire 26 de la page 638 du livre [L2].
Soit $A \subset \mathbb{R}^n$ une partie *bornée non vide*. On désigne par $\lambda^+(A)$ la *borne inférieure* des volumes des parties pavables *contenant* A et par $\lambda^-(A)$ la *borne supérieure* des volumes des parties pavables *contenues* dans A. Notons que si A est d'intérieur vide, toute partie pavable contenue dans A est de volume nul, donc $\lambda^-(A) = 0$.
On convient de $\lambda^+(\varnothing) = \lambda^-(\varnothing) = 0$.

Proposition 34. Soit $X \subset \mathbb{R}^n$ une partie *bornée* non vide. On a :

$$\lambda^+(X) = \overset{*}{\int} 1_X \qquad \text{et} \qquad \lambda^-(X) = \underset{*}{\int} 1_X.$$

Démonstration. On choisit un pavé fermé $\mathbf{P}$ contenant X.

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 parties pavables, les fonctions 1_A et 1_B sont en escalier (*cf.* la proposition 32). On suppose $A \subset X \subset B \subset \mathbf{P}$.
De $1_A \leqslant 1_X \leqslant 1_B$ on déduit alors que :

$$\lambda^-(X) \leqslant {}_*\!\!\int_{\mathbf{P}} 1_X \leqslant \overset{*}{\int_{\mathbf{P}}} 1_X \leqslant \lambda^+(X). \tag{8}$$

Notons $d(1_X)$ la borne *supérieure* des sommes de Darboux *inférieures* de la fonction 1_X restreinte à $\mathbf{P}$. Montrons que toute somme de Darboux *inférieure* de 1_X est le volume d'une partie pavable *contenue* dans X.
Soit $S := \sum_{k \in [\![1,N]\!]} m_k \operatorname{vol} \mathbf{P}_k$ une somme de Darboux inférieure de 1_X sur $\mathbf{P}$. Par définition $m_k := \inf_{\mathbf{x} \in \mathbf{P}_k} 1_X(\mathbf{x})$, donc $m_k = 0$ ou $m_k = 1$. On a $m_k = 1$ si, et seulement si, $\mathbf{P}_k \subset A$, donc S est le volume d'une partie pavable contenue dans X (la réunion des $\mathbf{P}_k$ contenus dans X).
On en déduit $d(1_X) \leqslant \lambda^-(X)$, puis, en utilisant le corollaire 17 de la page 186 et (8) :

$${}_*\!\!\int_{\mathbf{P}} 1_X = d(1_X) \leqslant \lambda^-(X) \leqslant {}_*\!\!\int_{\mathbf{P}} 1_X,$$

donc $\lambda^-(X) = {}_*\!\!\int_{\mathbf{P}} 1_X$.
On prouverait de même que toute somme de Darboux supérieure de 1_X est le volume d'une partie pavable contenant X, d'où, en utilisant le corollaire 17 de la page 186 et (8) :

$$\overset{*}{\int_{\mathbf{P}}} 1_X \leqslant \lambda^+(X) \leqslant D(1_X) = \overset{*}{\int_{\mathbf{P}}} 1_X,$$

donc $\lambda^+(X) = \overset{*}{\int_{\mathbf{P}}} 1_X$. ∎

Corollaire 35. Soit $A \subset \mathbb{R}^n$ une partie bornée non vide. Les deux conditions suivantes sont équivalentes :

(i) A est cubable ;

(ii) $\lambda^+(A) = \lambda^-(A)$.

Si elles sont satisfaites, on a $\int 1_A = \lambda^+(A) = \lambda^-(A)$.

Remarque. L'assertion (ii) du corollaire ci-dessus fournit une définition du « volume » très voisine de celle utilisée par Archimède dans les cas $n = 2$ (aire) ou 3. Si l'on a une « notion raisonnable de volume », on peut encadrer le volume *inconnu* d'une partie X par le volume *connu* d'une partie A contenue dans X et le volume *connu* d'une partie B contenant X (si $n = 2$, A et B sont par exemple des polygones) (*cf.* la figure III.1.1). Si l'on peut obtenir $\operatorname{vol} B - \operatorname{vol} A$ « aussi petit que l'on veut », alors on obtient le « volume » de X (chez Archimède par la méthode d'Eudoxe, *cf.* le premier *Point histoire* du module *Intégration* du livre [L2]).

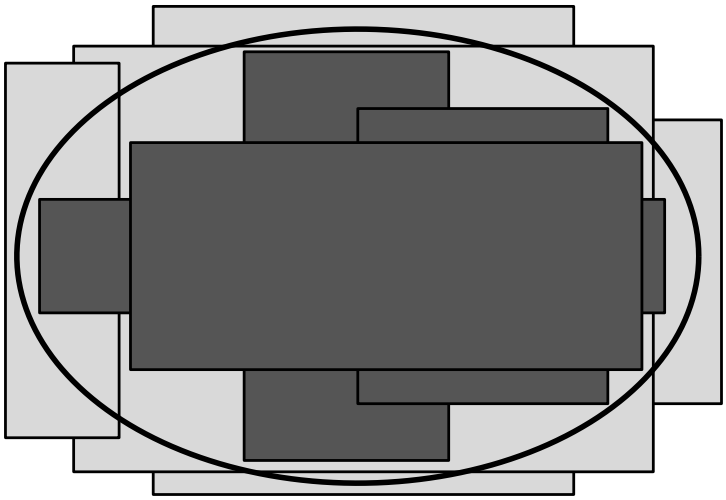


FIGURE III.1.1 – Partie quarrable

On en déduit facilement le résultat suivant.

Proposition 36. Soit $A \subset \mathbb{R}^n$ une partie bornée non vide. Les conditions suivantes sont équivalentes :

(i) A est $\mathscr{R}$-négligeable ;

(ii) $\lambda^+(A) = 0$;

(iii) A est cubable et $\displaystyle\int 1_A = \mathrm{vol}(A) = 0$.

Proposition 37.
Soient $\mathbf{P} \subset \mathbb{R}^n$ un pavé fermé et $f : \mathbf{P} \to \mathbb{R}_+$ une fonction $\mathscr{R}$-intégrable. On définit $A := \big\{(\mathbf{x}, y) \in \mathbf{P} \times \mathbb{R}_+ \;\big|\; 0 \leqslant y \leqslant f(\mathbf{x})\big\}$ et l'on note $\Gamma \subset \mathbf{P} \times \mathbb{R}_+$ le graphe de f. Alors A est cubable et Γ est $\mathscr{R}$-négligeable. De plus $\mathrm{vol}\, A = \displaystyle\int_{\mathbf{P}} f$.

Démonstration. Soit $\mathbf{u} := \{\mathbf{P}_k\}_{k \in [\![1,N]\!]}$ une subdivision de $\mathbf{P}$. On note :

$$\forall k \in [\![1, N]\!], \quad m_k := \inf_{\mathbf{x} \in \mathbf{P}_k} f(\mathbf{x}) \quad \text{et} \quad M_k = \sup_{\mathbf{x} \in \mathbf{P}_k} f(\mathbf{x}).$$

On définit $A^-(\mathbf{u}) := \displaystyle\bigcup_{k \in [\![1,N]\!]} (\mathbf{P}_k \times [0, m_k])$ et $A^+(\mathbf{u}) := \displaystyle\bigcup_{k \in [\![1,N]\!]} (\mathbf{P}_k \times [0, M_k])$. Les parties $A^-(\mathbf{u})$ et $A^+(\mathbf{u})$ de $\mathbb{R}^{n+1}$ sont *pavables* et l'on a $A^-(\mathbf{u}) \subset A \subset A^+(\mathbf{u})$. On a :

$$\lambda^+(A) - \lambda^-(A) \leqslant \mathrm{vol}\, A^+(\mathbf{u}) - \mathrm{vol}\, A^-(\mathbf{u}) = \sum_{k \in [\![1,N]\!]} (M_k - m_k)\, \mathrm{vol}\, \mathbf{P}_k.$$

On en déduit que A est cubable et que $\mathrm{vol}\, A = \displaystyle\int_{\mathbf{P}} f$ en utilisant le critère de Darboux (*cf.* le corollaire 18 de la page 186) et le corollaire 35.

Par ailleurs Γ est contenu dans l'ensemble pavable $\bigcup_{k\in[\![1,N]\!]} (\mathbf{P}_k \times [m_k, M_k])$ dont le volume est $\sum_{k\in[\![1,N]\!]} (M_k - m_k) \operatorname{vol} \mathbf{P}_k$. On en déduit qu'il est $\mathscr{R}$-négligeable en utilisant le critère de Darboux. ■

Autrement dit l'intégrale d'une fonction positive f $\mathscr{R}$-intégrable sur un pavé représente le « volume sous le graphe de f » et le graphe de f est $\mathscr{R}$-négligeable (on comparera à la proposition 22 de la page 188).

Exercice 4.
Montrer que si X est une partie cubable, son adhérence $\overline{X}$ est cubable.
Solution. Soit $X \subset \mathbb{R}^n$ une partie bornée. Si $A \subset X$, $A \subset \bar{X}$, donc $\lambda^-(X) \leqslant \lambda^-(\overline{X})$.
Si B est pavable il est fermé, donc si $X \subset B$, $\overline{X} \subset B$ et l'on en déduit $\lambda^+(X) = \lambda^+(\overline{X})$.
Donc $\lambda^-(X) \leqslant \lambda^-(\overline{X}) \leqslant \lambda^+(\overline{X}) = \lambda^+(X)$.
Si X est cubable $\lambda^+(X) = \lambda^-(X)$, donc $\lambda^+(\overline{X}) = \lambda^-(\overline{X})$ et $\overline{X}$ est cubable.

Exemple. Si $\overline{X}$ est cubable, X n'est pas nécessairement cubable. Voici un exemple.
Soit $n := 2$. On note $\Sigma := [0,1] \cap \mathbb{Q}$ et $X := \Sigma^2 \subset \mathbf{P} := [0,1]^2$. L'adhérence de X est $\mathbf{P}$, elle est donc cubable. On a $\lambda^+(X) = \lambda^+(\overline{X}) = 1$, mais X est d'intérieur vide (le vérifier) donc $\lambda^-(X) = 0 \neq \lambda^+(X)$. Ainsi X n'est pas cubable.

Proposition 38. Soit A une partie *bornée* de $\mathbb{R}^n$.
Les conditions suivantes sont équivalentes :

(i) A est cubable ;

(ii) ∂A est négligeable ;

(iii) ∂A est $\mathscr{R}$-négligeable ;

(iv) ∂A est cubable et de volume nul.

Démonstration. Soit $A \subset \mathbb{R}^n$ une partie bornée non vide, alors $\partial A \neq \varnothing$.

On a (ii) $\Leftrightarrow$ (iii) : ∂A est compact et l'on utilise la proposition 21 de la page 188.

On a (iii) $\Leftrightarrow$ (iv) d'après la proposition 36.

Montrons (i) $\Leftrightarrow$ (ii). D'après le théorème de Lebesgue-Vitali, il suffit de prouver que ∂A est l'ensemble des points de discontinuité de 1_A.

Les restrictions de 1_A aux intérieurs de A et $\mathbb{R}^n \setminus A$ sont constantes (respectivement égales à 1 ou 0) donc continues, ainsi 1_A est continue sur $\mathbb{R}^n \setminus \partial A$. Soit $\mathbf{a} \in \partial A$. Tout voisinage de $\mathbf{a}$ rencontre A et $\mathbb{R}^n \setminus A$, donc sur un tel voisinage la fonction 1_A prend les valeurs 1 et 0, donc elle n'est pas continue en $\mathbf{a}$. ■

On déduit facilement le résultat suivant des deux propositions ci-dessus.

Corollaire 39. Soit A une partie de $\mathbb{R}^n$. Les conditions suivantes sont équivalentes :

(i) A est $\mathscr{R}$-négligeable

(ii) $\bar{A}$ est $\mathscr{R}$-négligeable

(iii) A est bornée et $\bar{A}$ est négligeable ;

(iv) A est cubable et négligeable

(v) A est cubable et de volume nul.

En utilisant les propositions 23 et 25 de la page 191, on déduit de la proposition 38 le résultat suivant.

Corollaire 40.

(i) Une partie compacte de $\mathbb{R}^n$ dont la frontière est une variété lisse de classe $\mathcal{C}^1$ est cubable.

(ii) Une partie relativement compacte de $\mathbb{R}^n$ dont la frontière est une réunion finie de parties compactes chacune contenue dans une variété lisse de classe $\mathcal{C}^1$, de dimension strictement inférieure à n, est cubable.

(iii) Une partie convexe bornée de $\mathbb{R}^n$ est cubable.

Dans l'exercice suivant on donne des exemples de compacts non cubables en toute dimension. Nous en donnerons d'autres « plus compliqués » dans l'exercice III.1.45 en utilisant la théorie de Henstock-Kurzweil.

Exercice 5.

Soit $\mathbf{Q} \subset \mathbb{R}^q$ un pavé fermé non vide.

1. Soit $A \subset \mathbb{R}^p$ un compact. On suppose que $A \times \mathbf{Q} \subset \mathbb{R}^{p+q}$ est négligeable dans $\mathbb{R}^{p+q}$, montrer que A est négligeable dans $\mathbb{R}^p$.
2. Soient $K \subset [a,b]$ un compact non négligeable d'intérieur vide (il en existe, par exemple le compact de Smith-Volterra-Cantor, *cf.* l'exercice II.3.6 du livre [L2]). Montrer que $K \times \mathbf{Q} \subset \mathbb{R}^{n+1}$ n'est pas cubable dans $\mathbb{R}^{n+1}$.

Solution.

1. Moyennant un raisonnement par récurrence sur q, on peut supposer $q = 1$ et $\mathbf{Q} = [c,d]$. Pour simplifier les notations, nous supposerons $p = 1$, le lecteur modifiera facilement la preuve dans le cas général. Par hypothèse $A \times [c,d]$ est négligeable. Par suite, puisqu'il est compact, il est $\mathscr{R}$-négligeable.
   Soient $a, b \in \mathbb{R}$, $a < b$, tels que $A \subset [a,b]$. On note $\mathbf{P} := [a,b] \times [c,d]$. Soit $\varepsilon > 0$. Il existe une partie pavable C telle que $\operatorname{vol} C < \varepsilon(d-c)$ et $A \times \mathbf{P} \subset C$. Quitte à remplacer C par $C \cap \mathbf{P}$, on peut supposer $C \subset \mathbf{P}$. Soit $(\mathbf{P}_j)_{j \in J}$ (J fini) une famille de pavés fermés adaptée à la partie pavable C. On peut *compléter* cette famille en une subdivision $\mathbf{w} := (\mathbf{P}_i)_{i \in I}$ de $\mathbf{P}$, avec $J \subset I$. Il existe une subdivision produit $\mathbf{w}'$ de $\mathbf{P}$ plus fine que $\mathbf{w}$, elle induit une subdivision de chaque $\mathbf{P}_j$, $j \in J$, et la réunion de ces subdivisions est une famille adaptée à C. Autrement dit, on peut supposer que $\mathbf{w}$ *est une subdivision produit*. Elle est obtenue en faisant le produit d'une subdivision $\mathbf{u} := \{x_0, \ldots, x_r\}$ de $[a,b]$ et d'une subdivision $\mathbf{v} := \{y_0, \ldots, y_s\}$ de $[c,d]$.

On a $\operatorname{vol} C = \sum_{j\in J} \operatorname{vol} \mathbf{P}_j < \varepsilon(d-c)$. Les projections des $\mathbf{P}_j$, $j \in J$, recouvrent A, ce sont des segments $[x_h, x_{h+1}]$, on note $T \subset [\![1, r]\!]$ l'ensemble des indices h correspondants. Quitte à enlever un *nombre fini* de points à A (ce qui ne change rien au problème), on peut supposer que $(]x_h, x_{h+1}[)_{h\in T}$ est un recouvrement de A. Soient $h \in T$ et $x \in]x_h, x_{h+1}[$, les $\mathbf{P}_j$, $j \in J$, qui se projettent sur $[x_h, x_{h+1}]$ doivent recouvrir $\{x\} \times [c, d]$, ce sont donc *tous* les $([x_h, x_{h+1}] \times [y_k, x_{k+1}])_{k\in[\![1,s-1]\!]}$. Notons $J_h \subset J$ l'ensemble des indices correspondants, on a $\sum_{j\in J_h} \operatorname{vol} \mathbf{P}_j = (x_{h+1} - x_h) \times (d-c)$. Les $\{J_h\}_{h\in T}$ forment une *partition* de J, donc :

$$\operatorname{vol} C = \sum_{j\in J} \operatorname{vol} \mathbf{P}_j = \sum_{h\in T} \sum_{j\in J_h} \operatorname{vol} \mathbf{P}_j = (d-c) \times \sum_{h\in T} (x_{h+1} - x_h) < \varepsilon(d-c).$$

On en déduit $\sum_{h\in T} (x_{h+1} - x_h) < \varepsilon$. Ce qui prouve que A est $\mathscr{R}$-négligeable donc négligeable.

2. D'après 1., le compact $K \times \mathbf{Q} \subset \mathbb{R}^{n+1}$ n'est pas négligeable, sinon K le serait. Puisque K est d'intérieur vide, $K \times \mathbf{Q}$ est d'intérieur vide (l'image d'un pavé ouvert de $\mathbb{R}^{n+1}$ par la projection $\mathbb{R}^{n+1} \to \mathbb{R}$ est un intervalle ouvert). Ainsi $\partial(K \times \mathbf{Q}) = K \times \mathbf{Q}$ n'est pas négligeable et $K \times \mathbf{Q}$ n'est pas cubable.

L'image d'une partie négligeable par une application continue n'est pas en général négligeable. Par exemple $\mathbb{R} \times \{0\}$ est une partie négligeable de $\mathbb{R}^2$ mais son image par la projection $(x, y) \in \mathbb{R}^2 \mapsto x \in \mathbb{R}$ est $\mathbb{R}$ qui n'est pas une partie négligeable de $\mathbb{R}$.

Point Méthode

Pour les calculs d'intégrales multiples, nous aurons très souvent à vérifier que certaines parties sont cubables. Nous n'entrerons pas dans les détails, le lecteur fera chaque fois les vérifications nécessaires en utilisant le corollaire 40. Le plus souvent, la partie considérée est définie par une famille finie d'inéquations $(f_i \geqslant 0)_{i\in I}$, où les f_i sont des fonctions de classe $\mathcal{C}^1$, et l'on peut utiliser le théorème des fonctions implicites et le critère (ii) du corollaire 40. On n'oubliera pas d'utiliser la convexité quand c'est possible.

Exemple. Soient a, b, $c > 0$. L'ellipsoïde « plein » :

$$E := \left\{ (x, y, z) \in \mathbb{R}^3 \mid \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \leqslant 1 \right\}$$

est cubable.

Pour le montrer, le plus simple est de vérifier que E est borné et convexe. On peut aussi appliquer le théorème des fonctions implicites à la fonction $f(x, y, z) := \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2}$ ($0 \notin \partial E$, donc en un point de la frontière ∂E l'une au moins des dérivées partielles $\partial f/\partial x = 2x/a^2$, $\partial f/\partial y = 2y/b^2$, $\partial f/\partial z = 2z/c^2$ ne s'annule pas) et montrer que ∂E est une sous-variété lisse de dimension 1.

Nous verrons plus loin que, pour intégrer une fonction sur un ensemble cubable, on emploie souvent une formule de changement de variables (*cf.* le théorème 58 de la page 215). On peut alors utiliser le résultat suivant.

Proposition 41.
Soient $U \subset \mathbb{R}^n$ un ouvert et $\mathbf{h} : U \to \mathbb{R}^m$ une application de classe $\mathcal{C}^1$.

(i) Soit $A \subset U$ une partie négligeable telle que $\bar{A} \subset U$, alors $\mathbf{h}(A)$ est négligeable.

(ii) On suppose que $m = n$ et que $\mathbf{h}$ induit un difféomorphisme de classe $\mathcal{C}^1$ de U sur un ouvert $V \subset \mathbb{R}^n$. Soit $B \subset \mathbb{R}^n$ une partie cubable telle que $\bar{B} \subset U$, alors $\mathbf{h}(B)$ est une partie cubable.

Démonstration.

(i) La preuve de cette assertion est l'objet de l'exercice III.1.4.

(ii) L'application $\mathbf{h}$ induit un homéomorphisme de U sur V donc une bijection de ∂B sur $\partial(\mathbf{h}(B))$. Puisque B est cubable, la partie ∂B est négligeable. En utilisant (i), on en déduit que $\mathbf{h}(\partial B) = \partial(\mathbf{h}(B))$ est négligeable, donc $\mathbf{h}(B)$ est cubable. ∎

1.2.2 Intégration des fonctions sur un ensemble cubable

Soient $\mathbf{P} \subset \mathbb{R}^n$ un pavé fermé et $A \subset \mathbf{P}$ une partie cubable. Si $f : \mathbf{P} \to \mathbb{R}$ est une fonction $\mathscr{R}$-intégrable sur $\mathbf{P}$, $f_{|_A} = 1_A \times f$ est $\mathscr{R}$-intégrable sur $\mathbf{P}$ (un produit de fonctions $\mathscr{R}$-intégrables sur un pavé fermé est $\mathscr{R}$-intégrable) et l'on peut *définir* $\displaystyle\int_A f$ par $\displaystyle\int_A f := \int_{\mathbf{P}} f_{|_A}$. Nous allons généraliser cette situation (l'idée de la définition suivante remonte à Dirichlet 1839[2]).

Proposition et définition 42. Soient $A \subset \mathbb{R}^n$ une partie bornée et $f : A \to \mathbb{R}$. On prolonge f en une fonction $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$, appelée *prolongement canonique de* f définie par : $\tilde{f}(\mathbf{x}) := \begin{cases} f(\mathbf{x}), & \text{si} \quad \mathbf{x} \in A \\ 0 & \text{si} \quad \mathbf{x} \in \mathbb{R}^n \setminus A. \end{cases}$

Soit $\mathbf{P}$ *un* pavé fermé contenant A. Si $\tilde{f}$ est $\mathscr{R}$-intégrable sur $\mathbf{P}$, elle est alors $\mathscr{R}$-intégrable sur *tout* pavé fermé contenant A et l'intégrale $\displaystyle\int_{\mathbf{P}} \tilde{f}$ est *indépendante* du choix de $\mathbf{P}$, on l'appelle intégrale de f sur A et on la note $\displaystyle\int_A f$. Par abus de langage on dit que f est $\mathscr{R}$-intégrable sur A.

Démonstration. Il suffit de montrer le résultat pour deux pavés $A \subset \mathbf{P} \subset \mathbf{P}'$ et, dans ce cas, on utilise le résultat de la proposition 11 de la page 183. ∎

[2] *Werke*, vol. I, p. 377

On a $\widetilde{|f|} = |\tilde{f}|$, donc si f est $\mathscr{R}$-intégrable sur A, il en est de même de $|f|$.
Le théorème de Lebesgue-Vitali (*cf.* le théorème 29 de la page 193) permet de donner une caractérisation des fonctions intégrables sur une partie bornée.

Proposition 43. Soient $A \subset \mathbb{R}^n$ une partie bornée et $f : A \to \mathbb{R}$. Pour que f soit $\mathscr{R}$-intégrable sur A il faut et il suffit que les deux conditions suivantes soient réalisées :

(i) f est bornée sur A ;

(ii) l'ensemble des points de discontinuité de $\tilde{f}$ (dans $\mathbb{R}^n$) est négligeable.

Démonstration. Il existe un pavé fermé $\mathbf{P}$ tel que $\bar{A} \subset \operatorname{Int} \mathbf{P}$. La fonction $\tilde{f}$ est continue sur l'ouvert $\mathbb{R}^n \setminus \bar{A}$ (elle est nulle sur cet ouvert), l'ensemble de ses points de discontinuité dans $\mathbb{R}^n$ est donc le même que l'ensemble des points de discontinuité de sa restriction à $\mathbf{P}$. On termine la preuve en utilisant le théorème de Lebesgue-Vitali sur $\mathbf{P}$. ■

La proposition suivante permet d'intégrer une fonction *continue bornée* sur un ensemble *cubable*.

Proposition 44.
Toute fonction *continue bornée* sur un ensemble *cubable* A est $\mathscr{R}$-intégrable sur A.

Démonstration. Soit $f : A \to \mathbb{R}$ une application continue bornée. Si A est cubable, ∂A est négligeable. La fonction $\tilde{f}$ est continue sur les ouverts $\operatorname{Int} A$ et $\mathbb{R}^n \setminus \bar{A}$. Son ensemble de points de discontinuité S est donc contenu dans ∂A. Par suite, si ∂A est négligeable, S l'est aussi. On utilise alors la proposition précédente. ■

Exemple. *Masse, densité, moments d'inertie.* Soit $A \subset \mathbb{R}^3$ un compact cubable non négligeable. Une *masse volumique* sur A est une application *continue* $m : A \to \mathbb{R}_+$.
L'intégrale $m(A) := \iiint_A m(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ est appelée *masse* de A. On dit que $\frac{m(A)}{\operatorname{vol} A}$ est la *densité moyenne* de A. On note $\mathrm{d}m := m\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ (distribution de masses). Si Δ est une droite affine de $\mathbb{R}^3$ muni de sa structure euclidienne, la fonction :
$(x,y,z) \mapsto d\big((x,y,z),\Delta\big)^2 m(x,y,z)$ est continue, donc intégrable sur A. L'intégrale : $\iiint_A d\big((x,y,z),\Delta\big)^2\,\mathrm{d}m$ est appelée *moment d'inertie* de A par rapport à l'axe Δ.

Proposition 45. Si $A \subset \mathbb{R}^n$ est cubable et $f : A \to \mathbb{R}$ $\mathscr{R}$-intégrable sur A, alors :

$$\left|\int_A f\right| \leqslant \operatorname{vol} A \sup_{\mathbf{x}\in A} \big|f(\mathbf{x})\big|.$$

Démonstration. Soit $\mathbf{P}$ un pavé fermé contenant A. La fonction $\tilde{f}$ est $\mathscr{R}$-intégrable sur $\mathbf{P}$, donc $|\tilde{f}|$ est $\mathscr{R}$-intégrable sur $\mathbf{P}$ et l'on a $\left|\int_{\mathbf{P}} \tilde{f}\right| \leqslant \int_{\mathbf{P}} |\tilde{f}|$. Compte tenu de $|\tilde{f}| = \widetilde{|f|}$, cette inégalité s'écrit $\left|\int_A f\right| \leqslant \int_A |f|$.

La fonction $|f|$ étant $\mathscr{R}$-intégrable sur $\mathbf{P}$ est bornée sur $\mathbf{P}$, donc $M := \sup_{\mathbf{x}\in A}\big|f(\mathbf{x})\big| < +\infty$ et $\displaystyle\int_A |f| \leqslant \int_A M = M \operatorname{vol} A$. Par suite $\displaystyle\left|\int_A f\right| \leqslant M \operatorname{vol} A$. ■

Proposition 46. Toute fonction *continue bornée* sur un ensemble $\mathscr{R}$*-négligeable* A est intégrable et d'intégrale *nulle* sur A.

Démonstration. Soit $f : A \to \mathbb{R}$ une fonction continue bornée sur l'ensemble négligeable A. D'après le corollaire 39 de la page 202, la partie A est bornée et cubable de volume nul. Elle est donc d'intérieur vide, et $\bar{A} = \partial A$ est $\mathscr{R}$-négligeable.
On choisit un pavé fermé $\mathbf{P}$ contenant la partie bornée A. L'ensemble des points de discontinuité de $\tilde{f}$ est contenu dans ∂A, il est donc négligeable, donc, d'après le théorème 29 de la page 193, $\tilde{f}$ est $\mathscr{R}$-intégrable sur $\mathbf{P}$.
On montre $\displaystyle\int_A f = 0$ en utilisant la proposition 45. ■

Proposition 47. Soient A et B deux parties bornées de $\mathbb{R}^n$ et $f : A \to \mathbb{R}$.
On suppose que f est $\mathscr{R}$-intégrable sur A, que $B \subset A$ et que B est cubable.
Alors f est $\mathscr{R}$-intégrable sur B.

Démonstration. Soient $\mathbf{P}$ un pavé fermé contenant A et $\tilde{f}$ le prolongement canonique de f à $\mathbf{P}$. Puisque B est cubable, 1_B est $\mathscr{R}$-intégrable sur $\mathbf{P}$. La fonction $1_B\tilde{f}$ est le produit de deux fonctions $\mathscr{R}$-intégrables sur $\mathbf{P}$, elle est donc $\mathscr{R}$-intégrables sur $\mathbf{P}$, d'où le résultat (la restriction de f à B est égale à celle de $1_B\tilde{f}$). ■

On déduit des deux propositions précédentes la généralisation suivante de la relation de Chasles en dimension $n = 1$.

Proposition 48. Soient A et B deux parties cubables de $\mathbb{R}^n$ telles que $A\cap B$ soit $\mathscr{R}$-négligeable et $f : A \cup B \to \mathbb{R}$. Alors f est $\mathscr{R}$-intégrable sur A et B, si, et seulement si, elle l'est sur $A \cup B$. S'il en est ainsi, on a $\displaystyle\int_{A\cup B} f = \int_A f + \int_B f$.

Démonstration. Si f est $\mathscr{R}$-intégrable sur $A \cup B$, d'après la proposition 47, elle est intégrable sur A et B. On a $1_{A\cup B} = 1_A + 1_B - 1_{A\cap B}$, d'où $\displaystyle\int 1_{A\cup B} f = \int 1_A f + \int 1_B f$, en utilisant $\displaystyle\int 1_{A\cap B} f = 0$ (*cf.* la proposition 46). Ainsi $\displaystyle\int_{A\cup B} f = \int_A f + \int_B f$.
Inversement, si f est $\mathscr{R}$-intégrable sur A et B, alors $1_{A\cup B} f = 1_A f + 1_B f - 1_{A\cap B} f$.
D'après la proposition 46, la fonction $1_{A\cap B} f$ est intégrable et $\displaystyle\int 1_{A\cap B} f = 0$.
Donc $1_{A\cup B} f = 1_A f + 1_B f$ est intégrable et $\displaystyle\int_{A\cup B} f = \int_A f + \int_B f$. ■

De la proposition ci-dessus, on déduit, par récurrence, le résultat suivant.

Corollaire 49. Soit $(A_i)_{i\in I}$ une famille finie de parties cubables de $\mathbb{R}^n$ telles que les intersections deux à deux $A_i \cap A_j$, $i \neq j$, soient $\mathscr{R}$-négligeables et $f : \bigcup_{i\in I} A_i \to \mathbb{R}$. Alors f est $\mathscr{R}$-intégrable sur chacun des A_i, $i \in I$, si, et seulement si, elle l'est sur $\bigcup_{i\in I} A_i$. S'il en est ainsi, on a $\displaystyle\int_{\bigcup_{i\in I} A_i} f = \sum_{i\in I} \int_{A_i} f$.

Proposition 50. Soient $A \subset \mathbb{R}^n$ une partie cubable et $f : \bar{A} \to \mathbb{R}$ une fonction *bornée*. Alors f est $\mathscr{R}$-intégrable sur A si, et seulement si, elle l'est sur $\bar{A}$. s'il en est ainsi, on a $\displaystyle\int_A f = \int_{\bar{A}} f$.

Démonstration. La frontière ∂A de la partie cubable A est négligeable, donc la fonction f est $\mathscr{R}$-intégrable sur la partie cubable ∂A et l'on a $\displaystyle\int_{\partial A} f = 0$. La partie $A \cap \partial A$ est négligeable, donc, d'après la proposition 48, f est $\mathscr{R}$-intégrable sur A si, et seulement si, elle l'est sur $\bar{A} = A \cup \partial A$ et l'on a $\displaystyle\int_{\bar{A}} f = \int_A f + \int_{\partial A} f = \int_A f$. ■

Point Méthode

Soit $A \subset \mathbb{R}^n$ une partie cubable *ouverte*. On peut remplacer le calcul de l'intégrale $\displaystyle\int_{\bar{A}} f$ d'une fonction f continue sur $\bar{A}$ (qui est cubable) par celui de $\displaystyle\int_A f$ (qui lui est égale). On a alors la possibilité de calculer cette dernière en utilisant une *formule de changement de variables* (*cf.* la proposition 59 de la page 216).

1.3 Intégrales itérées et théorème de Fubini

Dans ce qui précède, nous avons *défini* la notion d'intégrale multiple, mais *nous ne savons pas calculer* les intégrales multiples, même dans des cas très simples. Pour ce faire, la principale méthode est de se ramener à des calculs *d'intégrales simples* par *un processus d'itération*. Les théorèmes justifiant cette méthode s'appellent théorèmes de Fubini. Il y a différents énoncés ; nous nous contenterons de quelques versions simples.

Pour le calcul des intégrales sur un domaine A de $\mathbb{R}^n$ par itération ($n \geqslant 2$), il est préférable d'utiliser la notation $\displaystyle\iint \ldots \int_A f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \ldots \mathrm{d}x_n$ au lieu de $\displaystyle\int_A f$ (le nombre de signes d'intégration $\displaystyle\int$ étant la dimension n de l'espace).

1.3.1 Le cas des fonctions continues sur un pavé

Intégrales doubles itérées des fonctions continues sur un rectangle

Théorème 51 (Théorème de Fubini). Soient $\mathbf{P} := [a,b] \times [c,d]$ et $f : \mathbf{P} \to \mathbb{R}$ une fonction *continue*.

(i) Pour tout $\mathbf{x} \in [a,b]$ la fonction $\mathbf{y} \mapsto f(\mathbf{x},\mathbf{y})$ est continue, donc $\mathscr{R}$-intégrable, sur $[c,d]$ et la fonction $\mathbf{x} \mapsto \displaystyle\int_c^d f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}$ est continue, donc $\mathscr{R}$-intégrable, sur $[a,b]$

(ii) Pour tout $\mathbf{y} \in [c,d]$ la fonction $\mathbf{x} \mapsto f(\mathbf{x},\mathbf{y})$ est continue, donc $\mathscr{R}$-intégrable, sur $[a,b]$ et la fonction $\mathbf{y} \mapsto \displaystyle\int_a^b f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}$ est continue, donc $\mathscr{R}$-intégrable, sur $[c,d]$.

(iii) On a :

$$\iint_{\mathbf{P}} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} = \int_a^b \mathrm{d}\mathbf{x} \int_c^d f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_c^d \mathrm{d}\mathbf{y} \int_a^b f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}. \tag{9}$$

Démonstration.

(i) Pour tout $\mathbf{x} \in [a,b]$ la fonction $\mathbf{y} \mapsto f(\mathbf{x},\mathbf{y})$ est évidemment continue et la fonction $\mathbf{x} \mapsto \displaystyle\int_c^d f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}$ est continue d'après le corollaire 127 de la page 717 (module *Intégration*) du livre [L2].

(ii) On raisonne comme dans (i).

(iii) Les trois membres des égalités (9) existent, il reste à prouver qu'ils sont égaux. On sait déjà que :

$$\int_a^b \mathrm{d}\mathbf{x} \int_c^d f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_c^d \mathrm{d}\mathbf{y} \int_a^b f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}$$

(c'est le théorème de Fubini démontré dans le module *Intégration* du livre [L2]).

Montrons : $\displaystyle\iint_{\mathbf{P}} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = \int_a^b \mathrm{d}\mathbf{x} \int_c^d f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}$.

On considère des subdivisions $\mathbf{u} := (x_i)_{i\in[\![0,p]\!]}$ et $\mathbf{v} := (y_i)_{j\in[\![0,q]\!]}$ respectivement de $[a,b]$ et de $[c,d]$ et la subdivision produit $\mathbf{w}$ de $\mathbf{P}$ correspondante :

$$\mathbf{w} := \big(\mathbf{P}_{ij}\big)_{i\in[\![0,p-1]\!],j\in[\![0,q-1]\!]}, \quad \text{avec} \quad \mathbf{P}_{ij} := [x_i, x_{i+1}] \times [y_j, y_{j+1}]\,.$$

Notons, pour tout $x \in [a,b]$, $F(\mathbf{x}) := \displaystyle\int_c^d f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}$.

On a, d'après la relation de Chasles $F(\mathbf{x}) = \displaystyle\sum_{j\in[\![0,q-1]\!]} \int_{y_j}^{y_{j+1}} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}$, puis :

$$\int_a^b F(\mathbf{x})\,\mathrm{d}\mathbf{x} = \sum_{i\in[\![0,p-1]\!]}\ \sum_{j\in[\![0,q-1]\!]} \int_{x_i}^{x_{i+1}} \mathrm{d}\mathbf{x} \int_{y_j}^{y_{j+1}} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y},$$

d'où, en utilisant les notations des sommes de Darboux (m_{ij} et M_{ij} désignant respectivement les bornes inférieures et supérieures de f sur $\mathbf{P}_{ij}$) :

$$\sum_{(i,j)\in[\![0,p-1]\!]\times[\![0,q-1]\!]} m_{ij}\,\mathrm{vol}(\mathbf{P}_{ij}) \leqslant \int_a^b F(\mathbf{x})\,\mathrm{d}\mathbf{x} \leqslant \sum_{(i,j)\in[\![0,p-1]\!]\times[\![0,q-1]\!]} M_{ij}\,\mathrm{vol}(\mathbf{P}_{ij}),$$

c'est-à-dire $d_{\mathbf{w}}(f) \leqslant \displaystyle\int_a^b F(\mathbf{x})\,\mathrm{d}\mathbf{x} \leqslant D_{\mathbf{w}}(f)$.

On en déduit $d(f) \leqslant \displaystyle\int_a^b F(\mathbf{x})\,\mathrm{d}\mathbf{x} \leqslant D(f)$, puis, en utilisant la proposition 19 de la page 187, $d(f) = D(f) = \displaystyle\iint_{\mathbf{P}} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} = \int_a^b F(\mathbf{x})\,\mathrm{d}\mathbf{x}$. ■

Point Méthode

Le cas particulier suivant du théorème de Fubini est très utile.
Si, sous les hypothèses du théorème 51, la fonction f est le produit d'une fonction g de $\mathbf{x}$ seul et d'une fonction h de $\mathbf{y}$ seul, on a :

$$\iint_{\mathbf{P}} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\mathrm{d}\,\mathbf{y} = \left(\int_a^b g(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)\left(\int_c^d h(\mathbf{y})\,\mathrm{d}\mathbf{y}\right). \tag{10}$$

Calcul des Intégrales multiples des fonctions continues sur un pavé

Le résultat suivant permet par récurrence de ramener le calcul de l'intégrale d'une fonction continue sur un pavé à des calculs successifs d'intégrales simples (intégrales itérées). La preuve est similaire à celle du théorème 51.

Théorème 52. Soient $n_1, n_2 \in \mathbb{N}^*$ et $\mathbf{P}_1 \subset \mathbb{R}^{n_1}$ et $\mathbf{P}_2 \subset \mathbb{R}^{n_2}$ des pavés fermés. On note $n := n_1 + n_2$ et $\mathbf{P} := \mathbf{P}_1 \times \mathbf{P}_2 \subset \mathbb{R}^n$. Soit $f : \mathbf{P} \to \mathbb{R}$ une fonction *continue*.

(i) Pour tout $x \in \mathbf{P}_1$ la fonction $y \mapsto f(x,y)$ est continue, donc $\mathscr{R}$-intégrable sur $\mathbf{P}_2$ et la fonction $x \mapsto \displaystyle\int_{\mathbf{P}_2} f(x,y)\,\mathrm{d}y$ est continue, donc $\mathscr{R}$-intégrable sur $\mathbf{P}_1$

(ii) Pour tout $y \in \mathbf{P}_2$ la fonction $x \mapsto f(x,y)$ est continue, donc $\mathscr{R}$-intégrable sur $\mathbf{P}_1$ et la fonction $y \mapsto \displaystyle\int_{\mathbf{P}_1} f(x,y)\,\mathrm{d}x$ est continue, donc $\mathscr{R}$-intégrable sur $\mathbf{P}_2$.

(iii) On a :

$$\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbf{P}_1} \mathrm{d}x \int_{\mathbf{P}_2} f(x,y)\,\mathrm{d}y = \int_{\mathbf{P}_2} \mathrm{d}y \int_{\mathbf{P}_1} f(x,y)\,\mathrm{d}x. \tag{11}$$

1.3.2 Le cas des fonctions $\mathscr{R}$-intégrables sur un pavé

On souhaiterait étendre le théorème 52 au cas des fonctions continues sur un domaine cubable. L'idée naturelle est de l'étendre d'abord au cas des fonctions $\mathscr{R}$-intégrables sur un pavé, malheureusement on rencontre certains obstacles qui nécessitent d'introduire des conditions restrictives. Nous allons d'abord étudier le cas de la dimension 2.

Intégrales doubles itérées des fonctions $\mathscr{R}$-intégrables

Commençons par signaler trois problèmes qui apparaissent quand on essaie de calculer par itération l'intégrale double d'une fonction sur un pavé $\mathbf{P}$. Ces problèmes montrent que l'intégrale de Riemann est peu adaptée au calcul par intégrales itérées et qu'il vaut mieux restreindre la classe des fonctions à intégrer (en se limitant aux fonctions continues) ou l'élargir en passant à l'intégrale de Lebesgue (*cf.* ci-dessous la section 4).

1. Il peut arriver que f soit $\mathscr{R}$-intégrable sur $\mathbf{P}$ mais que les intégrales partielles $\displaystyle\int_c^d f(x,y)\,\mathrm{d}y$ (respectivement $\displaystyle\int_a^b f(x,y)\,\mathrm{d}x$) ne soient pas définies pour toute valeur de $x \in [a,b]$ (respectivement $y \in [c,d]$).
2. Il peut arriver que l'une des intégrales itérées existe mais pas l'autre.
3. Il peut arriver que les deux intégrales itérées de (9) existent (et nous verrons plus loin qu'alors elles sont égales : *cf.* la proposition 55) sans que f soit $\mathscr{R}$-intégrable sur $\mathbf{P}$.

On trouvera des exemples illustrant ces situations dans les exercices III.1.5 à III.1.9.
La proposition suivante montre que l'on peut *toujours* calculer l'intégrale d'une fonction $\mathscr{R}$-intégrable sur un pavé par intégrales itérées en remplaçant convenablement dans l'énoncé du théorème de Fubini *intégrale* par *intégrale inférieure* ou *intégrale supérieure*.

Proposition 53. Soient $\mathbf{P} := [a,b]\times[c,d]$ et $f : \mathbf{P} \to \mathbb{R}$ une fonction $\mathscr{R}$-*intégrable*.

(i) Pour tout $x \in [a,b]$, les fonctions $x \mapsto \displaystyle{}^{*}\!\!\int_c^d f(x,y)\,\mathrm{d}y$ et $x \mapsto \displaystyle\int_{*c}^d f(x,y)\,\mathrm{d}y$ sont $\mathscr{R}$-intégrables sur $[a,b]$.

(ii) Pour tout $y \in [c,d]$, les fonctions $y \mapsto \displaystyle{}^{*}\!\!\int_a^b f(x,y)\,\mathrm{d}x$ et $y \mapsto \displaystyle\int_{*a}^b f(x,y)\,\mathrm{d}x$ sont $\mathscr{R}$-intégrables sur $[c,d]$.

(iii) On a :

$$\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \mathrm{d}x \,{}^{*}\!\!\int_c^d f(x,y)\,\mathrm{d}y = \int_c^d \mathrm{d}y \,{}^{*}\!\!\int_a^b f(x,y)\,\mathrm{d}x$$
$$= \int_a^b \mathrm{d}x \int_{*c}^d f(x,y)\,\mathrm{d}y = \int_c^d \mathrm{d}y \int_{*a}^b f(x,y)\,\mathrm{d}x.$$

Démonstration. La fonction f est bornée : $m \leqslant f \leqslant M$, donc, pour tout $x \in [a,b]$, $\displaystyle\int_{*c}^d f(x,y)\,\mathrm{d}y$ existe, on la note $F(x) := \displaystyle\int_{*c}^d f(x,y)\,\mathrm{d}y$. On a $m(d-c) \leqslant F \leqslant M(d-c)$,

ainsi F est bornée. Nous allons montrer que la fonction $x \in [a,b] \to F(x) \in \mathbb{R}$ est $\mathscr{R}$-intégrable sur $[a,b]$ et que $\displaystyle\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b F(x)\,\mathrm{d}x = \int_a^b \mathrm{d}x \,{}_*\!\!\int_c^d f(x,y)\,\mathrm{d}y$ en utilisant des sommes de Darboux.

Soient $\mathbf{u} := \{x_i\}_{i\in[\![0,p]\!]}$ et $\mathbf{v} := \{y_j\}_{j\in[\![0,q]\!]}$ des subdivisions respectives de $[a,b]$ et $[c,d]$. La relation de Chasles est vraie pour l'intégrale inférieure (*cf.* la proposition 18 de la page 633 du livre [L2]), donc $\displaystyle{}_*\!\!\int_c^d f(x,y)\,\mathrm{d}y = \sum_{j\in[\![0,q-1]\!]} {}_*\!\!\int_{y_j}^{y_{j+1}} f(x,y)\,\mathrm{d}y$. D'après les propositions 9 et 18 de ce même module *Intégration* du livre [L2], on a :

$$\begin{aligned}
{}_*\!\!\int_a^b F(x)\,\mathrm{d}x &\geqslant \sum_{j\in[\![0,q-1]\!]} {}_*\!\!\int_a^b \mathrm{d}x \,{}_*\!\!\int_{y_j}^{y_{j+1}} f(x,y)\,\mathrm{d}y \\
&\geqslant \sum_{i\in[\![0,p-1]\!]} \sum_{j\in[\![0,q-1]\!]} {}_*\!\!\int_{x_i}^{x_{i+1}} \mathrm{d}x \,{}_*\!\!\int_{y_j}^{y_{j+1}} f(x,y)\,\mathrm{d}y.
\end{aligned}$$

Avec les notations des sommes de Darboux de f pour la subdivision produit $\mathbf{s} := \mathbf{u} \times \mathbf{v}$ de $\mathbf{P} = [a,b] \times [c,d]$, on a $f(x,y) \geqslant m_{i,j}$ pour tous $x \in [x_i, x_{i+1}]$ et $y \in [y_j, y_{j+1}]$, $i \in [\![0,p-1]\!]$, $j \in [\![0,q-1]\!]$. On en déduit $\displaystyle{}_*\!\!\int_a^b F(x)\,\mathrm{d}x \geqslant d_{\mathbf{s}}(f)$. On montrerait de façon analogue, en utilisant les mêmes propositions 9 et 18 du module *Intégration* du livre [L2] :

$$\begin{aligned}
{}^*\!\!\int_a^b F(x)\,\mathrm{d}x &\leqslant \sum_{j\in[\![0,q-1]\!]} {}^*\!\!\int_a^b \mathrm{d}x \,{}_*\!\!\int_{y_j}^{y_{j+1}} f(x,y)\,\mathrm{d}y \\
&\leqslant \sum_{i\in[\![0,p-1]\!]} \sum_{j\in[\![0,q-1]\!]} {}^*\!\!\int_{x_i}^{x_{i+1}} \mathrm{d}x \,{}_*\!\!\int_{y_j}^{y_{j+1}} f(x,y)\,\mathrm{d}y,
\end{aligned}$$

puis $\displaystyle{}^*\!\!\int_a^b F(x) \leqslant D_{\mathbf{s}}(f)$. Ainsi :

$$d_{\mathbf{s}}(f) \leqslant {}_*\!\!\int_a^b F(x) \leqslant {}^*\!\!\int_a^b F(x) \leqslant D_{\mathbf{s}}(f).$$

Soit $\varepsilon > 0$; la fonction f étant $\mathscr{R}$-intégrable sur $\mathbf{P}$, on peut choisir $\mathbf{u}$ et $\mathbf{v}$ de telle sorte que $D_{\mathbf{s}}(f) - d_{\mathbf{s}}(f) \leqslant \varepsilon$. On en déduit $\displaystyle 0 \leqslant {}^*\!\!\int_a^b F(x) - {}_*\!\!\int_a^b F(x) \leqslant \varepsilon$.

Par suite F est $\mathscr{R}$-intégrable sur $[a,b]$. On a, de plus :

$$d_{\mathbf{s}}(f) \leqslant \iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \leqslant D_{\mathbf{s}}(f).$$

On en déduit $\displaystyle\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \mathrm{d}x \,{}_*\!\!\int_c^d f(x,y)\,\mathrm{d}y$.

On prouverait de façon similaire les autres résultats d'intégrabilité énoncés dans (i) et (ii) et les autres égalités :

$$\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \mathrm{d}x \,{}_*\!\!\int_c^d f(x,y)\,\mathrm{d}y = \int_c^d \mathrm{d}y \,{}_*\!\!\int_a^b f(x,y)\,\mathrm{d}x$$
$$= \int_c^d \mathrm{d}y \,{}^*\!\!\int_a^b f(x,y)\,\mathrm{d}x.$$

Corollaire 54. Soient $\mathbf{P} := [a,b] \times [c,d]$ et $f : \mathbf{P} \to \mathbb{R}$ une fonction $\mathscr{R}$-intégrable. On *suppose* que :

(i) pour tout $x \in [a,b]$, les fonctions $y \mapsto f(x,y)$ sont $\mathscr{R}$-intégrables sur $[c,d]$;

(ii) pour tout $y \in [c,d]$, les fonctions $x \mapsto f(x,y)$ sont $\mathscr{R}$-intégrables sur $[a,b]$.

Alors :

- la fonction $x \mapsto \displaystyle\int_c^d f(x,y)\,\mathrm{d}y$ est $\mathscr{R}$-intégrable sur $[a,b]$,
- la fonction $y \mapsto \displaystyle\int_a^b f(x,y)\,\mathrm{d}x$ est $\mathscr{R}$-intégrable sur $[c,d]$ et :

$$\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \mathrm{d}x \int_c^d f(x,y)\,\mathrm{d}y = \int_c^d \mathrm{d}y \int_a^b f(x,y)\,\mathrm{d}x.$$

La preuve du résultat suivant (dû à André Warusfel) est l'objet de l'exercice III.1.11.

Proposition 55. Soient $\mathbf{P} := [a,b] \times [c,d]$ et $f : \mathbf{P} \to \mathbb{R}$ une fonction bornée. On *suppose* que :

(i) pour tout $x \in [a,b]$, les fonctions $y \mapsto f(x,y)$ sont $\mathscr{R}$-intégrables sur $[c,d]$;

(ii) pour tout $y \in [c,d]$, les fonctions $x \mapsto f(x,y)$ sont $\mathscr{R}$-intégrables sur $[a,b]$.

Alors :

- la fonction $x \mapsto \displaystyle\int_c^d f(x,y)\,\mathrm{d}y$ est $\mathscr{R}$-intégrable sur $[a,b]$,
- la fonction $y \mapsto \displaystyle\int_a^b f(x,y)\,\mathrm{d}x$ est $\mathscr{R}$-intégrable sur $[c,d]$ et :

$$\int_a^b \mathrm{d}x \int_c^d f(x,y)\,\mathrm{d}y = \int_c^d \mathrm{d}y \int_a^b f(x,y)\,\mathrm{d}x.$$

Le résultat suivant montre que l'on peut calculer par intégrales itérées l'intégrale double d'une fonction continue sur un ensemble cubable « convenable ». C'est encore une version du théorème de Fubini.

Théorème 56. Soient $\mathbf{P} := [a,b] \times [c,d]$, $A \subset \mathbf{P}$ un ensemble cubable et $f : \mathbf{P} \to \mathbb{R}$ une fonction *continue bornée* sur A et nulle sur $\mathbf{P} \setminus A$. On suppose que, pour tout $x \in [a,b]$, $A_x := \{y \in [c,d] \mid (x,y) \in A\}$ est une partie cubable de $[c,d]$. On a alors :

$$\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \mathrm{d}x \int_c^d f(x,y)\,\mathrm{d}y = \int_a^b \mathrm{d}x \int_{A_x} f(x,y)\,\mathrm{d}y.$$

On a un résultat analogue en échangeant les rôles de x et y.

Démonstration. Pour tout $x \in [a,b]$, la fonction $y \mapsto f(x,y)$ est $\mathscr{R}$-intégrable sur $[c,d]$. En effet la fonction $y \mapsto f(x,y)$ est continue et bornée sur A_x et A_x est cubable par hypothèse. Par suite $\displaystyle\mathop{\int}_{*}{}_c^d f(x,y)\,\mathrm{d}y = \int_c^d f(x,y)\,\mathrm{d}y$.

Par ailleurs f est intégrable sur $\mathbf{P}$, donc, d'après la proposition 53 :

$$\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \mathrm{d}x \mathop{\int}_{*}{}_c^d f(x,y)\,\mathrm{d}y = \int_a^b \mathrm{d}x \int_c^d f(x,y)\,\mathrm{d}y. \qquad \blacksquare$$

Exemple. *Ouverts verticalement ou horizontalement simples*

On rappelle qu'un ouvert $\Omega \subset \mathbb{R}^2$ est dit *verticalement simple* (*cf.* la définition 22 de la page 827 du module *Fonctions de plusieurs variables* du livre [L2]) si :

$$\Omega = \big\{(x,y) \in \mathbb{R}^2 \mid a < x < b,\ \varphi_1(x) < y < \varphi_2(x)\big\},$$

où $a < b$ et où φ_1 et φ_2 sont des fonctions de classe $\mathcal{C}^1$ par morceaux vérifiant $\varphi_1 < \varphi_2$ sur $]a,b[$. On définit de façon analogue la notion d'ouvert *horizontalement simple* en échangeant les rôles de x et y.

D'après la proposition 24 de la page 191, un ouvert verticalement ou horizontalement simple est cubable.

On peut calculer l'intégrale double d'une fonction continue bornée f sur un ouvert verticalement simple en utilisant le théorème 56. En effet, pour tout $x \in]a,b[$, l'ensemble $\Omega_x := \{y \in \mathbb{R} \mid (x,y) \in A\}$ est l'intervalle ouvert $\varphi_1(x) < y < \varphi_2(x)$, c'est donc une partie cubable de $\mathbb{R}$. On a $\displaystyle\iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_a^b \mathrm{d}x \int_{\varphi_1(x)}^{\varphi_2(x)} f(x,y)\,\mathrm{d}y$.

On a un résultat similaire pour les ouverts horizontalement simples.

Exercice 6.

Soit $\Delta := \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leqslant 1,\ y \geqslant 0\} \subset \mathbb{R}^2$. Calculer $\displaystyle\iint_\Delta y\,\mathrm{d}x\mathrm{d}y$.

Solution. D'après le théorème 56 ci-dessus :

$$\iint_\Delta y\,\mathrm{d}x\mathrm{d}y = \int_{-1}^1 \mathrm{d}x \int_0^{\sqrt{1-x^2}} y\,\mathrm{d}y = \frac{1}{2}\int_{-1}^1 \big[y^2\big]_0^{\sqrt{1-x^2}}\,\mathrm{d}x = \frac{1}{2}\int_{-1}^1 (1-x^2)\,\mathrm{d}x$$
$$= \frac{1}{2}\left(2 - \frac{2}{3}\right) = \frac{2}{3}.$$

Réduction des intégrales multiples des fonctions $\mathscr{R}$-intégrables

Le résultat suivant généralise le théorème 56. Il permet de calculer par récurrence l'intégrale multiple d'une fonction continue sur un ensemble cubable vérifiant certaines conditions. Nous admettrons ce résultat : la preuve est similaire à celle du cas $n = 2$ et repose sur la généralisation (facile) du théorème 53. Nous laissons les détails au lecteur.

Théorème 57. Soient $n_1, n_2 \in \mathbb{N}^*$, $\mathbf{P}_1 \subset \mathbb{R}^{n_1}$ et $\mathbf{P}_2 \subset \mathbb{R}^{n_2}$ des pavés fermés. On note $n := n_1 + n_2$ et $\mathbf{P} := \mathbf{P}_1 \times \mathbf{P}_2 \subset \mathbb{R}^n$. Soient $A \subset \mathbf{P}$ un ensemble cubable et $f : \mathbf{P} \to \mathbb{R}$ une fonction continue sur A et nulle sur $\mathbf{P} \setminus A$. On suppose que, pour tout $x \in \mathbf{P}_1$, $A_x = A \cap \big(\{x\} \times \mathbb{R}^{n_2}\big)$ est un ensemble cubable « de » $\mathbb{R}^{n_2}$. On a alors :

$$\iint_{\mathbf{P}} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbf{P}_1} \mathrm{d}x \int_{\mathbf{P}_2} f(x,y)\,\mathrm{d}y = \int_{\mathbf{P}_1} \mathrm{d}x \int_{A_x} f(x,y)\,\mathrm{d}y \qquad (12)$$

Formule de Green-Riemann

Munis de nouveaux outils, nous revenons sur la formule de Green-Riemann traitée dans le livre [L2]. Nous reprenons les définitions et notations données dans la partie du module *Fonctions de plusieurs variables* traitant de cette question, afin de préciser certains points.
On considère un ouvert borné Ω à frontière $\mathcal{C}^1$ par morceaux. D'après la proposition 24 de la page 191, $\partial\Omega$ est négligeable, donc Ω est cubable.
Soient $U \subset \mathbb{R}^2$ un ouvert contenant $\overline{\Omega}$ et $P,\ Q : U \to \mathbb{R}$ deux fonctions de classe $\mathcal{C}^1$.
Alors la fonction $(x,y) \mapsto \frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y)$ est *continue bornée* donc $\mathscr{R}$-intégrable sur Ω. On note $\omega := P\,\mathrm{d}x + Q\,\mathrm{d}y$, $\mathrm{d}\omega := \big(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\big)\,\mathrm{d}x\,\mathrm{d}y$.
On a $\displaystyle\int_\Omega \mathrm{d}\omega = \int_\Omega \Big(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y)\Big)\,\mathrm{d}x\,\mathrm{d}y$. On suppose que le bord de Ω est orienté dans le sens positif, alors, la formule de Green-Riemann est $\displaystyle\int_\Omega \mathrm{d}\omega = \int_{\partial\Omega} \omega$ (*cf.* [MPA1]).
Nous allons la prouver dans le cas particulier d'un ouvert verticalement *et* horizontalement simple (*cf.* l'exemple 1.3.2 de la page précédente).
Soit :

$$\begin{aligned}\Omega &= \big\{(x,y) \in \mathbb{R}^2 \mid a < x < b,\ \varphi_1(x) < y < \varphi_2(x)\big\}\\ &= \big\{(x,y) \in \mathbb{R}^2 \mid c < y < d,\ \psi_1(y) < x < \psi_2(y)\big\},\end{aligned}$$

avec $\varphi_1,\ \varphi_2, \psi_1, \psi_2$ de classe $\mathcal{C}^1$. Soient $U \subset \mathbb{R}^2$ un ouvert contenant $\overline{\Omega}$ et $P,\ Q : U \to \mathbb{R}$ deux fonctions de classe $\mathcal{C}^1$, on a (*cf.* les lemmes 34 et 35 du module *Fonctions de plusieurs variables* du livre [L2]) :

$$\int_{\partial\Omega} P(x,y)\,\mathrm{d}x = \int_a^b \Big(P\big(x,\varphi_1(x)\big) - P\big(x,\varphi_2(x)\big)\Big)\,\mathrm{d}x$$

et :

$$\int_{\partial\Omega} Q(x,y)\,\mathrm{d}y = \int_c^d \Big(-Q\big(y,\psi_1(y)\big) + Q\big(y,\psi_2(y)\big)\Big)\,\mathrm{d}y.$$

D'après le théorème 56 (*cf.* l'exemple de la page 213) :

$$\begin{aligned}\int_\Omega \frac{\partial P}{\partial y}(x,y)\big)\,\mathrm{d}x\,\mathrm{d}y &= \int_a^b \mathrm{d}x \int_{\varphi_1(x)}^{\varphi_2(x)} \frac{\partial P}{\partial y}(x,y)\big)\,\mathrm{d}y \\ &= \int_a^b \Big(P\big(x,\varphi_2(x)\big) - P\big(x,\varphi_1(x)\big)\Big)\,\mathrm{d}x = -\int_{\partial\Omega} P(x,y)\,\mathrm{d}x.\end{aligned}$$

On prouve de façon similaire $\displaystyle\int_\Omega \frac{\partial Q}{\partial y}(x,y)\big)\,\mathrm{d}x\,\mathrm{d}y = \int_{\partial\Omega} Q(x,y)\,\mathrm{d}y$, d'où :

$$\int_\Omega \mathrm{d}\omega = \int_\Omega \Big(\frac{\partial Q}{\partial x}(x,y) - \frac{\partial P}{\partial y}(x,y)\Big)\,\mathrm{d}x\,\mathrm{d}y = \int_{\partial\Omega} \big(P(x,y)\,\mathrm{d}x + Q(x,y)\,\mathrm{d}y\big) = \int_{\partial\Omega} \omega.$$

1.4 Formule de changement de variables

Nous conclurons cette section par la formule de changement de variables pour les intégrales multiples dans un cadre restreint mais suffisant pour beaucoup d'applications. La preuve de la formule de changement de variables (en réalité d'une version plus générale de cette formule) sera donnée plus loin. Cette preuve utilise la formule de Fubini et, de ce fait, il y a une version « plus satisfaisante » dans le cadre de l'intégrale de Henstock-Kurzweil. Nous reviendrons sur cette question dans la partie 4.4.

1.4.1 Formule de changement de variables

On rappelle qu'une application $\mathbf{h} : U_1 \to U_2$ est un difféomorphisme de classe $\mathcal{C}^1$ entre deux ouverts U_1 et U_2 de $\mathbb{R}^n$ si $\mathbf{h}$ est bijective, et si $\mathbf{h}$ et $\mathbf{h}^{-1}$ sont de classe $\mathcal{C}^1$. En particulier, le déterminant jacobien :

$$\operatorname{Jac}\mathbf{h}(\mathbf{x}) := \det\big(D_{\mathbf{x}}\mathbf{h}\big) = \det\left(\frac{\partial h_i}{\partial x_j}\right)_{1\leqslant i,j\leqslant n}$$

est partout non nul.

Nous donnerons plus loin la preuve d'une version plus générale du résultat suivant que nous admettons provisoirement (*cf.* le théorème 189 de la page 353).

Théorème 58. Soient U_1 et U_2 deux ouverts *bornés* de $\mathbb{R}^n$ et $\mathbf{h} : U_1 \to U_2$ un difféomorphisme de classe $\mathcal{C}^1$ de jacobien $\operatorname{Jac}\mathbf{h}$. Soit $f : U_2 \to \mathbb{R}$ une application satisfaisant les deux conditions suivantes :

(i) f est $\mathscr{R}$-intégrable sur U_1 ;

(ii) $f \circ \mathbf{h}\,|\operatorname{Jac}\mathbf{h}|$ est $\mathscr{R}$-intégrable sur U_2.

Alors on a l'égalité :

$$\int_{U_2} f(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_{U_1} f\circ\mathbf{h}(\mathbf{x})\,\big|\operatorname{Jac}\mathbf{h}(\mathbf{x})\big|\,\mathrm{d}\mathbf{x}. \tag{13}$$

Il est possible (et très utile...) d'étendre la formule de changement de variables à certains cas d'intégrales généralisées. Nous ne ferons pas de théorie générale et traiterons chaque cas en passant à la limite (*cf.* les exemples de la partie 1.4.2 ci-dessous). Nous reviendrons plus loin sur la question dans le cas de l'intégrale de Henstock-Kurzweil (*cf.* la section 4). Le résultat suivant se déduit du théorème 58 ci-dessus.

Proposition 59 (Changement de variables). Soient U_1 et U_2 deux ouverts *cubables* de $\mathbb{R}^n$ et $\mathbf{h} : U_1 \to U_2$ un difféomorphisme de classe $\mathcal{C}^1$ dont le jacobien est *borné*. Si $f : U_2 \to \mathbb{R}$ est continue et bornée sur U_2, alors f et $f \circ \mathbf{h} \,|\operatorname{Jac}\mathbf{h}|$ sont intégrables au sens de Riemann respectivement sur U_2 et U_1 et l'on a :

$$\int_{U_2} f(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_{U_1} f \circ \mathbf{h}(\mathbf{x})\,\big|\operatorname{Jac}\mathbf{h}(\mathbf{x})\big|\,\mathrm{d}\mathbf{x}. \tag{14}$$

Démonstration. Si $f : U_2 \to \mathbb{R}$ est continue et bornée sur U_2, alors $f \circ \mathbf{h} \,|\operatorname{Jac}\mathbf{h}|$ est continue et bornée sur U_1, donc f et $f \circ \mathbf{h} \,|\operatorname{Jac}\mathbf{h}|$ sont absolument intégrables au sens de Riemann respectivement sur U_2 et U_1. ■

Point Méthode

1. Dans la pratique, on veut souvent calculer l'intégrale d'une fonction bornée f sur un *compact* cubable Δ. On ne peut pas appliquer directement le théorème de changement de variables sur Δ. On peut parfois le faire sur un voisinage ouvert de Δ, mais, si possible, on préfère utiliser un $\mathcal{C}^1$-difféomorphisme $\mathbf{h}$ de $\operatorname{Int}\Delta$ sur un ouvert cubable $U \subset \mathbb{R}^n$. D'après la proposition 50 de la page 207, on a alors (sous réserve de la vérification des conditions du théorème 58) :

$$\int_{\Delta} f(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_{\operatorname{Int}\Delta} f(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_U f \circ \mathbf{h}\,|\operatorname{Jac}\mathbf{h}|.$$

   De plus $\overline{U}$ est alors cubable et, pour tout prolongement *borné* g de $f \circ \mathbf{h}|\operatorname{Jac}\mathbf{h}|$ à $\overline{U}$, on a $\displaystyle\int_U f \circ \mathbf{h}(\mathbf{x})\,|\operatorname{Jac}\mathbf{h}((x)|\,\mathrm{d}\mathbf{x} = \int_{\overline{U}} g(\mathbf{x})\,\mathrm{d}\mathbf{x}$, on est ainsi ramené au calcul de cette dernière intégrale. Si par exemple $\overline{U}$ est un pavé fermé, on peut alors utiliser le théorème de Fubini.

2. On peut parfois simplifier le calcul de l'intégrale d'une fonction f sur un domaine A en fractionnant A en parties d'intersections deux à deux $\mathscr{R}$-négligeables (*cf.* le corollaire 49 de la page 207) et, si c'est possible, utiliser des symétries sur A et f pour faire des changements de variables.

Remarque. Dans la situation de la proposition ci-dessus, il est possible que $\operatorname{Jac}\mathbf{h}$ soit borné mais que $\operatorname{Jac}\mathbf{h}^{-1}$ ne le soit pas. Cette situation peut se rencontrer dans un passage en coordonnées polaires (r^{-1} n'est pas borné au voisinage de 0).

De nombreux systèmes « concrets » (que l'on rencontre par exemple en physique ou en mécanique) ont une symétrie centrale dans le plan ou dans l'espace, ou un axe de symétrie dans l'espace. Dans ces cas, il est utile de travailler en coordonnées polaires, sphériques ou cylindriques. Nous allons maintenant expliciter la formule de changement de variables dans ces trois cas.

Exemples.

1. Le changement de variables en coordonnées polaires est donné par l'application :
$$P : \begin{pmatrix} r \\ \theta \end{pmatrix} \mapsto \begin{pmatrix} x = r\cos\theta \\ y = r\sin\theta \end{pmatrix}$$
où (r, θ) sont les coordonnées polaires du point $(x, y) \in \mathbb{R}^2$. La matrice jacobienne de l'application continûment différentiable P est :
$$J_{(r,\theta)}P = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$
et son déterminant vaut :
$$\det D_{(r,\theta)}P = r.$$
Par conséquent, si $V \subset]0, +\infty[\times]0, 2\pi[$ est un ouvert quarrable du plan (r, θ) paramétrant un ouvert quarrable U du plan (x, y) et si $f : U \to \mathbb{R}$ est une fonction continue, on a $r > 0$ et :
$$\int_U f \begin{pmatrix} x \\ y \end{pmatrix} \mathrm{d}x\,\mathrm{d}y = \int_V f \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix} r\,\mathrm{d}r\,\mathrm{d}\theta.$$

2. En coordonnées sphériques, un point de $\mathbb{R}^3$ est repéré par sa distance r à l'origine, sa colatitude θ et sa longitude φ. Le changement de coordonnées est donné par l'application :
$$S : \begin{pmatrix} r \\ \theta \\ \varphi \end{pmatrix} \mapsto \begin{pmatrix} x = r\sin\theta\cos\varphi \\ y = r\sin\theta\sin\varphi \\ z = r\cos\theta \end{pmatrix}.$$
Le déterminant de la matrice jacobienne est :
$$\det \begin{pmatrix} \sin\theta\cos\varphi & r\cos\theta\cos\varphi & -r\sin\theta\sin\varphi \\ \sin\theta\sin\varphi & r\cos\theta\sin\varphi & r\sin\theta\cos\varphi \\ \cos\theta & -r\sin\theta & 0 \end{pmatrix} = r^2\sin\theta.$$
Par conséquent, si $V \subset]0, +\infty[\times]0, \pi[\times]0, 2\pi[$ est un ouvert cubable de l'espace (r, θ, φ) paramétrant un ouvert U cubable de l'espace (x, y, z) et si $f : U \to \mathbb{R}$ est une fonction continue, on a $r^2\sin\theta > 0$, $r^2\sin\theta$ est borné sur V et :
$$\int_U f \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \int_V f \begin{pmatrix} r\sin\theta\cos\varphi \\ r\sin\theta\sin\varphi \\ r\cos\theta \end{pmatrix} r^2\sin\theta\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\varphi.$$

3. En coordonnées cylindriques, un point $(x, y, z) \in \mathbb{R}^3$ est repéré par les coordonnées polaires (r, θ) du point (x, y) et son altitude z. D'après le théorème

de Fubini et la formule de changement de variables en coordonnées polaires, si $V \subset]0,+\infty[\times]0,2\pi[\times\mathbb{R}$ est un ouvert cubable de l'espace (r,θ,z) paramétrant un ouvert cubable U de l'espace (x,y,z) et si $f : U \to \mathbb{R}$ est une fonction continue, on a :

$$\int_U f\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \int_V f\begin{pmatrix} r\cos\theta \\ r\sin\theta \\ z \end{pmatrix} r\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}z.$$

1.4.2 Exemples de calculs d'intégrales par changement de variables

Dans les exercices suivants, on utilisera les *Point méthode* des pages 203 et 216.

Exercice 7.

Soit $\Delta := \{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 \leqslant 1,\ y \geqslant 0\} \subset \mathbb{R}^2$. Calculer $\displaystyle\iint_\Delta y\,\mathrm{d}x\mathrm{d}y$ en utilisant un changement de variables.

Solution. On utilise le passage en coordonnées polaires $x := r\cos\theta,\ y := r\sin\theta$.
Le pavé *ouvert* $U := \{(r,\theta) \in \mathbb{R}^2 \mid 0 < r < 1,\ 0 < \theta < \pi\}$ est transformé en l'ouvert cubable $\operatorname{Int}\Delta = \{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 < 1,\ y > 0\}$. D'après la formule de changement de variables et la proposition 50 :

$$\iint_\Delta y\,\mathrm{d}x\mathrm{d}y = \iint_{\operatorname{Int}\Delta} y\,\mathrm{d}x\mathrm{d}y = \iint_U r^2\sin\theta\,\mathrm{d}r\mathrm{d}\theta = \iint_{\overline{U}} r^2\sin\theta\,\mathrm{d}r\mathrm{d}\theta.$$

D'après le théorème de Fubini :

$$\iint_{\overline{U}} r^2\sin\theta\,\mathrm{d}r\mathrm{d}\theta = \left(\int_0^1 r^2 dr\right)\left(\int_0^\pi \sin\theta\,\mathrm{d}\theta\right) = \frac{1}{3}\int_0^\pi \sin\theta\,\mathrm{d}\theta = -\frac{1}{3}\left[\cos\theta\right]_0^\pi = \frac{2}{3}\cdot$$

On retrouve le résultat de l'exercice 6 de la page 213.

Exercice 8.

Soient $0 < a \leqslant b$. Soit Δ le compact du premier quadrant de $\mathbb{R}^2$ délimité par les courbes $y = x$, $y^2 - x^2 = 1$, $xy = a$, $xy = b$. Calculer $\displaystyle\iint_\Delta (y^2-x^2)^{xy}\,(x^2+y^2)\,\mathrm{d}x\mathrm{d}y$.

Solution. On utilise le changement de variables $u := xy,\ v := y^2 - x^2$.
L'ouvert cubable $\operatorname{Int}\Delta = \{(x,y) \in \mathbb{R}^2 \mid a < xy < b,\ 0 < y^2 - x^2 < 1\}$ est transformé en le pavé ouvert $U := \{(u,v) \in \mathbb{R}^2 \mid a < u < b,\ 0 < v < 1\}$. On a :

$$\frac{\partial u}{\partial x} = y,\ \frac{\partial u}{\partial y} = x,\ \frac{\partial v}{\partial x} = -2x,\ \frac{\partial v}{\partial y} = 2y,$$

le jacobien de la transformation est donc $\begin{pmatrix} y & x \\ -2x & 2y \end{pmatrix} = 2(x^2+y^2) > 0$.

D'après la formule de changement de variables et la proposition 50 de la page 207 :

$$\begin{aligned}\iint_\Delta (y^2-x^2)^{xy}\,(x^2+y^2)\,\mathrm{d}x\mathrm{d}y &= \iint_{\operatorname{Int}\Delta} (y^2-x^2)^{xy}\,(x^2+y^2)\,\mathrm{d}x\mathrm{d}y \\ &= \frac{1}{2}\iint_U v^u\,\mathrm{d}u\mathrm{d}v = \frac{1}{2}\iint_{\overline{U}} v^u\,\mathrm{d}u\mathrm{d}v.\end{aligned}$$

D'après le théorème de Fubini :

$$\iint_{\overline{U}} v^u \,\mathrm{d}u\mathrm{d}v = \int_a^b \mathrm{d}u \int_0^1 v^u \,\mathrm{d}v = \int_a^b \mathrm{d}u \left[\frac{v^{u+1}}{u+1}\right]_0^1 = \int_a^b \frac{\mathrm{d}u}{u+1} = \ln\frac{b+1}{a+1}\cdot$$

Donc $\displaystyle\iint_\Delta (y^2-x^2)^{xy}\,(x^2+y^2)\,\mathrm{d}x\mathrm{d}y = \frac{1}{2}\ln\frac{b+1}{a+1}\cdot$

Exercice 9.

Soient $a,\ b>0$. On note $\Delta := \{(x,y)\in\mathbb{R}^2 \mid \frac{x^2}{a^2}+\frac{y^2}{b^2}\leqslant 1\}$.

Calculer $\displaystyle\iint_\Delta \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y$

Solution. Notons : $\Delta^+ := \{(x,y)\in\mathbb{R}^2 \mid \dfrac{x^2}{a^2}+\dfrac{y^2}{b^2}\leqslant 1,\ y\geqslant 0\}$

$$\Delta^- := \{(x,y)\in\mathbb{R}^2 \mid \frac{x^2}{a^2}+\frac{y^2}{b^2}\leqslant 1,\ y\leqslant 0\}.$$

Le compact $\Delta^+\cap\Delta^- = \{(x,y)\in\mathbb{R}^2 \mid -a\leqslant x\leqslant a,\ y=0\}$ est négligeable, donc, d'après la proposition 48 de la page 206 :

$$\iint_\Delta \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y = \iint_{\Delta^+} \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y + \iint_{\Delta^-} \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y.$$

On vérifie (en changeant y en $-y$) que :

$$\iint_{\Delta^+} \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y = \iint_{\Delta^-} \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y,$$

$$\text{donc}\qquad \iint_\Delta \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y = 2\iint_{\Delta^+} \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y.$$

Il suffit donc de calculer l'intégrale sur Δ^+. On utilise une variante du passage en coordonnées polaires transformant Δ^+ en un pavé fermé : $x := ar\cos\varphi,\ y := br\sin\varphi$.
Le pavé ouvert $U := \{(r,\varphi)\in\mathbb{R}^2 \mid 0<r<1,\ 0<\varphi<\pi\}$ est transformé en l'ouvert cubable $\operatorname{Int}\Delta^+ = \left\{(x,y)\in\mathbb{R}^2 \mid \frac{x^2}{a}+\frac{y^2}{b^2}<1,\ y>0\right\}$. D'après la formule de changement de variables et la proposition 50 :

$$\iint_{\Delta^+} \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y = \iint_{\operatorname{Int}\Delta^+} \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y = ab\iint_U \sqrt{1-r^2}\,r\,\mathrm{d}r\mathrm{d}\varphi$$
$$= ab\iint_{\overline{U}} \sqrt{1-r^2}\,r\,\mathrm{d}r\mathrm{d}\varphi.$$

D'après le théorème de Fubini :

$$\iint_{\overline{U}} \sqrt{1-r^2}\,r\,\mathrm{d}r\mathrm{d}\varphi = \pi\int_0^1 \sqrt{1-r^2}\,r\,\mathrm{d}r = \pi\left[-\frac{1}{3}(1-r^2)^{3/2}\right]^1 = \frac{\pi}{3}\cdot$$

On en déduit $\displaystyle\iint_\Delta \sqrt{1-\frac{x^2}{a^2}-\frac{y^2}{b^2}}\,\mathrm{d}x\mathrm{d}y = \frac{2\pi ab}{3}\cdot$

1.4.3 Calculs d'aires et de volumes

Cette partie est consacrée à quelques calculs d'aires et de volumes.

Proposition 60. Soit $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ une *isométrie affine*.
Soient $A \subset \mathbb{R}^n$ et $f : A \to \mathbb{R}$ une fonction continue. Alors :

(i) A est cubable si, et seulement si, $\varphi(A)$ est cubable ;

(ii) si A est cubable $\displaystyle\int_A f = \int_{\varphi(A)} f \circ \varphi^{-1}$. En particulier $\mathrm{vol}(A) = \mathrm{vol}\big(\varphi(A)\big)$.

Démonstration.

(i) On peut appliquer le théorème 59 de la page 216. Plus simplement on peut vérifier qu'une partie $X \subset \mathbb{R}^n$ est $\mathscr{R}$-négligeable si, et seulement si, $\varphi(X)$ est $\mathscr{R}$-négligeable. On utilise ensuite la proposition 38 de la page 201, avec $X = \partial A$.

(ii) On a $|\mathrm{Jac}\,\varphi| = 1$. On utilise le théorème 59 de la page 216. ■

Ainsi pour $n = 2$ et $n = 3$, en particulier, les notions d'aire et de volume sont *invariantes par isométrie affine*. Pour calculer des aires ou des volumes on peut choisir une base orthonormée adaptée au problème.

Exemple : calcul du volume de la boule unité euclidienne de $\mathbb{R}^3$.
On peut calculer le volume de la boule unité de $\mathbb{R}^3$ en utilisant des coordonnées sphériques (ou cylindriques). On peut aussi utiliser des coordonnées cartésiennes et le théorème 57 de la page 214. Les domaines d'intégration sont des intervalles ou des disques, ils sont donc évidemment quarrables.
Soit $B := \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 \leqslant 1\}$. Son volume est égal à :

$$\iiint_B \mathrm{d}x\mathrm{d}y\mathrm{d}z = \int_0^1 \mathrm{d}r \int_0^\pi \mathrm{d}\theta \int_0^{2\pi} r^2 \sin\theta \,\mathrm{d}\varphi,$$

(coordonnées sphériques), ou :

$$\iiint_B \mathrm{d}x\mathrm{d}y\mathrm{d}z = \int_{-1}^1 \mathrm{d}z \iint_{x^2+y^2\leqslant 1-z^2} \mathrm{d}x\mathrm{d}y,$$

(sommation par plaques ou tranches) ou :

$$\iiint_B \mathrm{d}x\mathrm{d}y\mathrm{d}z = \iint_{x^2+y^2\leqslant 1} \mathrm{d}x\mathrm{d}y \int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}} \mathrm{d}z$$

(sommation par piles).
Voici les trois calculs :

$$\int_0^1 \mathrm{d}r \int_0^\pi \mathrm{d}\theta \int_0^{2\pi} r^2 \sin\theta \,\mathrm{d}\varphi = 2 \times 2\pi \times \frac{1}{3} = \frac{4\pi}{3}$$

$$\iiint_B \mathrm{d}x\mathrm{d}y\mathrm{d}z = \int_{-1}^1 \mathrm{d}z \iint_{x^2+y^2\leqslant 1-z^2} \mathrm{d}x\mathrm{d}y = \int_{-1}^1 \pi(1 - z^2)\mathrm{d}z = \frac{4\pi}{3}$$

$$\iiint_B \mathrm{d}x\mathrm{d}y\mathrm{d}z = \iint_{x^2+y^2\leqslant 1} \mathrm{d}x\mathrm{d}y \int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}} \mathrm{d}z = \iint_{x^2+y^2\leqslant 1} 2\sqrt{1-x^2-y^2}\mathrm{d}x\mathrm{d}y,$$

soit (en passant en coordonnées polaires dans le plan des (x, y), c'est-à-dire en coordonnées cylindriques) :

$$\iiint_B \mathrm{d}x\mathrm{d}y\mathrm{d}z = \int_0^{2\pi} \mathrm{d}\theta \int_0^1 2\sqrt{1-r^2}\, r\mathrm{d}r = 2\pi \int_0^1 \sqrt{t}\,\mathrm{d}t = \frac{4\pi}{3}\cdot$$

Dans l'exemple suivant, nous allons généraliser cette dernière méthode pour le calcul du volume de la boule euclidienne en dimension quelconque.

Exemple : calcul du volume de la boule unité euclidienne de $\mathbb{R}^n$.
Notons $B_n(R)$ la boule de centre 0 et de rayon $R > 0$ de $\mathbb{R}^n$ muni de la norme euclidienne canonique et $B_n(1) = B_n$. On se propose de calculer son volume. On notera $V_n(R) := \mathrm{vol}\, B_n(R)$ et $V_n := \mathrm{vol}\, B_n$
D'après la formule de changement de variables, on a $V_n(R) = R^n V_n$ (le jacobien de $\mathbf{x} \mapsto R\mathbf{x}$ est R^n).
On utilise, pour $n \geqslant 3$, des *coordonnées cylindriques généralisées* $(r, \theta, x_3, \ldots, x_n)$, c'est-à-dire le changement de variables :

$$x_1 = r\cos\theta, \quad x_2 = r\sin\theta, \quad x_3 = x_3, \quad \ldots \quad x_n = x_n.$$

Le jacobien est $r \geqslant 0$, donc, en utilisant le théorème de Fubini :

$$\begin{aligned} V_n &= \int_0^1 \int_0^{2\pi} V_{n-2}\big(\sqrt{1-r^2}\big) r\,\mathrm{d}r\,\mathrm{d}\theta = 2\pi V_{n-2} \int_0^1 (1-r^2)^{\frac{n-2}{2}}\, r\,\mathrm{d}r \\ &= \pi V_{n-2} \int_0^1 u^{\frac{n-2}{2}}\,\mathrm{d}u = \frac{2\pi}{n}\, V_{n-2}. \end{aligned}$$

Cette formule permet de calculer V_n par récurrence respectivement pour les valeurs paires et impaires de n à partir de $V_2 = \pi$ et $V_1 = 2$. On obtient :

$$V_{2p} = \frac{\pi}{p} V_{2p-2} = \frac{\pi^2}{p(p-1)} V_{2p-4} = \cdots = \frac{\pi^{p-1}}{p!} V_2 = \frac{\pi^p}{p!}$$

et :

$$\begin{aligned} V_{2p+1} = \frac{2\pi}{2p+1} V_{2p-1} = \frac{(2\pi)^2}{(2p+1)(2p-1)} V_{2p-3} &= \frac{2^3\pi^2 p}{(2p+1)2p(2p-1)} V_{2p-3} \\ &= \cdots = \frac{2^{2p+1}\pi^p p!}{(2p+1)!}\cdot \end{aligned}$$

On peut donner une formule indépendante de la parité de n en utilisant la fonction Γ (*cf.* le module *Intégration* du livre [L2], exercices 16 de la page 687 et 27 de la page 708). À partir de la formule de récurrence pour Γ et $\Gamma(1/2) = \sqrt{\pi}$, on vérifie :

$$V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}\cdot$$

En effet le second membre de cette égalité vérifie la même formule de récurrence que V_n et les mêmes conditions initiales :

$$V_2 = \pi = \frac{\pi}{\Gamma(2)} \quad \text{et} \quad V_1 = 2 = 2\,\frac{\sqrt{\pi}}{\sqrt{\pi}} = \frac{\sqrt{\pi}}{\Gamma(\frac{3}{2})}\cdot$$

Exercice 10.

Calculer le volume du sous-ensemble borné Δ de $\mathbb{R}^3$ délimité par le paraboloïde de révolution $\mathcal{P}$ d'équation $x^2 + y^2 - 2pz = 0$ $(p > 0)$ et le plan P d'équation $y - z + a = 0$.

Solution.

Première méthode On vérifie que Δ est cubable (le paraboloïde et le plan sont des surfaces lisses), donc $\operatorname{vol}\Delta = \displaystyle\iiint_\Delta \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$.

On vérifie que l'intersection de $\mathcal{P}$ et P est la courbe d'équations :

$$x^2 + (y-p)^2 = p^2 + 2ap \quad \text{et} \quad z = y + a.$$

C'est l'intersection d'un plan et d'un cylindre de révolution (on montre facilement par un argument géométrique que c'est une ellipse, nous le prouverons plus loin par le calcul). On pose $r^2 := p^2 + 2ap$. La base du cylindre est le cercle de centre $(0,p)$ et de rayon r d'équation $x^2+(y-p)^2 = r^2$. La projection de Δ sur le plan « horizontal » parallèlement à l'axe des z est le disque D correspondant. En appliquant le théorème de Fubini, on obtient :

$$\begin{aligned}\operatorname{vol}\Delta = \iiint_\Delta \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z &= \iint_D \mathrm{d}x\,\mathrm{d}y \int_{\frac{x^2+y^2}{2p}}^{y+a} \mathrm{d}z = \iint_D \big(y + a - \frac{x^2+y^2}{2p}\big)\,\mathrm{d}x\,\mathrm{d}y \\ &= -\frac{1}{2p}\iint_D \big(x^2 + (y-p)^2 - r^2\big)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Pour calculer l'intégrale $I := \displaystyle\iint_D \big(x^2 + (y-p)^2 - r^2\big)\,\mathrm{d}x\,\mathrm{d}y$ on fait le changement de variables $X := x$, $Y := y - p$, puis l'on passe en coordonnées polaires :

$$X = \rho\cos\theta \quad \text{et} \quad Y = \rho\sin\theta.$$

On obtient $I = \displaystyle\int_0^r\int_0^{2\pi} (\rho^2 - r^2)\rho\,\mathrm{d}\rho\,\mathrm{d}\theta = 2\pi\big(\frac{r^4}{4} - \frac{r^2}{2}\big) = -2\pi\frac{r^4}{4}$, puis :

$$\operatorname{vol}\Delta = \frac{2\pi}{2p}\frac{r^4}{4} = \frac{\pi}{p}\frac{r^4}{4} = \frac{\pi}{4}(p^3 + 4ap^2 + 4a^2p).$$

Deuxième méthode

On effectue une *transformation orthogonale* (rotation autour de l'axe des x) :

$$x = X, \quad y = \frac{Y - Z}{\sqrt{2}}, \quad z = \frac{Y + Z}{\sqrt{2}},$$

transformant le plan P en le plan d'équation $Z = \frac{a}{\sqrt{2}}$.

La courbe intersection de $\mathcal{P}$ et P est transformée en la courbe d'équations :

$$X^2 + \left(\frac{Y - Z}{\sqrt{2}} - p\right)^2 = r^2 \quad \text{et} \quad Z = \frac{a}{\sqrt{2}},$$

ou

$$X^2 + \frac{1}{2}\left(Y - \frac{a}{\sqrt{2}} - p\sqrt{2}\right)^2 = r^2 \quad \text{et} \quad Z = \frac{a}{\sqrt{2}}\cdot$$

Cette courbe est une ellipse de petit axe r et de grand axe $r\sqrt{2}$.
Sa surface est $\pi r^2\sqrt{2} = \pi\sqrt{2}(p^2 + 2ap)$.
On montrerait de même que l'intersection de $\mathcal{P}$ et du plan d'équation $z = y + h$ est, pour tout $h > -p/2$ une ellipse de surface $S(h) := \pi\sqrt{2}(p^2 + 2hp)$. Pour $h := p/2$ l'intersection est réduite à un point.
On en déduit, en posant $h := \sqrt{2}Z$:

$$\begin{aligned}\operatorname{vol}\Delta = \int_{-p/2}^{a} S(h)\,\mathrm{d}h = \pi\int_{-p/2}^{a}(p^2 + 2hp)\,\mathrm{d}h &= \pi\left(p^2\big(a + \frac{p}{2}\big) + p\big(a^2 - \frac{p^2}{4}\big)\right)\\ &= \frac{\pi}{4}(p^3 + 4ap^2 + 4a^2p).\end{aligned}$$

Exemple : aires de la boucle d'une strophoïde et de la portion comprise entre la strophoïde et son asymptote.
Soit $a > 0$. Une paramétrisation de la strophoïde est :

$$\sigma : t \in \mathbb{R} \mapsto \begin{cases} x(t) = a\frac{1-t^2}{1+t^2} \\ y(t) = tx = at\frac{1-t^2}{1+t^2}\cdot \end{cases}$$

En posant $\theta := \arctan t$, on obtient :

$$x(\theta) = a\cos 2\theta,\ y(\theta) = a\tan\theta\cos 2\theta$$

et la représentation de la strophoïde en paramétrage polaire $\rho = a\frac{\cos 2\theta}{\cos\theta}$, $\theta \in \,]-\pi/2, \pi/2[$.

1. Calculons l'aire $\mathcal{A}$ de la boucle de la strophoïde (dont la frontière est $\sigma\big([0,1]\big)$).
   On utilise les coordonnées polaires. Par symétrie ($\theta \mapsto -\theta$) on a :

$$\begin{aligned}\mathcal{A} &= 2\int_0^{\pi/4}\int_0^{\frac{a\cos 2\theta}{\cos\theta}} \rho\,\mathrm{d}\rho\,\mathrm{d}\theta = \int_0^{\pi/4}\left[\rho^2\right]_0^{\frac{a\cos 2\theta}{\cos\theta}}\mathrm{d}\theta = a^2\int_0^{\pi/4}\frac{\cos^2 2\theta}{\cos^2\theta}\,\mathrm{d}\theta\\ &= a^2\int_0^{\pi/4}\frac{1-\sin^2 2\theta}{\cos^2\theta}\,\mathrm{d}\theta = a^2\int_0^{\pi/4}\left(\frac{1}{\cos^2\theta} - 4\sin^2\theta\right)\mathrm{d}\theta\\ &= a^2\left[\tan\theta + \sin 2\theta - 2\theta\right]_0^{\pi/4} = a^2\Big(2 - \frac{\pi}{2}\Big).\end{aligned}$$

2. Calculons « l'aire » $\mathcal{A}'$ de l'ensemble compris entre la strophoïde et son asymptote (dont la frontière est la réunion de $\sigma\left(]-\infty,-1]\right) \cup \sigma\left([1,+\infty[\right)$ et de la droite d'équation $x = -a$).
   Première méthode
   Il y a un problème de définition de l'aire $\mathcal{A}'$ car le domaine compris entre la strophoïde et son asymptote *n'est pas borné*. Dans la partie 1.5 nous définirons cette aire

en utilisant l'intégrale de Riemann généralisée. Pour l'instant nous procéderons de la façon suivante. L'arc de strophoïde contenu dans $]-a,0]\times\mathbb{R}_+$ est le graphe de la fonction :

$$x\in\,]-a,0]\mapsto y(x):=-x\sqrt{\frac{a-x}{a+x}}.$$

Soient $D:=\left\{(x,y)\in\mathbb{R}^2\;\middle|\;-a\leqslant x\leqslant 0,\ 0\leqslant y\leqslant -x\sqrt{\frac{a-x}{a+x}}\right\}$ et, pour tout b tel que $-a<-b<0$:

$$D(b):=\left\{(x,y)\in\mathbb{R}^2\;\middle|\;-b\leqslant x\leqslant 0,\ 0\leqslant y\leqslant -x\sqrt{\frac{a-x}{a+x}}\right\}.$$

La partie $D(b)$ est quarrable d'aire :

$$S(b)=\iint_{D(b)}\mathrm{d}x\,\mathrm{d}y=\int_{-b}^{0}-x\sqrt{\frac{a-x}{a+x}}\,\mathrm{d}x.$$

On *définit* (en utilisant la symétrie) l'aire $\mathcal{A}'$ comme la limite de $2S(b)$ quand b tend vers a^- (*cf.* l'exemple de la page 230). Ainsi $\mathcal{A}'=2\displaystyle\int_{-a}^{0}-x\sqrt{\frac{a-x}{a+x}}\,\mathrm{d}x$, l'intégrale étant à prendre au sens généralisé (nous allons voir qu'elle converge).

Pour étudier cette intégrale généralisée sur $]-a,0]$, on utilise le changement de variables $x=a\cos 2\varphi,\ \varphi\in\,]-\pi/2,-\pi/4]$ (*cf.* le module *Intégration* du livre [L2]). On a $\sqrt{\frac{a-x}{a+x}}=\tan\varphi$ et l'obtient :

$$\begin{aligned}\mathcal{A}'&=-4a^2\int_{-\pi/2}^{-\pi/4}\cos 2\varphi\sin 2\varphi\tan\varphi\,\mathrm{d}\varphi=8a^2\int_{\pi/2}^{\pi/4}\cos 2\varphi\sin^2\varphi\,\mathrm{d}\varphi\\&=4a^2\int_{\pi/2}^{\pi/4}(2\cos 2\varphi-1+\cos 4\varphi)\,\mathrm{d}\varphi=a^2\big(2+\frac{\pi}{2}\big).\end{aligned}$$

Deuxième méthode

On « constate » que $\mathcal{A}'=2(\mathcal{A}'_1+\mathcal{A}'_2)$, où $\mathcal{A}'_1=a^2/2$ est l'aire du triangle de sommets $(-a,0)$, $(0,0)$, $(-a,a)$ et $\mathcal{A}'_2$ « l'aire » du domaine (non borné...) limité par la strophoïde, son asymptote et le segment joignant $(0,0)$ à $(-a,a)$. On calcule ensuite $\mathcal{A}'_2$ en utilisant les coordonnées polaires. Le lecteur pourra formaliser ce raisonnement en utilisant les techniques de la partie 1.5 ci-dessous.

On obtient :

$$\begin{aligned}\mathcal{A}'_2&=\frac{1}{2}\int_{\pi/4}^{\pi/2}\int_{\frac{a\cos 2\theta}{\cos\theta}}^{-\frac{a}{\cos\theta}}\mathrm{d}(\rho^2)\,\mathrm{d}\theta=\frac{a^2}{2}\int_{\pi/4}^{\pi/2}\frac{1-\cos^2 2\theta}{\cos^2\theta}\,\mathrm{d}\theta=\frac{a^2}{2}\int_{\pi/4}^{\pi/2}\frac{\sin^2 2\theta}{\cos^2\theta}\,\mathrm{d}\theta\\&=a^2\int_{\pi/4}^{\pi/2}2\sin^2\theta\,\mathrm{d}\theta=a^2\int_{\pi/4}^{\pi/2}(1-\cos 2\theta)\,\mathrm{d}\theta=a^2\left(\frac{\pi}{4}+\frac{1}{2}\right).\end{aligned}$$

D'où $\mathcal{A}'=2\left(\frac{a^2}{2}+a^2\big(\frac{\pi}{4}+\frac{1}{2}\big)\right)=a^2\big(2+\frac{\pi}{2}\big)$.

1.5 Intégrales multiples généralisées

Dans le module *Intégration* du livre [L2] nous avons introduit la notion d'intégrale généralisée ou impropre, étendant la notion d'intégrale de Riemann sur un segment à une notion d'intégrale sur un intervalle réel arbitraire. Il est naturel de se demander s'il est possible de procéder de façon « analogue » pour les intégrales multiples. Nous allons voir que ce n'est pas si simple...

Une présentation approfondie du sujet serait nécessairement très technique, nous nous limiterons donc au strict minimum permettant de traiter rigoureusement des problèmes « usuels » (en particulier en géométrie, physique, mécanique...). Nous ne considérerons que le cas des intégrales *absolument convergentes*. La notion d'intégrale semi-convergente est en effet *beaucoup plus délicate* qu'en dimension 1 (*cf.* l'exemple ci-dessous page 227 et l'exercice III.1.17) et elle n'est raisonnablement formulable que dans le cadre de l'intégrale multiple de Henstock-Kurzweil. Elle ne sera pas traitée dans cet ouvrage.

En revanche, nous reviendrons plus loin, dans la section 4, sur l'étude des intégrales généralisées *absolument convergentes* dans le cadre plus général de l'intégrale de Henstock-Kurzweil.

Nous ne donnerons pas dans cette partie de formule de Fubini pour les intégrales généralisées. Dans la pratique, on utilisera les formules usuelles et un passage à la limite. On peut procéder de façon similaire pour la formule de changement de variables, mais nous donnerons une formule valable pour les intégrales généralisées (que nous prouverons dans la section 4 en utilisant l'intégrale de Henstock-Kurzweil à plusieurs variables) qui permet d'éviter les passages à la limite.

1.5.1 Suites exhaustives de compacts

Définition 13. Soit $A \subset \mathbb{R}^n$. On appelle *suite exhaustive de compacts* de A toute suite croissante $(K_k)_{k\in\mathbb{N}}$ de compacts de $\mathbb{R}^n$ contenus dans A, telle que, pour tout compact K de $\mathbb{R}^n$ inclus dans A, il existe $m \in \mathbb{N}$ tel que $K \subset K_m$.

Les singletons de A sont compacts, donc si $(K_k)_{k\in\mathbb{N}}$ est une suite exhaustive de compacts, on a $\bigcup\limits_{k\in\mathbb{N}} K_k = A$.

On trouve dans de nombreux ouvrages une autre définition, *plus restrictive*, de la notion de suite exhaustive de compacts :

$$\text{pour tout } k \in \mathbb{N},\ K_k \text{ est un compact contenu dans } A \text{ et } K_k \subset \operatorname{Int} K_{k+1}. \qquad (\star)$$

Le lecteur vérifiera que si la condition $(\star)$ est vérifiée, la définition 13 est satisfaite (*cf.* aussi l'exercice III.1.15).

La définition $(\star)$ est bien adaptée à l'analyse fonctionnelle mais elle est trop restrictive pour l'étude des intégrales multiples généralisées où l'on intègre sur des domaines qui ne sont pas nécessairement ouverts.

Exemples.

1. On note, pour tout $k \in \mathbb{N}$, $\mathbf{Q}_k := [-k,k]^n$ et $B'(0,k)$ la boule euclidienne fermée de centre 0 et de rayon k. Les suites $(\mathbf{Q}_k)_{k\in\mathbb{N}}$ et $\big(B'(0,k)\big)_{k\in\mathbb{N}}$ sont des suites exhaustives de compacts de $\mathbb{R}^n$.
2. Notons $\Delta := \{(x,y) \in \mathbb{R}^2 \mid x \geqslant 0,\ y \geqslant 0\}$ le premier quadrant fermé du plan. Pour tout $k \in \mathbb{N}$ on note :

   $K_k := \{(x,y) \in \Delta \mid x^2 + y^2 \leqslant k^2\}$

   $K'_k := \{(x,y) \in \Delta \mid \sup(x,y) \leqslant k\} = [0,k]^2$ (c'est un pavé carré fermé).

   On vérifie que $(K_k)_{k\in\mathbb{N}}$ et $(K'_k)_{k\in\mathbb{N}}$ sont des suites exhaustives de compacts de Δ.
3. Notons $U := \{(x,y) \in \mathbb{R}^2 \mid 0 < x,\ 0 < y\}$ le premier quadrant ouvert du plan. Pour tout $k \in \mathbb{N}^*$ on note $L_k := \{(x,y) \in U \mid 1/k \leqslant x \leqslant k,\ 1/k \leqslant y \leqslant k\}$ (c'est un pavé carré fermé). On vérifie que $(L_k)_{k\in\mathbb{N}^*}$ est une suite exhaustive de compacts de U.

En pratique, comme dans les exemples ci-dessus, on utilisera seulement des suites exhaustives de compacts *cubables* de A, par contre A lui même ne sera pas nécessairement cubable (et pas nécessairement borné). Dans les exemples et applications, nous ne ferons pas de vérifications détaillées, laissant ce soin au lecteur (*cf.* le *Point Méthode* de la page 203).

Exercice 11.

(i) On note $\Delta := [0,1]^2 \setminus \{(1,1)\}$ et :

$\forall\, k \in \mathbb{N}^*, \quad C'_k := [0,1]\times[0,1-1/k]\,,\ C''_k := [0,1-1/k]\times[0,1]\,, C_k := C'_k \cup C''_k.$

Montrer que $(C_k)_{k\in\mathbb{N}^*}$ est une suite exhaustive de compacts pavables de Δ.

(ii) Pour tout $k \in \mathbb{N}^*$, on note $K_k := \{(x,y) \in \Delta \mid x + y \leqslant 2 - 1/k\}$ Montrer que $(K_k)_{k\in\mathbb{N}^*}$ est une suite exhaustive de compacts cubables de Δ.

Solution.

(i) Le compact C_k est réunion de deux carés fermés, il est donc pavable.
On a :

$$\bigcup_{k\in\mathbb{N}} C_k = \left(\bigcup_{k\in\mathbb{N}} C'_k\right) \cup \left(\bigcup_{k\in\mathbb{N}} C''_k\right) = ([0,1]\times[0,1[) \cup ([0,1[\times[0,1]) = \Delta.$$

Si $K \subset \Delta$ est compact, on a $(1,1) \notin K$, donc $d\big((1,1),K\big) = \delta > 0$ (la distance étant calculée avec la norme $\sup$). Il existe $m \in \mathbb{N}^*$ tel que $1/m < \delta$ et l'on a $K \subset C_m$.

(ii) Le compact K_k est défini par un système de 5 inégalités affines, il est donc cubable. On a $C_k \subset K_k$, donc $\bigcup\limits_{k\in\mathbb{N}} K_k = \Delta$. Si $K \subset \Delta$ est compact, la fonction continue $(x,y) \mapsto x + y$ est bornée par $0 < M < 2$ sur K. Notons $\chi = 2 - M$, il existe $m \in \mathbb{N}^*$ tel que $1/m < \chi$ et l'on a $K \subset K_m$.

1.5.2 Intégrales généralisées absolument convergentes

Définition 14. Soient $A \subset \mathbb{R}^n$ et une application $f : A \to \mathbb{R}$.

(i) L'application f est dite *intégrable au sens de Riemann* sur la suite exhaustive de compacts $(K_k)_{k\in\mathbb{N}}$ de A si, pour tout $k \in \mathbb{N}$, f est intégrable au sens de Riemann sur K_k.

(ii) L'application f est dite *localement intégrable au sens de Riemann* sur A s'il existe *une* suite exhaustive de compacts de A sur laquelle f est intégrable au sens de Riemann.

La terminologie « f est localement intégrable au sens de Riemann sur A » est un peu trompeuse. Cette condition n'implique pas que f est intégrable au sens de Riemann sur tout compact $K \subset A$.

Si f est *intégrable au sens de Riemann* sur la suite exhaustive de compacts $(K_k)_{k\in\mathbb{N}}$ de A, $|f|$ l'est aussi. Si f est *localement intégrable au sens de Riemann* sur A, $|f|$ l'est aussi.

Si f est *continue* et s'il existe une suite exhaustive de compacts *cubables* de A, f est intégrable au sens de Riemann sur cette suite (*cf.* la proposition 44 de la page 205), elle est donc localement intégrable au sens de Riemann sur A.

Compte tenu des résultats connus en dimension 1, quand $S := \lim\limits_{k\to+\infty} \displaystyle\int_{K_k} f$ existe pour une suite exhaustive de compacts $(K_k)_{k\in\mathbb{N}}$ de A, il serait « tentant » de dire que f est intégrable au sens de Riemann généralisé sur A et que son intégrale est S. En fait ce n'est pas possible car la limite S peut dépendre de la suite exhaustive de compacts choisie ou exister pour une suite et pas pour une autre comme le montre l'exemple suivant (dû à Cayley).

Exemple. Soit $A := \{(x,y) \in \mathbb{R}^2 \mid x \geqslant 0,\ y \geqslant 0\}$ le premier quadrant de $\mathbb{R}^2$. On définit $f : (x,y) \in A \mapsto \sin(x^2+y^2) \in \mathbb{R}$ et l'on considère deux suites de compacts de A : $\big(K_k := [0,k]^2\big)_{k\in\mathbb{N}}$ et $\big(K'_k := \{(x,y) \in A \mid x^2+y^2 \leqslant k^2\}\big)_{k\in\mathbb{N}}$. On vérifie qu'elles sont exhaustives.

On a, en utilisant la formule de Fubini :

$$\begin{aligned}\int_{K_k} f &= \int_{[0,k]^2} \sin(x^2+y^2)\,\mathrm{d}x\,\mathrm{d}y = \int_{[0,k]^2} (\sin x^2 \cos y^2 + \cos x^2 \sin y^2)\,\mathrm{d}x\,\mathrm{d}y \\ &= \Big(\int_0^k \cos y^2 \mathrm{d}y\Big)\Big(\int_0^k \sin x^2 \mathrm{d}x\Big) + \Big(\int_0^k \cos x^2 \mathrm{d}x\Big)\Big(\int_0^k \sin y^2 \mathrm{d}y\Big) \\ &= 2\Big(\int_0^k \sin x^2 \mathrm{d}x\Big)\Big(\int_0^k \cos x^2 \mathrm{d}x\Big)\end{aligned}$$

On rappelle que $x \mapsto \sin x^2$ et $x \mapsto \cos x^2$ sont intégrables au sens généralisé sur $[0,+\infty[$ et que $\displaystyle\int_0^{+\infty} \sin x^2\,\mathrm{d}x = \int_0^{+\infty} \cos x^2\,\mathrm{d}x = \sqrt{\pi/8}$ (ce sont les intégrales de Fresnel, *cf.* l'exercice 21 de la page 696 du livre [L2] et l'exercice 19 de la page 241 du présent ouvrage). On en déduit $\lim\limits_{k\to+\infty} \displaystyle\int_{K_k} f = \pi/4$.

Par ailleurs on a, en utilisant un passage en coordonnées polaires :

$$\int_{K'_k} f = \int_{\{x^2+y^2\leqslant k^2\}} \sin(x^2+y^2)\,\mathrm{d}x\,\mathrm{d}y = \int_0^k \int_0^{\pi/2} \sin r^2\, r\mathrm{d}r\mathrm{d}\theta$$
$$= \frac{\pi}{4}\int_0^{k^2} \sin u\,\mathrm{d}u = \frac{\pi}{4}\big(1-\cos k^2\big).$$

On vérifie que la suite $(1-\cos k^2)_{k\in\mathbb{N}}$ n'a pas de limite (*cf.* l'exercice III.1.16). Ainsi la suite $\left(\int_{K'_k} f\right)_{k\in\mathbb{N}}$ n'a pas de limite.

L'exemple précédent suggère de se limiter à des fonctions *absolument intégrables*, ce que nous ferons dans la suite.

Proposition 61. Soient $A\subset\mathbb{R}^n$ et une application $f : A\to\mathbb{R}$. On suppose qu'il existe une suite exhaustive de compacts $(K_k)_{k\in\mathbb{N}}$ de A sur laquelle f est intégrable au sens de Riemann telle que la suite $\sigma := \left(\int_{K_k} |f|\right)_{k\in\mathbb{N}}$ soit convergente de limite Σ. Alors :

(i) la suite $s := \left(\int_{K_k} f\right)_{k\in\mathbb{N}}$ a une limite S ;

(ii) pour toute suite exhaustive de compacts $(K'_k)_{k\in\mathbb{N}}$ de A sur laquelle f est intégrable au sens de Riemann :

a) la suite $\sigma' := \left(\int_{K'_k} |f|\right)_{k\in\mathbb{N}}$ est convergente et admet pour limite Σ ;

b) la suite $s' := \left(\int_{K'_k} f\right)_{k\in\mathbb{N}}$ est convergente et admet pour limite S.

Notons que la suite à termes positifs $\sigma' := \left(\int_{K'_k} |f|\right)_{k\in\mathbb{N}}$ est convergente si, et seulement si, elle est majorée. En pratique on vérifie cette deuxième condition.

Démonstration.

(i) Montrons que la suite s est de Cauchy.
Pour tous $0\leqslant k\leqslant k'$, on a :

$$\Big|\int_{K_{k'}} f - \int_{K_k} f\Big| = \Big|\int_{K_{k'}\setminus K_k} f\Big| \leqslant \int_{K_{k'}\setminus K_k} |f| = \int_{K_{k'}} |f| - \int_{K_k} |f|.$$

La suite σ est convergente donc de Cauchy et, d'après les inégalités ci-dessus, la suite s est aussi de Cauchy, donc convergente.

(ii) a) Les suites σ et σ' sont croissantes. Pour tout $k \in \mathbb{N}$, il existe $h \in \mathbb{N}$ tel que $K'_k \subset K_h$. On en déduit $\displaystyle\int_{K'_k} |f| \leqslant \int_{K_h} |f| \leqslant \Sigma$. La suite σ' étant croissante et majorée est convergente, notons Σ' sa limite. On a $\Sigma' \leqslant \Sigma$. On montrerait de même que $\Sigma \leqslant \Sigma'$. Ainsi $\Sigma' = \Sigma$.

b) On définit par récurrence une application $\psi : \mathbb{N} \to \mathbb{N}$ par :

$$\psi(0) := \min\{j \in \mathbb{N} \mid K'_0 \subset K_j\}$$

$$\forall n \geqslant 1\,,\ \psi(n) := \max\big\{\psi(n-1)+1, \min\{j \in \mathbb{N} \mid K'_k \subset K_j\}\big\}.$$

On a :

$$\Big|\int_{K_{\psi(k)}} f - \int_{K'_k} f\Big| \leqslant \Big|\int_{K_{\psi(k)} \setminus K'_k} f\Big| \leqslant \int_{K_{\psi(k)} \setminus K'_k} |f| = \int_{K_{\psi(k)}} |f| - \int_{K'_k} |f|.$$

Les suites $\displaystyle\Big(\int_{K_{\psi(k)}} f\Big)$ et $\displaystyle\Big(\int_{K_{\psi(k)}} |f|\Big)$ sont respectivement des suites extraites des suites convergentes $\displaystyle\Big(\int_{K_k} f\Big)$ et $\displaystyle\Big(\int_{K_k} |f|\Big)$. On déduit alors des inégalités ci-dessus par passage à la limite (en utilisant a)) : $|S - S'| = 0$, c'est-à-dire $S = S'$.

La proposition ci-dessus justifie la définition suivante.

Proposition et définition 62. Soient $A \subset \mathbb{R}^n$ et une application $f : A \to \mathbb{R}$. On dit que f est *intégrable au sens de Riemann généralisé* sur A (ou que l'intégrale impropre $\displaystyle\int_A f$ converge) s'il existe une suite exhaustive de compacts $(K_k)_{k\in\mathbb{N}}$ de A sur laquelle f est intégrable au sens de Riemann et telle que la suite $\displaystyle\Big(\int_{K_k} |f|\Big)_{k\in\mathbb{N}}$ soit convergente. La suite $\displaystyle\Big(\int_{K_k} f\Big)_{k\in\mathbb{N}}$ est alors convergente et sa limite est appelée intégrale de f sur A et notée $\displaystyle\int_A f$.

Si f est *intégrable au sens de Riemann généralisé* sur A, $|f|$ l'est aussi.
La proposition et définition ci-dessus généralise la définition introduite pour la dimension $n = 1$ dans le module *Intégration* du livre [L2]. En dimension $n = 1$ les seules suites de compacts utilisées étaient des suites de *segments*.
On montre facilement le résultat suivant.

Proposition 63. Soit $A \subset \mathbb{R}^n$ admettant une suite exhaustive de compacts, alors l'ensemble des applications de A dans $\mathbb{R}$ qui admettent une intégrale au sens de Riemann généralisé est un sous-espace vectoriel de $\mathcal{F}(A)$ et l'application $f \mapsto \displaystyle\int_A f$ de cet espace dans $\mathbb{R}$ est linéaire et positive.

Exemple. Soient $a \in \mathbb{R}$ et $b \in \mathbb{R} \cup \{+\infty\}$, $a < b$ et $f : [a, b[\to \mathbb{R}_+$ une fonction *localement intégrable* sur $[a, b[$. On note $A := \{(x, y) \in [a, b[\times \mathbb{R}_+ \mid 0 \leqslant y \leqslant f(x)\}$.
On va montrer que les deux conditions suivantes sont équivalentes.

(i) La fonction f est intégrable au sens de Riemann généralisé sur $[a, b[$;

(ii) La fonction 1 est intégrable au sens de Riemann généralisé sur A.

S'il en est ainsi $\displaystyle\int_{[a,b[} f = \int_A 1$. On dit alors, comme dans le cas des fonctions intégrables sur un segment, que $\displaystyle\int_A 1$ est l'aire comprise entre l'axe des abscisses et le graphe de f et l'on note $\operatorname{vol} A := \displaystyle\int_A 1$ (ou mieux, puisque $n = 2$, aire $A := \displaystyle\int_A 1$).
Soit $(b_k)_{k\in\mathbb{N}}$ une suite strictement croissante de $]a, b[$ convergeant vers b quand $k \to +\infty$. On note $A_k := \{(x, y) \in [a, b_k] \times \mathbb{R}_+ \mid 0 \leqslant y \leqslant f(x)\}$. Les A_k sont des compacts *cubables* (*cf.* la proposition 37 de la page 200) et l'on a $\displaystyle\int_{A_k} 1 = \int_{[a,b_k]} f$. La suite $(A_k)_{k\in\mathbb{N}}$ est une suite exhaustive de compacts de A. En effet, si $K \subset A$ est compact, sa projection sur $[a, b[$ est compacte donc contenue dans un segment $[a, c]$ avec $c < b$. On a :

$$\lim_{k\to+\infty} \int_{A_k} 1 = \lim_{k\to+\infty} \int_{[a,b_k]} f.$$

Le résultat s'en déduit.

Proposition 64. Soient $A \subset \mathbb{R}^n$ une partie *cubable* et une application $f : A \to \mathbb{R}$. On suppose que f est intégrable au sens de Riemann sur A et qu'il existe une suite exhaustive de compacts cubables $(K_k)_{k\in\mathbb{N}}$ de A sur laquelle f est intégrable au sens de Riemann. Alors f est intégrable au sens de Riemann généralisé sur A et les intégrales au sens de Riemann et au sens de Riemann généralisé coïncident.

Démonstration. Notons $I := \displaystyle\int_A f$ et $J := \displaystyle\int_A |f|$, les intégrales étant prises *au sens de Riemann*. Pour tout $k \in \mathbb{N}$, $\displaystyle\int_{K_k} |f| \leqslant J$, donc f est intégrable au sens de Riemann généralisé sur A.
La fonction f est bornée sur A : il existe $M > 0$ tel que $|f| \leqslant M$. On a $\operatorname{vol} A = \lambda^-(A)$, donc, puisque les K_k sont cubables, $\operatorname{vol} A = \displaystyle\lim_{k\to+\infty} \operatorname{vol} K_k$. On a :

$$\forall k \in \mathbb{N}, \quad \left| I - \int_{K_k} f \right| \leqslant J - \int_{K_k} |f| \leqslant M(\operatorname{vol} A - \operatorname{vol} K_k).$$

Donc $\displaystyle\lim_{k\to+\infty} \int_{K_k} f = I$. ∎

1.5.3 Un critère d'existence des intégrales généralisées

Proposition 65. Soient $A \subset \mathbb{R}^n$ et une application $f : A \to \mathbb{R}$ localement intégrable sur A. Alors f admet une intégrale au sens de Riemann généralisé sur A si, et seulement s'il existe une fonction $g : A \to \mathbb{R}_+$ admettant une intégrale au sens de Riemann généralisé sur A et telle que $|f| \leqslant g$.

Démonstration. La condition est évidemment nécessaire : il suffit de prendre $g := |f|$. Montrons qu'elle est suffisante. Supposons-la satisfaite, il existe une suite exhaustive de compacts $(K'_k)_{k\in\mathbb{N}}$ de A sur laquelle g est intégrable au sens de Riemann et qui est telle que la suite $\left(\int_{K'_k} g\right)_{k\in\mathbb{N}}$ soit convergente. Notons Σ sa limite.

Par hypothèse la fonction f est intégrable sur *une* suite exhaustive de compacts $(K_k)_{k\in\mathbb{N}}$ de A. Soit $k \in \mathbb{N}$; il existe $h \in \mathbb{N}$ tel que $K_k \subset K'_h$. Soit $\mathbf{P}$ un pavé fermé de $\mathbb{R}^n$ tel que $K'_h \subset \mathbf{P}$. On a $|\tilde{f}|1_{K_k} \leqslant \tilde{g}1_{K'_h}$, donc $\displaystyle\int_{K_k} |f| = \int_{\mathbf{P}} |\tilde{f}|1_{K_k} \leqslant \int_{\mathbf{P}} \tilde{g}1_{K'_h} = \int_{K'_h} g$.

On en déduit que la suite croissante $\left(\int_{K_k} |f|\right)_{k\in\mathbb{N}}$ est majorée par Σ ; elle est donc convergente. ∎

Corollaire 66 (Règle de comparaison).
Soient $A \subset \mathbb{R}^n$ et $f,\ g : A \to \mathbb{R}_+$ deux applications localement intégrables sur A telles que $0 \leqslant f \leqslant g$. Alors :

(i) si g admet une intégrale au sens de Riemann généralisé sur A, il en est de même de f ;

(i) si f n'admet pas d'intégrale au sens de Riemann généralisé sur A, il en est de même de g.

Proposition 67. Soient $A \subset \mathbb{R}^n$ un sous-ensemble admettant une suite exhaustive de compacts *cubables* $(K_k)_{k\in\mathbb{N}^*}$ et une application $f : A \to \mathbb{R}$ *continue* sur A.

(i) On suppose que la suite numérique $\left(\int_{K_k} |f|\right)_{k\in\mathbb{N}}$ est *bornée* par $C > 0$. Alors f admet une intégrale au sens de Riemann généralisé sur A et $\left|\int_A f\right| \leqslant C$.

(ii) On suppose f bornée par $M > 0$ et A bornée : $A \subset \mathbf{P}$ (où $\mathbf{P}$ est un pavé fermé). Alors f admet une intégrale au sens de Riemann généralisée sur A et $\left|\int_A f\right| \leqslant M \operatorname{vol} P$.

(iii) Soit $B \subset A$ admettant une suite exhaustive de compacts *cubables*. Si f est *positive* et admet une intégrale au sens de Riemann généralisé sur A, alors elle admet une intégrale au sens de Riemann généralisé sur B et $\displaystyle\int_B f \leqslant \int_A f$.

Démonstration.

(i) La suite $\left(\int_{K_k} |f|\right)_{k\in\mathbb{N}}$ est croissante et majorée par C, elle est donc convergente. Ainsi, f est intégrable au sens de Riemann généralisé sur A et l'on a :

$$\left|\int_A f\right| \leqslant \int_A |f| \leqslant C.$$

(ii) On a :

$$\forall\, k \in \mathbb{N}^*,\ K_k \subset A \subset \mathbf{P}, \quad \text{donc} \quad \left|\int_{K_k} f\right| \leqslant \int_{K_k} |f| \leqslant M \operatorname{vol} K_k \leqslant M \operatorname{vol} P.$$

On termine la preuve en utilisant (i).

(iii) Si $K \subset B$ est un compact cubable, on a $\int_K f \leqslant \int_A f$.

On termine en appliquant (i) à la restriction de f à B avec $C := \int_A f$. ■

Si $U \subset \mathbb{R}^n$ est un ouvert, on peut montrer qu'il admet une suite exhaustive de compacts *pavables*. Nous l'admettrons provisoirement (*cf.* la proposition 186 de la page 347).

Exercice 12.

Soit $U \subset \mathbb{R}^n$ un ouvert *borné*. Montrer que la fonction constante 1 est intégrable au sens de Riemann généralisé sur U et que $\int_U 1 = \lambda^-(U)$.

Solution. On a, par définition, $\lambda^-(U) := \sup_C \operatorname{vol} C = \sup_C \int_C 1$, les sup portant sur les parties pavables $C \subset A$. Ainsi, pour toute partie pavable $C \subset A$, $\int_C 1 \leqslant \lambda^-(U)$, donc 1 est intégrable au sens de Riemann généralisé sur U d'après la proposition ci-dessus et l'on a $\int_U 1 \leqslant \lambda^-(U)$.

Soit $(K_k)_{k\in\mathbb{N}}$ une suite exhaustive de compacts pavables de U. On a $\int_U 1 = \lim_{k\to+\infty} \int_{K_k} 1$. Par ailleurs, si $C \subset A$ est une partie pavable, il existe $m \in \mathbb{N}$ tel que $C \subset K_m$, donc $\int_C 1 \leqslant \int_{K_m} 1$. On en déduit $\int_U 1 = \lambda^-(U)$.

Soit $U \subset \mathbb{R}^n$ un ouvert borné. Si 1_U est $\mathscr{R}$-intégrable sur $\mathbb{R}^n$, on a $\lambda^+(U) = \lambda^-(U) = \int 1_U$ (*cf.* le corollaire 35 de la page 199), donc $\int 1_U = \int_U 1$, la deuxième intégrale étant prise au sens de Riemann généralisé. Il est, en revanche, possible que 1_U ne soit pas $\mathscr{R}$-intégrable, c'est-à-dire que $\lambda^+(U) > \lambda^-(U)$. Nous en avons vu un exemple en dimension $n = 1$ dans le module *Intégration* du livre [L2] (*cf.* l'exercice II.3.5)

et nous en donnons d'autres ici dans l'exercice type corrigé III.1.0. Ainsi l'intégrabilité au sens de Riemann de la fonction caractéristique 1_U sur un pavé fermé contenant A *n'est pas équivalente* à l'intégrabilité au sens de Riemann généralisé de la fonction 1 sur U (qui est toujours vraie). Nous verrons que la situation est différente dans le cadre de l'intégrale de Henstock-Kurzweil (*cf.* la partie 4.3).

Exercice 13.

(i) Soient $U \subset \mathbb{R}^n$ un *ouvert* et $(U_h)_{h\in\mathbb{N}}$ une suite croissante *d'ouverts cubables* contenus dans U telle que $U = \bigcup\limits_{h\in\mathbb{N}} U_h$. Soit $f : U \to \mathbb{R}$ une fonction *continue positive* intégrable sur U au sens de Riemann généralisé. Montrer que f est intégrable sur les U_h et que $\displaystyle\int_U f = \lim_{h\to+\infty} \int_{U_h} f$.

(ii) Soient $L := [0,1[^2$, $U := \operatorname{Int} L =]0,1[^2$ et $f : \Delta \to \mathbb{R}$ une fonction *continue positive* intégrable au sens de Riemann généralisé sur L. Montrer que f est intégrable au sens de Riemann généralisé sur U et que $\displaystyle\int_U f = \int_L f$.

Solution.

(i) Si $K \subset U_h \subset U$ est un compact *pavable* $\displaystyle\int_K f \leqslant \int_U f$, on en déduit que f est intégrable sur U_h et que $\displaystyle\int_{U_h} f \leqslant \int_U f$.

Il existe dans U une suite exhaustive $(K_k)_{k\in\mathbb{N}}$ de compacts *pavables*. Soit $K \subset U$ un compact. La famille $\{U_h\}_{h\in\mathbb{N}}$ est un recouvrement ouvert de K, on peut en extraire un sous-recouvrement fini et, puisque la suite $(U_h)_{h\in\mathbb{N}}$ est croissante, il existe $h(K) \in \mathbb{N}$ tel que $K \subset U_{h(K)}$. Si, de plus, K est cubable, on a $\displaystyle\int_K f \leqslant \int_{U_{h(K)}} f$. Notons, pour tout $k \in \mathbb{N}$, $h_k := h(K_k)$, on a :

$$\int_U f = \lim_{k\to+\infty} \int_{K_k} f \leqslant \lim_{k\to+\infty} \int_{U_{h_k}} f \leqslant \lim_{h\to+\infty} \int_{U_h} f \leqslant \int_U f.$$

Ainsi $\displaystyle\int_U f = \lim_{h\to+\infty} \int_{U_h} f$.

(i) Pour tout $h \in \mathbb{N}^*$, on note $L_h := [0, 1-1/h[^2$ et $U_h := \operatorname{Int} L_h =]0, 1-1/h[^2$. La suite $(L_h)_{h\in\mathbb{N}^*}$ est une suite exhaustive de compacts (pavés fermés) de L et l'on a $\bigcup\limits_{h\in\mathbb{N}^*} U_h = U$.

Pour tout compact cubable $K \subset U \subset \Delta$, l'on a $\displaystyle\int_K f \leqslant \int_\Delta f$, donc f est intégrable au sens de Riemann généralisé sur U.

D'après la proposition 50 de la page 207, pour tout $h \in \mathbb{N}^*$, $\displaystyle\int_{L_h} f = \int_{U_h} f$.

En utilisant (i), on en déduit :

$$\int_L f = \lim_{h\to+\infty} \int_{L_h} f = \lim_{h\to+\infty} \int_{U_h} f = \int_U f.$$

Exercice 14.

On rappelle $\zeta(2) := \sum\limits_{k\in\mathbb{N}^*} 1/k^2$. Pour tout $0 < a < 1$, on note :

$$I(a) := \int_{[0,a]^2} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy} \qquad I_1(a) := \int_{[0,1]\times[0,a]} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy} \quad \text{et} \quad I_2(a) := \int_{[0,a]\times[0,1]} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy}.$$

(i) Montrer que $\lim\limits_{a\to1^-} I(a) = \zeta(2)$. En déduire que l'intégrale $I := \displaystyle\int_{[0,1[^2} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy}$ existe au sens de Riemann généralisé et la calculer. Montrer que $I = \displaystyle\int_{]0,1[^2} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy}$.

(ii) Montrer que $I_1(a) = I_2(a)$ et que $\lim\limits_{a\to1^-} I_1(a) = I$.

(iii) On note $\Delta := [0,1]^2 \setminus \{(1,1)\}$ et :

$\forall\, k \in \mathbb{N}^*, \quad C'_k := [0,1]\times[0,1-1/k]\,,\ C''_k := [0,1-1/k]\times[0,1]\,, C_k := C'_k \cup C''_k.$

Montrer que la fonction $(x,y) \mapsto \frac{1}{1-xy}$ est intégrable sur la suite de compacts pavables $(C_k)_{k\in\mathbb{N}^*}$, puis que cette fonction est intégrable au sens de Riemann généralisé sur Δ et que $I := \displaystyle\int_\Delta \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy}$.

Solution.

(i) Si $0 \leqslant x \leqslant a$ et $0 \leqslant y \leqslant a$, on a $\frac{1}{1-xy} = \sum\limits_{k\in\mathbb{N}} (xy)^k$. Des égalités :

$$\int_{[0,a]^2} (xy)^k\,\mathrm{d}x\,\mathrm{d}y = \left(\int_0^a x^k\ \mathrm{d}x\right)^2 = \frac{a^{2k+2}}{(k+1)^2} \quad \text{et} \quad \sum_{k\in\mathbb{N}} \frac{a^{2k+2}}{(k+1)^2} \leqslant \zeta(2).$$

l'on déduit

$$\begin{aligned} I(a) &= \sum_{k\in[\![0,p]\!]} \int_{[0,a]^2} (xy)^k\,\mathrm{d}x\,\mathrm{d}y + \int_{[0,a]^2} \frac{(xy)^{p+1}}{1-xy}\,\mathrm{d}x\,\mathrm{d}y \\ &= \sum_{k\in[\![0,p]\!]} \frac{a^{2k+2}}{(k+1)^2} + \int_{[0,a]^2} \frac{(xy)^{p+1}}{1-xy}\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

De $\left|\displaystyle\int_{[0,a]^2} \frac{(xy)^{p+1}}{1-xy}\,\mathrm{d}x\,\mathrm{d}y\right| \leqslant \dfrac{a^{2p+4}}{1-a^2}$, l'on déduit alors $I(a) = \displaystyle\sum_{k\in\mathbb{N}} \frac{a^{2k+2}}{(k+1)^2}$ et $I(a) \leqslant \zeta(2)$. Donc, d'après la proposition 67, l'intégrale I existe et $I \leqslant \zeta(2)$.

La limite de $I(a)$ quand a tend vers 1^- est $\zeta(2)$ (d'après la convergence normale de la série), donc $I = \zeta(2)$.

On a $I = \displaystyle\int_{]0,1[^2} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy}$ d'après la question (ii) de l'exercice ci-dessus.

(ii) En intervertissant x et y, on montre $I_1(a) = I_2(a)$.

On a $I_1(a) = \left(\displaystyle\int_0^1 x^k\,\mathrm{d}x\right)\left(\displaystyle\int_0^a x^k\,\mathrm{d}x\right) = \dfrac{a^{k+1}}{(k+1)^2}$.

On en déduit $\displaystyle\lim_{a\to 1^-} I_1(a) = \zeta(2) = I$.

(iii) On a vu dans l'exercice 11 de la page 226 que $(C_k)_{k\in\mathbb{N}}$ est une suite exhaustive de compacts pavables.

On a $C'_k \cap C''_k = [0, 1-1/k]^2$, donc :

$$\int_{C_k} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy} = \int_{C'_k} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy} + \int_{C''_k} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy} - I(1-1/k) = 2I_1(1-1/k) - I(1-1/k).$$

On en déduit $\displaystyle\lim_{a\to 1^-} \int_{C_k} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy} = 2I - I = I$ et $\displaystyle\int_\Delta \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy} = I$.

Exercice 15.

Le but de l'exercice est de donner une preuve de la formule $\zeta(2) = \pi^2/6$ (la solution du *problème de Bâle*, posé en 1644 par P. Mengoli et résolu par L. Euler en 1735) en utilisant la formule de changement de variables pour les intégrales doubles.

On utilise les résultats de l'exercice précédent *dont on reprend les notations*.

(i) Montrer que les intégrales :

$$J_1 := \int_0^{1/2} \mathrm{d}u \int_0^u \frac{\mathrm{d}v}{1-u^2+v^2} \qquad \text{et} \qquad J_2 := \int_{1/2}^1 \mathrm{d}u \int_0^u \frac{\mathrm{d}v}{1-u^2+v^2}$$

existent en un sens que l'on précisera et les calculer.

(ii) En utilisant le changement de variables $x := u+v$, $y := u-v$, montrer que :

$$I/4 = J_1 + J_2 \qquad \text{et en déduire} \qquad \zeta(2) = \pi^2/6.$$

Solution.

La fonction $(u,v) \mapsto \frac{1}{1-u^2+v^2}$ est continue sur $[0,1]^2 \setminus (1,0)$.

(i) On définit les deux domaines triangulaires :

$$F_1 := \{(u,v) \in \mathbb{R}_+^2 \mid v \leqslant u \leqslant 1/2\}$$
$$F_2 := \{(u,v) \in \mathbb{R}_+^2 \mid u+v \leqslant 1,\ 1/2 \leqslant u \leqslant 1\}.$$

La fonction $(u,v) \mapsto \frac{1}{1-u^2+v^2}$ est continue sur le compact cubable F_1, donc elle est $\mathscr{R}$-intégrable sur F_1 et l'on a, d'après le théorème 56 de la page 213 :

$$\int_{F_1} \frac{\mathrm{d}u\,\mathrm{d}v}{1-u^2+v^2} = \int_0^{1/2} \mathrm{d}u \int_0^u \frac{\mathrm{d}v}{1-u^2+v^2} = J_1.$$

Calculons J_1. Pour tout $c > 0$, on a $\displaystyle\int_0^u \frac{\mathrm{d}v}{c^2+v^2} = \frac{1}{c}\int_0^{u/c} \frac{\mathrm{d}v}{1+v^2} = \frac{1}{c}\arctan\frac{u}{c}\cdot$

On en déduit $J_1 = \displaystyle\int_0^{1/2} \arctan\frac{u}{\sqrt{1-u^2}}\frac{\mathrm{d}u}{\sqrt{1-u^2}}$, puis, en utilisant le changement de variables $u := \sin\theta$, $J_1 = \displaystyle\int_0^{\pi/6} \theta\,\mathrm{d}\theta = \frac{1}{2}\left[\theta^2\right]_0^{\pi/6} = \pi^2/72$.

Pour J_2 la situation est un peu plus délicate car la fonction $(u,v) \mapsto \frac{1}{1-u^2+v^2}$ n'est pas continue sur le compact F_2 : elle n'est pas bornée au voisinage du point $(1,0) \in F_2$. Il faut donc utiliser des intégrales généralisées.

On note :

$$\forall\, 1/2 < b < 1, \quad F_2(b) := \{(u,v) \in \mathbb{R}_+^2 \mid u+v \leqslant 1,\ 1/2 \leqslant u \leqslant b\}.$$

La fonction $(u,v) \mapsto \frac{1}{1-u^2+v^2}$ est continue sur la compact cubable $F_2(b)$ (en effet, on a $1-u^2+v^2 \geqslant 1-b^2$ sur $F_2(b)$), donc elle est $\mathscr{R}$-intégrable sur $F_2(b)$ et l'on a, d'après le théorème 56 de la page 213 :

$$\int_{F_2(b)} \frac{\mathrm{d}u\,\mathrm{d}v}{1-u^2+v^2} = \int_{1/2}^b \mathrm{d}u \int_0^{1-u} \frac{\mathrm{d}v}{1-u^2+v^2}\cdot$$

On note $J_2(b)$ cette intégrale et on l'interprète comme une intégrale *simple* en u sur $[0,b]$.

Calculons $J_2(b)$. On a $J_2(b) = \displaystyle\int_{1/2}^b \frac{1}{\sqrt{1-u^2}} \arctan\sqrt{\frac{1-u}{1+u}}\,\mathrm{d}u$, puis, en utilisant le changement de variables $u := \cos(2\theta)$, $J_2(b) = 2\displaystyle\int_{\frac{1}{2}\arccos b}^{\pi/6} \theta\,\mathrm{d}\theta$. Par suite l'intégrale *simple* généralisée J_2 est convergente sur $[1/2,1[$ et $J_2 = \lim\limits_{b\to 1^-} J_2(b) = \pi^2/36$ (en fait J_2 s'interprète comme une intégrale de Riemann sur $[1/2,1]$ en prolongeant par continuité la fonction à intégrer en $u=1$). On a $J_1 + J_2 = \pi^2/24$.

(ii) On ne peut pas appliquer directement la formule de changement de variables car la fonction $(x,y) \mapsto \frac{1}{1-xy}$ n'est pas bornée au voisinage du point $(1,1)$.

On note $\Delta := [0,1]^2 \setminus \{(1,1)\}$ et, pour tout $k \in \mathbb{N}^*$:

$$K_k := \{(x,y) \in \Delta \mid x+y \leqslant 2-1/k\}, \quad H_k := \int_{K_k} \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy}\cdot$$

D'après les résultats de l'exercice 11 de la page 226, $(K_k)_{k\in\mathbb{N}^*}$ est une suite exhaustive de compacts cubables de Δ. D'après les résultats de la question (iii) de l'exercice précédent on a $\lim\limits_{k\to+\infty} H_k = \displaystyle\int_\Delta \frac{\mathrm{d}x\,\mathrm{d}y}{1-xy} = I$.

Le changement de variables $x := u+v$, $y := u-v$ est *linéaire* bijectif, c'est un difféomorphisme de classe $\mathcal{C}^1$ de $\mathbb{R}^2$ sur $\mathbb{R}^2$ de jacobien $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = -2$. Il transforme le pavé fermé $[0,1]^2$ en la réunion des triangles F_1, F_2 et de leurs symétriques par rapport à l'axe des v. Pour tout $k \geqslant 2$, il transforme K_k en la réunion des domaines F_1, $F_2(1-1/2k)$ et de leurs symétriques par rapport à l'axe des v. En utilisant un argument de symétrie ($v \mapsto -v$)

et le théorème de changement de variables on en déduit $H_k = 4\big(J_1 + J_2(1-1/2k)\big)$. Par suite :

$$I = \lim_{k\to+\infty} H_k = \lim_{k\to+\infty} 4\big(J_1 + J_2(1-1/2k)\big) = 4(J_1+J_2) = 4\pi^2/24 = \pi^2/6.$$

Ainsi $\zeta(2) = \pi^2/6$.

Remarque. Euler, après avoir deviné la formule $\zeta(2) = \pi^2/6$, en a donné, plusieurs années plus tard, trois preuves différentes. Nous ignorons l'origine de la preuve citée dans l'exercice ci-dessus[3] qui semble avoir été redécouverte plusieurs fois par divers auteurs… Il existe d'autres preuves de la formule $\zeta(2) = \pi^2/6$ également basées sur des changements de variables dans des intégrales doubles (*cf.* les exercices III.1.19, III.1.20 et III.1.22) ou sur la formule de Fubini (*cf.* l'exercice III.1.21). Il existe aussi des preuves utilisant de tout autres outils, comme par exemple les séries de Fourier grâce à la formule de Parseval (*cf.* le module *Séries de Fourier* du livre [L2]).

Exercice 16.

Soient $t > 0$ et $\Delta := \{(x,y) \in \mathbb{R}^2 \mid 0 \leqslant x \leqslant t,\ y \geqslant 0\}$.

Montrer que l'intégrale généralisée $\displaystyle\iint_\Delta \mathrm{e}^{-xy} \sin x \,\mathrm{d}x\,\mathrm{d}y$ existe et la calculer de deux façons. En déduire la valeur de $\displaystyle\int_0^{+\infty} \frac{\sin x}{x}\,\mathrm{d}x$.

Solution. Notons :

$$f : (x,y) \in \mathbb{R}_+^2 \mapsto \mathrm{e}^{-xy}\sin x \in \mathbb{R} \quad \text{et} \quad g : (x,y) \in \mathbb{R}_+^2 \mapsto x\mathrm{e}^{-xy} \in \mathbb{R}.$$

On a $|f| \leqslant g$. Pour montrer que f admet une intégrale de Riemann généralisée sur $\Delta \subset \mathbb{R}_+^2$ il suffit de montrer que g admet une intégrale de Riemann généralisée sur Δ.
Les fonctions f et g sont continues sur Δ.
Notons, pour tout $k \in \mathbb{N}^*$, $K_k := [0,t] \times [0,k]$, $(K_k)_{k\in\mathbb{N}^*}$ est une suite exhaustive de compacts de Δ. Calculons $J_k := \displaystyle\int_{K_k} g$ en utilisant le théorème de Fubini :

$$J_k = \int_0^t \mathrm{d}x \int_0^k \mathrm{e}^{-xy}\, x\,\mathrm{d}y = \int_0^t (1-\mathrm{e}^{-kx})\,\mathrm{d}x = t + \frac{\mathrm{e}^{-kt}-1}{k}.$$

Le réel $t > 0$ étant fixé, $\displaystyle\lim_{k\to+\infty} \frac{\mathrm{e}^{-kt}-1}{k} = 0$, donc la suite (J_k) est convergente. Ainsi g admet une intégrale de Riemann généralisée sur Δ et il en est de même de f.

Notons $I_k(t) := \displaystyle\int_{K_k} f$. On peut calculer $I(t) := \displaystyle\int_\Delta f = \lim_{k\to+\infty} f$ de deux façons en utilisant le théorème de Fubini pour $I_k(t)$ et en passant à la limite :

$$I_k(t) = \int_0^k \mathrm{d}y \int_0^t \mathrm{e}^{-xy}\sin x\,\mathrm{d}x = \int_0^t \sin x\,\mathrm{d}x \int_0^k \mathrm{e}^{-xy}\,\mathrm{d}y.$$

---

[3] Cf. William Judson LeVeque, *Topics in Number Theory*, 1956.

On a d'une part :

$$\int_0^t \mathrm{e}^{-xy}\sin x\ \mathrm{d}x = \mathrm{Im}\left(\int_0^t \mathrm{e}^{-(y-i)x}\ \mathrm{d}x\right) = \mathrm{Im}\left(\frac{1-\mathrm{e}^{-(y-i)t}}{y-i}\right)$$
$$= \frac{1}{y^2+1}(1-\cos t\,\mathrm{e}^{-ty} - y\sin t\,\mathrm{e}^{-ty}).$$

D'où :

$$I_k(t) = \int_0^k \mathrm{d}y \int_0^t \mathrm{e}^{-xy}\sin x\ \mathrm{d}x = \int_0^k \frac{\mathrm{d}y}{y^2+1} - \int_0^k \frac{\cos t + y\sin t}{y^2+1}\mathrm{e}^{-ty}\ \mathrm{d}y.$$

La fonction $y \mapsto \frac{\cos t + y\sin t}{y^2+1}\mathrm{e}^{-ty}$ est intégrable au sens de Riemann généralisé sur $[0,+\infty[$.

On note $J(t) := \displaystyle\int_0^{+\infty} \frac{\cos t + y\sin t}{y^2+1}\mathrm{e}^{-ty}\ \mathrm{d}y$.

On vérifie l'inégalité $\sup\limits_{y>0} \frac{y+1}{y^2+1} \leqslant 2$ qui donne :

$$\left|\int_0^k \frac{\cos t + y\sin t}{y^2+1}\,\mathrm{e}^{-ty}\ \mathrm{d}y\right| \leqslant \int_0^k \frac{y+1}{y^2+1}\,\mathrm{e}^{-ty}\ \mathrm{d}y \leqslant 2\int_0^k \mathrm{e}^{-ty}\ \mathrm{d}y = 2\,\frac{1-\mathrm{e}^{-tk}}{t} \leqslant \frac{2}{t}\cdot$$

On en déduit :

$$I(t) = \lim_{k\to+\infty} I_k = \frac{\pi}{2} - J(t), \quad \text{avec} \quad \big|J(t)\big| \leqslant 2/t.$$

On a d'autre part :

$$I_k(t) = \int_0^t \mathrm{d}x \int_0^k \mathrm{e}^{-xy}\sin x\ \mathrm{d}y = \int_0^t \frac{\sin x}{x}(1-\mathrm{e}^{-kx})\ \mathrm{d}x$$
$$= \int_0^t \frac{\sin x}{x}\ \mathrm{d}x - \int_0^t \frac{\sin x}{x}\mathrm{e}^{-kx}\ \mathrm{d}x.$$

De $\left|\displaystyle\int_0^t \frac{\sin x}{x}\mathrm{e}^{-kx}\ \mathrm{d}x\right| \leqslant \displaystyle\int_0^t \mathrm{e}^{-kx}\ \mathrm{d}x = \frac{1}{k}$, on déduit :

$$I(t) = \lim_{k\to+\infty} I_k = \int_0^t \frac{\sin x}{x}\ \mathrm{d}x.$$

Par suite $\displaystyle\int_0^t \frac{\sin x}{x}\ \mathrm{d}x = \frac{\pi}{2} - J(t)$. Quand t tend vers $+\infty$, $J(t)$ tend vers 0, donc $\displaystyle\lim_{t\to+\infty}\int_0^t \frac{\sin x}{x}\ \mathrm{d}x$ existe et l'on a $\displaystyle\int_0^{+\infty} \frac{\sin x}{x}\ \mathrm{d}x = \frac{\pi}{2}\cdot$

On peut aussi déduire ce résultat de la solution de l'exercice II.3.35 du livre [L2]).

Pour faire un changement de variables dans une intégrale généralisée on peut, comme nous l'avons vu plus haut, utiliser le théorème 58 de la page 215 (ou la proposition 59) sur des sous-ensembles convenables puis passer à la limite (*cf.* ci-dessous 1.5.4). Cette méthode *élémentaire* (que nous conseillons au débutant) peut être délicate à mettre en œuvre ou le

lecteur peut être paresseux, il pourra alors utiliser le résultat suivant que nous admettrons provisoirement. Nous le prouverons dans la section 4 en utilisant l'intégrale de Henstock-Kurzweil à plusieurs variables (*cf.* le corollaire 192 de la page 357).

Proposition 68. Soient U_1 et U_2 deux ouverts de $\mathbb{R}^n$ et $\mathbf{h} : U_1 \to U_2$ un difféomorphisme de classe $\mathcal{C}^1$. Une fonction continue $f : U_2 \to \mathbb{R}^n$ est intégrable au sens de Riemann généralisé sur U_2 si, et seulement si, $f \circ \mathbf{h} \cdot |\operatorname{Jac} \mathbf{h}|$ est intégrable au sens de Riemann généralisé sur U_1. Dans ce cas :

$$\int_{U_2} f(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_{U_1} f \circ \mathbf{h}(\mathbf{x}) \cdot \big|\operatorname{Jac} \mathbf{h}(\mathbf{x})\big|\,\mathrm{d}\mathbf{x}.$$

1.5.4 Quelques applications des intégrales doubles généralisées

Intégrale de Gauß– Fonctions Gamma et Beta

Nous proposons d'abord deux exercices d'application. Le résultat du premier exercice est aussi en fait une conséquence de celui du second.

Exercice 17.
En passant en coordonnées polaires, calculer l'intégrale sur $\Delta := [0, +\infty[\times [0, +\infty[$ de la fonction $f : (x, y) \mapsto \mathrm{e}^{-(x^2+y^2)}$. En déduire que $\displaystyle\int_{-\infty}^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = \sqrt{\pi}$.

Solution. Observons que la fonction $t \mapsto e^{-t^2}$ est continue, donc localement intégrable sur $[0, +\infty[$. En $+\infty$, on a $\mathrm{e}^{-t^2} = \mathrm{o}(1/t^2)$ et donc, l'intégrale $\displaystyle\int_0^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t$ est convergente.
La fonction $f : (x, y) \mapsto \mathrm{e}^{-(x^2+y^2)}$ est continue et positive sur Δ. Soit, pour tout $k \in \mathbb{N}$, $\mathbf{P}_k := [0, k] \times [0, k] \subset \Delta$, $(\mathbf{P}_k)_{k\in\mathbb{N}}$ est une suite exhaustive de compacts de Δ et, d'après le théorème de Fubini :

$$\int_{\mathbf{P}_k} f = \left(\int_0^k \mathrm{e}^{-x^2}\,\mathrm{d}x\right)\left(\int_0^k \mathrm{e}^{-y^2}\,\mathrm{d}y\right) \leqslant I := \left(\int_0^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t\right)^2 < +\infty.$$

La suite $\displaystyle\left(\int_{\mathbf{P}_k} f\right)_{k\in\mathbb{N}}$ est croissante et majorée par I, elle est donc convergente. Par suite la fonction f est absolument intégrable sur Δ. Par passage à la limite quand k tend vers $+\infty$, on a :

$$\int_\Delta f = \left(\lim_{k\to+\infty} \int_0^k \mathrm{e}^{-x^2}\,\mathrm{d}x\right)\left(\lim_{k\to+\infty} \int_0^k \mathrm{e}^{-y^2}\,\mathrm{d}y\right) = \left(\int_0^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t\right)^2. \tag{15}$$

Passons en coordonnées polaires : $x = \rho\cos\theta$ et $y = \rho\sin\theta$. Soit, pour tout $k \in \mathbb{N}^*$, $V_k := \big\{(\rho\cos\theta, \rho\sin\theta) \;\big|\; 0 < \theta < \pi/2,\ 0 < \rho < k\big\}$, $(\overline{V_k})_{k\in\mathbb{N}^*}$ est une suite exhaustive de compacts de Δ et l'on a donc $\displaystyle\int_\Delta f = \lim_{k\to+\infty} I_k$ avec $\displaystyle I_k := \int_{\overline{V_k}} f = \int_{V_k} f$.

D'après le théorème de changement de variables :

$$I_k = \int_{V_k} f = \int_{(\rho,\theta)\in]0,k[\times]0,\pi/2[} \mathrm{e}^{-\rho^2}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta = \int_{(\rho,\theta)\in[0,k[\times[0,\pi/2[} \mathrm{e}^{-\rho^2}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta.$$

D'après le théorème de Fubini :

$$I_k = \int_0^k \mathrm{e}^{-\rho^2}\rho\,\mathrm{d}\rho \int_0^{\pi/2} \mathrm{d}\theta = \frac{1}{2}\frac{\pi}{2}\int_0^{k^2} \mathrm{e}^{-t}\,\mathrm{d}t = \frac{\pi}{4}(1-\mathrm{e}^{-k^2}).$$

Par suite $\displaystyle\int_\Delta f = \lim_{k\to+\infty} I_k = \frac{1}{2}\cdot\frac{\pi}{2} = \frac{\pi}{4}\cdot$ D'où, en utilisant (15) :

$$\left(\int_0^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t\right)^2 = \frac{\pi}{4} \quad \text{et} \quad \int_{-\infty}^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = 2\int_0^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = \sqrt{\pi}.$$

Exercice 18.

On rappelle que les fonctions Gamma et Beta d'Euler sont définies par :

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t \quad \text{et} \quad B(r,s) = \int_0^1 t^{r-1}(1-t)^{s-1}\,\mathrm{d}t$$

avec $x>0$, $r>0$ et $s>0$ (voir les exercices 16 et 17 du livre [L2], page 687).

1. Montrer que $\Gamma(x) = \displaystyle\int_0^{+\infty} 2\mathrm{e}^{-u^2}u^{2x-1}\,\mathrm{d}u$.
2. Posons $\Delta := [0,+\infty[\times[0,+\infty[$ et considérons la fonction $f:\Delta\to\mathbb{R}$ définie par :
$$f(u,v) = 4\mathrm{e}^{-(u^2+v^2)}u^{2r-1}v^{2s-1}.$$
Montrer que f est absolument intégrable sur Δ et que $\displaystyle\int_\Delta f = \Gamma(r)\Gamma(s)$.
3. En passant en coordonnées polaires, montrer que pour tout $r,s>0$, on a :
$$\Gamma(r)\Gamma(s) = 2\,\Gamma(r+s)\int_0^{\pi/2}(\cos\theta)^{2r-1}(\sin\theta)^{2s-1}\,\mathrm{d}\theta.$$
4. En faisant le changement de variables $t=\cos^2\theta$, montrer que :
$$B(r,s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}\cdot$$

Solution. Pour 2 et 3 on reprend les notations et les méthodes de la solution de l'exercice précédent.

1. Faire le changement de variable $t=u^2$.
2. La fonction f est continue et positive sur Δ. Soit $\mathbf{P}_k := [0,k]\times[0,k]\subset\Delta$, d'après le théorème de Fubini :
$$\int_{\mathbf{P}_k} f = \left(\int_0^k 2\mathrm{e}^{-u^2}u^{2r-1}\,\mathrm{d}u\right)\left(\int_0^k 2\mathrm{e}^{-v^2}v^{2s-1}\,\mathrm{d}v\right) \leqslant J := \Gamma(r)\Gamma(s).$$

On en déduit, en raisonnant comme dans l'exercice précédent, que la fonction f est absolument intégrable sur Δ. Puis, par passage à la limite quand k tend vers $+\infty$, on obtient :

$$\int_\Delta f = \left(\lim_{k\to+\infty}\int_0^k 2\mathrm{e}^{-u^2}u^{2r-1}\,\mathrm{d}u\right)\left(\lim_{k\to+\infty}\int_0^k 2\mathrm{e}^{-v^2}v^{2s-1}\,\mathrm{d}v\right) = \Gamma(r)\Gamma(s).$$

3. Passons en coordonnées polaires : $u = \rho\cos\theta$ et $v = \rho\sin\theta$.
D'après le théorème de changement de variables :

$$J_k = \int_{V_k} f = \int_{(\rho,\theta)\in]0,k[\times]0,\pi/2[} 4\mathrm{e}^{-\rho^2}\rho^{2(r+s)-2}(\cos\theta)^{2r-1}(\sin\theta)^{2s-1}\rho\,\mathrm{d}\rho\,\mathrm{d}\theta.$$

D'après le théorème de Fubini :

$$J_k = \left(\int_0^k 4\mathrm{e}^{-\rho^2}\rho^{2(r+s)-2}\rho\,\mathrm{d}\rho\right)\left(\int_0^{\pi/2}(\cos\theta)^{2r-1}(\sin\theta)^{2s-1}\,\mathrm{d}\theta\right).$$

Par suite :

$$\begin{aligned}\Gamma(r)\Gamma(s) &= \int_\Delta f = \lim_{k\to+\infty} J_k\\ &= \left(\int_0^{+\infty} 4e^{-\rho^2}\rho^{2(r+s)-2}\rho\,\mathrm{d}\rho\right)\left(\int_0^{\pi/2}(\cos\theta)^{2r-1}(\sin\theta)^{2s-1}\,\mathrm{d}\theta\right)\\ &= 2\,\Gamma(r+s)\int_0^{\pi/2}(\cos\theta)^{2r-1}(\sin\theta)^{2s-1}\,\mathrm{d}\theta.\end{aligned}$$

4. Le changement de variable $t = \cos^2\theta$ donne :

$$\Gamma(r)\Gamma(s) = \Gamma(r+s)\int_0^1 t^{r-1}(1-t)^{s-1}\,\mathrm{d}t = \Gamma(r+s)B(r,s).$$

Intégrales de Fresnel

Les intégrales doubles généralisées permettent de calculer les intégrales de Fresnel (obtenues par une autre méthode dans le module *Intégration* du livre [L2]). C'est l'objet de l'exercice suivant.

Exercice 19.

1. a) Montrer que, pour tout $t \geqslant 0$, l'intégrale $\displaystyle\int_0^{+\infty}\mathrm{e}^{(i-t)x^2}\,\mathrm{d}x$ est convergente. En déduire, pour tout $t \geqslant 0$, l'existence de :

$$F_2(t) := \int_0^{+\infty}\mathrm{e}^{-tx^2}\sin x^2\,\mathrm{d}x \quad\text{et}\quad G_2(t) := \int_0^{+\infty}\mathrm{e}^{-tx^2}\cos x^2\,\mathrm{d}x.$$

On rappelle que $F_2(0) := \displaystyle\int_0^{+\infty}\sin x^2\,\mathrm{d}x$ et $G_2(0) := \displaystyle\int_0^{+\infty}\cos x^2\,\mathrm{d}x$ sont les intégrales de Fresnel.

b) Montrer que, pour tout $t \geqslant 0$ et tout $a > 0$, $\left|\displaystyle\int_a^{+\infty} \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x\right| \leqslant \dfrac{1}{a}$.

2. Montrer que, pour tout $t > 0$, l'intégrale $\displaystyle\int_0^{+\infty} \mathrm{e}^{(i-t)x}\,\mathrm{d}x$ est convergente.

   En déduire l'existence et le calcul de :

$$F_1(t) := \int_0^{+\infty} \mathrm{e}^{-tx}\sin x\,\mathrm{d}x \qquad \text{et} \qquad G_1(t) := \int_0^{+\infty} \mathrm{e}^{-tx}\cos x\,\mathrm{d}x.$$

3. On note $\Delta := [0,+\infty[^2$. Pour tout $t > 0$ on définit des applications :

$$\begin{array}{rccl} f_t : & \Delta & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & \mathrm{e}^{-t(x^2+y^2)}\sin(x^2+y^2) \end{array}$$

$$\text{et} \quad \begin{array}{rccl} g_t : & \Delta & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & \mathrm{e}^{-t(x^2+y^2)}\cos(x^2+y^2). \end{array}$$

   a) Montrer que f_t et g_t sont $\mathscr{R}$-intégrables sur $\Delta := [0,+\infty[^2$.

   b) Calculer, pour tout $t > 0$, $\displaystyle\iint_\Delta f_t(x,y)\,\mathrm{d}x\,\mathrm{d}y$ et $\displaystyle\iint_\Delta g_t(x,y)\,\mathrm{d}x\,\mathrm{d}y$ en fonction d'une part de $F_1(t)$, $G_1(t)$ et d'autre part $F_2(t), G_2(t)$.
   Calculer $2F_2(t)G_2(t)$ et $G_2^2(t) - F_2^2(t)$ en fonction de t.

4. Montrer que $\displaystyle\lim_{t\to 0^+} F_2(t) = F_2(0)$ et $\displaystyle\lim_{t\to 0^+} G_2(t) = G_2(0)$.

5. Calculer les intégrales de Fresnel $F_2(0)$ et $G_2(0)$.

Solution.

1. a) Si $t > 0$, $|\mathrm{e}^{(i-t)x^2}| = \mathrm{e}^{-tx^2}$, donc l'intégrale est convergente.
   Si $t = 0$, on utilise $\left(\mathrm{e}^{ix^2}\right)' = 2ix\mathrm{e}^{ix^2}$.
   On a, pour tous $a > 0$ et $A > 0$, $a < A$, par intégration par parties ;

$$\int_a^A \mathrm{e}^{ix^2}\,\mathrm{d}x = \int_a^A \frac{\left(\mathrm{e}^{ix^2}\right)'}{2ix}\,\mathrm{d}x = \left[\frac{\left(\mathrm{e}^{ix^2}-1\right)}{2ix}\right]_a^A + \int_a^A \frac{\left(\mathrm{e}^{ix^2}-1\right)}{2ix^2}\,\mathrm{d}x.$$

   En passant à la limite quand a tend vers 0 et A tend vers $+\infty$, le second membre tend vers l'intégrale convergente $\displaystyle\frac{1}{2i}\int_0^{+\infty} \frac{\left(\mathrm{e}^{ix^2}-1\right)}{2ix^2}\,\mathrm{d}x$, l'intégrale $\displaystyle\int_0^{+\infty} \mathrm{e}^{ix^2}\,\mathrm{d}x$ est donc convergente et l'on a :

$$\int_0^{+\infty} \mathrm{e}^{ix^2}\,\mathrm{d}x = \frac{1}{2i}\int_0^{+\infty} \frac{\left(\mathrm{e}^{ix^2}-1\right)}{2ix^2}\,\mathrm{d}x.$$

En prenant les parties réelle et imaginaire, on obtient, pour tout $t \geqslant 0$, l'existence des intégrales $\displaystyle\int_0^{+\infty} \mathrm{e}^{-tx^2} \sin x^2 \,\mathrm{d}x$ et $\displaystyle\int_0^{+\infty} \mathrm{e}^{-tx^2} \cos x^2 \,\mathrm{d}x$.

Pour $t = 0$, on obtient : $F_2(0) = \displaystyle\int_0^{+\infty} \sin x^2 \,\mathrm{d}x = \frac{1}{2}\int_0^{+\infty} \frac{1-\cos x^2}{x^2}\,\mathrm{d}x$ et :

$$G_2(0) = \int_0^{+\infty} \cos x^2 \,\mathrm{d}x = \frac{1}{2}\int_0^{+\infty} \frac{\sin x^2}{x^2}\,\mathrm{d}x.$$

Il en résulte que $F_2(0) > 0$.

b) Soit $t > 0$, on a une variante de la méthode d'intégration par parties utilisée en a). On utilise $\left(\mathrm{e}^{(i-t)x^2}\right)' = 2(i-t)x\mathrm{e}^{(i-t)x^2}$. On a, pour tous $a > 0$ et $A > 0$, $a < A$, par intégration par parties :

$$2(i-t)\int_a^A \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x = \int_a^A \frac{\left(\mathrm{e}^{(i-t)x^2}\right)'}{x}\,\mathrm{d}x = \left[\frac{\mathrm{e}^{(i-t)x^2}}{x}\right]_a^A + \int_a^A \frac{\mathrm{e}^{(i-t)x^2}}{x^2}\,\mathrm{d}x$$

En passant à la limite quand A tend vers $+\infty$ le second membre tend vers :

$$-\frac{\mathrm{e}^{(i-t)a^2}}{a} + \int_a^{+\infty} \frac{\mathrm{e}^{(i-t)x^2}}{x^2}\,\mathrm{d}x$$

et l'on a :

$$\int_a^{+\infty} \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x = \frac{1}{2(t-i)}\frac{\mathrm{e}^{(i-t)a^2}}{a} + \frac{1}{2(i-t)}\int_a^{+\infty} \frac{\mathrm{e}^{(i-t)x^2}}{x^2}\,\mathrm{d}x.$$

On a $\left|\displaystyle\int_a^{+\infty} \frac{\mathrm{e}^{(i-t)x^2}}{x^2}\,\mathrm{d}x\right| \leqslant \int_a^{+\infty} \frac{1}{x^2}\,\mathrm{d}x = \frac{1}{a}$, d'où :

$$\left|\int_a^{+\infty} \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x\right| \leqslant \frac{1}{\sqrt{1+t^2}}\left(\frac{1}{2a} + \frac{1}{2a}\right) \leqslant \frac{1}{a}\cdot$$

2. Soit $t > 0$. On a $\left|\mathrm{e}^{(i-t)x}\right| = \mathrm{e}^{-tx}$, donc la fonction $x \mapsto \mathrm{e}^{(i-t)x}$ est intégrable sur $[0, +\infty[$ et il en est de même de ses parties réelle et imaginaire. On a, pour tout $a > 0$, $\displaystyle\int_0^a \mathrm{e}^{(i-t)x}\,\mathrm{d}x = \left[\frac{\mathrm{e}^{(i-t)x}}{i-t}\right]_0^a$.

On en déduit $\displaystyle\int_0^{+\infty} \mathrm{e}^{(i-t)x}\,\mathrm{d}x = \frac{1}{t-i} = \frac{i+t}{1+t^2}$, d'où :

$$\int_0^{+\infty} \mathrm{e}^{-tx}\sin x\,\mathrm{d}x = \frac{1}{1+t^2} \quad \text{et} \quad \int_0^{+\infty} \mathrm{e}^{-tx}\cos x\,\mathrm{d}x = \frac{t}{1+t^2}\cdot$$

3. a) Montrons que, pour tout $t > 0$, la fonction positive :

$$h_t : (x, y) \in \Delta \mapsto \mathrm{e}^{-t(x^2+y^2)}$$

est intégrable sur Δ. On note :

$$\forall k \in \mathbb{N}^*,\ \Delta_k := \{(x, y) \in \mathbb{R}^2 \mid x \geqslant 0,\ y \geqslant 0,\ x^2 + y^2 \leqslant k^2\}.$$

La suite $(\Delta_k)_{k\in\mathbb{N}^*}$ est une suite exhaustive de compacts de Δ.
Par passage en coordonnées polaires on obtient :

$$\iint_{\Delta_k} h_t(x, y)\,\mathrm{d}x\,\mathrm{d}y = \frac{\pi}{2}\int_0^k \mathrm{e}^{-tr^2} r\,\mathrm{d}r \leqslant \frac{\pi}{2}\int_0^{+\infty} \mathrm{e}^{-tr^2} r\,\mathrm{d}r < +\infty.$$

Donc h_t est $\mathscr{R}$-intégrable sur Δ. On a $|f_t| \leqslant h_t$ et $|g_t| \leqslant h_t$, donc f_t et g_t sont $\mathscr{R}$-intégrables sur Δ.

b) On définit une suite exhaustive de compacts de Δ : $\big(K_k := [0, k]^2\big)_{k\in\mathbb{N}^*}$. En utilisant l'intégrabilité de f_t, l'égalité :

$$\sin(x^2 + y^2) = \sin x^2 \cos y^2 + \sin y^2 \cos x^2$$

et le théorème de Fubini, on obtient :

$$\begin{aligned}
\iint_{\Delta} f_t(x, y)\,\mathrm{d}x\,\mathrm{d}y &= \lim_{k\to+\infty} \iint_{\Delta_k} f_t(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \lim_{k\to+\infty} \left(\int_0^k \sin x^2\,\mathrm{d}x\right)\left(\int_0^k \cos y^2\,\mathrm{d}y\right) \\
&\qquad + \lim_{k\to+\infty} \left(\int_0^k \cos x^2\,\mathrm{d}x\right)\left(\int_0^k \sin y^2\,\mathrm{d}y\right) \\
&= 2F_2(t)G_2(t).
\end{aligned}$$

On montre de façon similaire $\displaystyle\iint_{\Delta} g_t\,\mathrm{d}x\,\mathrm{d}y = G_2^2(t) - F_2^2(t)$.

On peut également calculer les intégrales $\displaystyle\iint_{\Delta} f_t\,\mathrm{d}x\,\mathrm{d}y$ et $\displaystyle\iint_{\Delta} g_t\,\mathrm{d}x\,\mathrm{d}y$ en utilisant un passage en coordonnées polaires. On obtient :

$$\begin{aligned}
\iint_{\Delta} f_t(x, y)\,\mathrm{d}x\,\mathrm{d}y &= \lim_{k\to+\infty} \iint_{K_k} f_t(x, y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \lim_{k\to+\infty} \iint_{K_k} \mathrm{e}^{-tr^2} \sin r^2 r\,\mathrm{d}r\,\mathrm{d}\theta \\
&= \iint_{\Delta} \mathrm{e}^{-tr^2} \sin r^2\, r\,\mathrm{d}r\,\mathrm{d}\theta = \frac{\pi}{4}F_1(t) = \frac{\pi}{4}\frac{1}{1+t^2}\cdot
\end{aligned}$$

On montre de façon similaire $\displaystyle\iint_{\Delta} g_t\,\mathrm{d}x\,\mathrm{d}y = \frac{\pi}{4}G_1(t) = \frac{\pi}{4}\frac{t}{1+t^2}\cdot$

On en déduit, pour $t > 0$:

$$2F_2(t)G_2(t) = \frac{\pi}{4}\frac{1}{1+t^2} \quad \text{et} \quad G_2^2(t) - F_2^2(t) = \frac{\pi}{4}\frac{t}{1+t^2}\cdot$$

4. Pour tout $a > 0$ fixé, on a $\lim\limits_{t\to 0^+} \displaystyle\int_0^a \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x = \int_0^a \mathrm{e}^{ix^2}\,\mathrm{d}x$.

   Soit $\varepsilon > 0$. On choisit $a > 0$ tel que $\frac{1}{a} \leqslant \varepsilon/4$. Il existe $\delta > 0$ tel que, pour tout $0 < t < \delta$:
$$\left|\int_0^a \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x - \int_0^a \mathrm{e}^{ix^2}\,\mathrm{d}x\right| \leqslant \varepsilon/2.$$
   On a alors en utilisant deux fois 1.b) :
$$\begin{aligned}\left|\int_0^{+\infty} \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x - \int_0^{+\infty} \mathrm{e}^{ix^2}\,\mathrm{d}x\right| &\leqslant \left|\int_0^a \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x - \int_0^a \mathrm{e}^{ix^2}\,\mathrm{d}x\right| \\ &\quad + \left|\int_a^{+\infty} \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x\right| + \left|\int_a^{+\infty} \mathrm{e}^{ix^2}\,\mathrm{d}x\right| \\ &\leqslant \varepsilon/2 + \varepsilon/4 + \varepsilon/4 = \varepsilon.\end{aligned}$$
   Le résultat s'en déduit en passant aux parties réelle et imaginaire de :
$$\int_0^{+\infty} \mathrm{e}^{(i-t)x^2}\,\mathrm{d}x - \int_0^{+\infty} \mathrm{e}^{ix^2}\,\mathrm{d}x.$$

5. D'après 3.b) on a, pour tout $t > 0$:
$$2F_2(t)G_2(t) = \frac{\pi}{4}\frac{1}{1+t^2} \quad \text{et} \quad G_2^2(t) - F_2^2(t) = \frac{\pi}{4}\frac{t}{1+t^2}\cdot$$
   On en déduit par passage à la limite, lorsque t tend vers 0, $F_2^2(0) = G_2^2(0)$ et $2F_2(0)G_2(0) = \frac{\pi}{4} > 0$.

   Nous avons vu que $F_2(0) = \dfrac{1}{2}\displaystyle\int_0^{+\infty} \frac{1-\cos x^2}{x^2}\,\mathrm{d}x > 0$, donc $G_2(0) > 0$. Il en résulte que $F_2(0) = G_2(0) = \sqrt{\dfrac{\pi}{8}}\cdot$

---

2 L'intégrale de Henstock-Kurzweil

Cette partie est consacrée à une présentation de l'intégrale de Henstock-Kurzweil pour les fonctions *d'une variable*. Pour un bref historique de cette intégrale nous renvoyons le lecteur au module *Intégration* du livre [L2] (*Histoire. Développements récents*). La définition que nous allons utiliser a été introduite indépendamment par Ralph Henstock et Jaroslaw Kurzweil au milieu du XXe siècle (en 1955 et en 1957). Même s'il a fallu beaucoup de temps pour qu'elle soit dégagée, cette définition est *simple* et *très proche de la définition initiale de Riemann*. Selon les auteurs, l'intégrale de Henstock-Kurzweil est aussi appelée intégrale de jauge, intégrale de Denjoy-Perron ou intégrale de Riemann complète. En 1964, Henstock a intitulé une conférence à Leicester présentant « son » intégrale : *Integration without tears* (*L'intégration sans larmes*), nous nous sommes efforcés de justifier ce titre.

2.1 Intégrale de Henstock-Kurzweil sur un segment

Dans la théorie de Henstock-Kurzweil les principaux résultats (théorème de Hake, lemme de Henstock, théorème de convergence monotone) se démontrent « assez facilement ». Les preuves reposent en grande partie sur quelques « astuces » dans le maniement des jauges (procédé droite-gauche, jauges forçant des points de marquage,...) que nous avons soigneusement mises en évidence et dont le lecteur devrait facilement acquérir la maîtrise. Nous avons cependant signalé des passages un peu techniques qui peuvent être omis en première lecture.

Dans l'absolu, il n'est pas nécessaire de connaître l'intégrale de Riemann pour aborder l'étude de celle de Henstock-Kurzweil et l'on pourrait commencer directement par celle-ci. Nous conseillons toutefois la démarche contraire, plus progressive, que nous avons choisie dans nos livres.

2.1.1 Subdivisions et sommes de Riemann

Nous faisons tout d'abord quelques brefs rappels sur les subdivisions (pointées ou non) et les sommes de Riemann associées. Pour plus de détails, nous renvoyons le lecteur au module *Intégration* du livre [L2].

Dans toute cette partie $[a, b]$ est un segment tel que $a < b$ sauf mention explicite du contraire.

Définition 15.

1. On appelle *subdivision* d'un segment $[a, b]$ toute *famille* $\mathbf{u} := (x_i)_{i\in[\![0,m]\!]}$ de points de ce segment telle que $m \in \mathbb{N}^*$ et :
$$x_0 := a < x_1 < \ldots < x_m := b$$
Les x_i sont appelés *points de subdivision* de $\mathbf{u}$.
2. On appelle *subdivision pointée* d'un segment $[a, b]$ la donnée d'une subdivision $\mathbf{u} := (x_i)_{i\in[\![0,m]\!]}$ et, pour tout $i \in [\![0, m-1]\!]$, d'un point $\xi_i \in [x_i, x_{i+1}]$. On note $\mathcal{D} := \big(([x_i, x_{i+1}], \xi_i)\big)_{i\in[\![0,m-1]\!]}$ une telle subdivision (c'est une *famille de segments pointés*). Les ξ_i sont appelés *points marqués* de $\mathcal{D}$.

On rappelle que si $f : [a, b] \to \mathbb{R}$ et si $\mathcal{D} := (J_i := ([x_i, x_{i+1}], \xi_i))_{i\in[\![0,m-1]\!]}$ est une subdivision pointée du segment $[a, b]$, la somme de Riemann associée est :
$$S_{\mathcal{D}}(f) := \sum_{i\in[\![0,m-1]\!]} f(\xi_i) \operatorname{long} J_i = \sum_{i\in[\![0,m-1]\!]} f(\xi_i)(x_{i+1} - x_i).$$

Les deux paragraphes suivants sont consacrés à des « exercices d'assouplissement » sur les sommes de Riemann. Les résultats sont élémentaires mais ils seront *très utilisés* dans la suite. Ils sont résumés ci-dessous dans un *Point Méthode*, page 248.

Découpage d'une subdivision en subdivisions pointées

D'abord un conseil au lecteur : il faut garder en mémoire que $m+1$ *points sont nécessaires* pour définir *une subdivision en* m *segments* !

Soit $\mathcal{D} := \big(([x_i, x_{i+1}], \xi_i)\big)_{i\in[\![0,m-1]\!]}$ une subdivision pointée de $[a,b]$. On choisit une sous-famille de p indices de $[\![1, m-1]\!]$, soit : $i_1 < i_2 < \ldots < i_p$, que l'on complète par $i_0 := a$ et $i_{p+1} := b$. Pour tout $k \in [\![0,p]\!]$, on définit une subdivision pointée $\mathcal{D}_k := \big(([x_i, x_{i+1}], \xi_i)\big)_{i\in[\![i_k,i_{k+1}]\!]}$ de $\big[x_{i_k}, x_{i_{k+1}}\big]$. On dit que l'on a *découpé* $\mathcal{D}$ en les subdivisions $\mathcal{D}_k$ à partir des points de subdivision i_k de $\mathcal{D}$ ($k \in [\![1,p]\!]$). On a évidemment, pour toute fonction $f : [a,b] \to \mathbb{R}$:

$$S_{\mathcal{D}}(f) = \sum_{k\in[\![0,p-1]\!]} S_{\mathcal{D}_k}(f). \tag{16}$$

Inversement, si $\mathbf{v} := (c_k)_{k\in[\![0,p-1]\!]}$ est une subdivision non pointée de $[a,b]$ et si, pour tout $k \in [\![0,p]\!]$, $\mathcal{D}_k$ est une subdivision pointée de $[c_k, c_{k+1}]$, la réunion $\mathcal{D} := \bigcup_{k\in[\![1,p]\!]} \mathcal{D}_k$ des familles $\mathcal{D}_k$ est une subdivision pointée de $[a,b]$ dont les c_k sont des points de subdivision. On retrouve les D_k en découpant $\mathcal{D}$ à partir des c_k.

Le découpage d'une subdivision pointée $\mathcal{D}$, décrit ci-dessus, repose sur un « bon choix » des i_k : ce sont des points de subdivision de $\mathcal{D}$. Nous allons voir qu'il est difficile de s'affranchir de cette hypothèse si l'on veut conserver l'égalité (16) (et même si l'on accepte une « petite différence » entre les deux membres).

Si $\mathbf{u}$ est une subdivision *non pointée* de $[a,b]$, en ajoutant arbitrairement un nombre fini de points de subdivision on obtient une subdivision plus fine. Par contre, on ne peut pas en général procéder ainsi « raisonnablement » à partir d'une subdivision *pointée*. Soit en effet $\mathcal{D} := \big\{([x_i, x_{i+1}], \xi_i)\big\}_{i\in[\![0,m-1]\!]}$ une telle subdivision. Si $c \in \,]x_i, x_{i+1}[$, on peut subdiviser $[x_i, x_{i+1}]$ en les deux segments $[x_i, c]$ et $[c, x_{i+1}]$, mais il y a un *problème de choix de points marqués*. Supposons, par exemple, que $\xi_i \in [x_i, c]$. Alors il est naturel de marquer $[x_i, c]$ en ξ_i, mais on ne sait pas où marquer $[c, x_{i+1}]$. On peut bien sûr décider de le faire par un point χ choisi *arbitrairement*, mais un tel choix peut être, en pratique, désastreux. En effet la contribution $f(\xi_i)(x_{i+1} - x_i)$ à la somme de Riemann $S_{\mathcal{D}}(f)$ est remplacée, après division, par $f(\xi_i)(c - x_i) + f(\chi)(x_{i+1} - c)$. La différence est $\big(f(\chi) - f(\xi_i)\big)(x_{i+1} - c)$, et si f est *très irrégulière* sur $[c, x_{i+1}]$ (comme très souvent dans le cadre de la théorie de Henstock-Kurzweil), cette différence est « incontrôlable ». Nous n'avons pas rencontré ce problème dans la théorie de Riemann, en effet si f est $\mathscr{R}$-intégrable sur $[a,b]$, on a $|f| \leqslant M$ sur $[a,b]$ et l'on contrôle la différence par :

$$\big|f(\chi) - f(\xi_i)\big|(x_{i+1} - c) \leqslant 2M(x_{i+1} - c).$$

Le procédé « droite-gauche »

Il existe heureusement un cas où la difficulté que nous venons de signaler disparaît. C'est celui *où l'on coupe en un point de marquage* $c = \xi_i$. Il est alors naturel de pointer chacun des deux segments $[x_i, \xi_i]$ et $[\xi_i, x_{i+1}]$ en ξ_i et, grâce à ce choix, nous n'avons plus de problème avec les sommes de Riemann, puisque $f(\xi_i)(x_{i+1} - x_i) = f(\xi_i)(\xi_i - x_i) + f(\xi_i)(x_{i+1} - \xi_i)$. Notons $\mathcal{D}'$ la nouvelle subdivision pointée obtenue en remplaçant le segment pointé $[x_i, x_{i+1}]$ par les *deux* segments $[x_i, \xi_i]$ et $[\xi_i, x_{i+1}]$ tous deux pointés en ξ_i, on a $S_{\mathcal{D}'}(f) = S_{\mathcal{D}}(f)$. C'est ce que l'on appelle le *procédé droite-gauche*.

Plus généralement, soit $\mathcal{D} = \big\{[x_i, x_{i+1}], \xi_i)\big\}_{i\in[\![0,m-1]\!]}$ une subdivision pointée de $[a,b]$. On considère l'ensemble des points de marquage ξ_i de $\mathcal{D}$ qui *ne sont pas des points de subdivision* (c'est-à-dire qui sont tels que $\xi_i \in]x_i, x_{i+1}[$) et l'on en *choisit* un sous-ensemble de p éléments (dans les applications ce choix sera fait en fonction du problème étudié) : $\xi_{i_1} < \xi_{i_2} < \ldots < \xi_{i_p}$. On peut appliquer le procédé droite-gauche à tous les segments $[x_{i_k}, x_{i_k+1}]$, on obtient ainsi une subdivision $\mathcal{D}'$ telle que tous les ξ_{i_k} en soient, à la fois, des points de subdivision *et* des points de marquage. On vérifie $S_{\mathcal{D}'}(f) = S_{\mathcal{D}}(f)$.
On peut découper $\mathcal{D}'$ à partir des ξ_{i_k}, on obtient des subdivisions $\mathcal{D}_k$ et l'on a, d'après (16) :

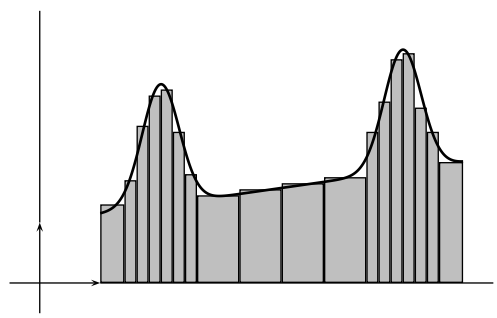
$$S_{\mathcal{D}}(f) = S_{\mathcal{D}'}(f) = \sum_{k\in[\![1,p-1]\!]} S_{\mathcal{D}_k}(f). \tag{17}$$

Point Méthode

Soit $\mathcal{D} := \big([x_i, x_{i+1}], \xi_i\big)_{i\in[\![0,m-1]\!]}$ une subdivision pointée de $[a,b]$. On a un procédé droite-gauche. On choisit $j \in [\![0, m-1]\!]$ tel que $\xi_j \in]x_j, x_{j+1}[$ (s'il en existe) et l'on considère la nouvelle subdivision $\mathcal{D}'$ obtenue en remplaçant le segment $[x_j, x_{j+1}]$ par les *deux* segments $[x_j, \xi_j]$ et $[\xi_j, x_{j+1}]$ tous deux pointés en ξ_j. La somme de Riemann correspondante associée à une fonction $f : [a,b] \to \mathbb{R}$ (arbitraire) *ne change pas* : $S_{\mathcal{D}}(f) = S_{\mathcal{D}'}(f)$. Plus généralement, on peut appliquer le procédé à p points de marquage $\{\xi_{i_k}\}_{k\in[\![1,p]\!]}$ de $\mathcal{D}$ qui *ne sont pas des points de subdivision*. On peut ainsi « ajouter » les ξ_{i_k} aux points de subdivision de $\mathcal{D}$, on obtient une subdivision pointée $\mathcal{D}'$ de $[a,b]$ dont les ξ_{i_k} sont des points de subdivision *et* des points de marquage et l'on a $S_{\mathcal{D}}(f) = S_{\mathcal{D}'}(f)$.

2.1.2 Jauges et subdivisions

L'idée centrale de la théorie développée par Henstock et Kurzweil est de *ne pas considérer toutes les sommes de Riemann* (de pas suffisamment petit) pour définir l'intégrale d'une fonction. Intuitivement, si la fonction f « oscille » plus à certains endroits qu'à d'autres ou si elle n'est pas bornée, pour approximer l'aire algébrique comprise entre le graphe de f et l'axe des abscisses, on a intérêt à resserrer d'avantage les points x_k de la subdivision en certains endroits.



Ainsi, les intervalles pointés (J, ξ) des subdivisions que nous allons maintenant rencontrer seront soumis à une condition liant la longueur de l'intervalle J au point marqué $\xi \in J$. Ce lien va se traduire par l'inégalité $\operatorname{long}(J) \leqslant \delta(\xi)$ où δ est une *fonction strictement positive* (arbitraire) sur $[a,b]$, que nous appellerons une « jauge ». Les jauges que nous manipulerons seront généralement des fonctions discontinues. Le lecteur est invité à penser que la jauge δ est petite là où f oscille vite, et que δ peut même tendre vers 0 sur un ensemble dense de $[a,b]$.

Définition 16. Une *jauge* sur un intervalle $I \subset \mathbb{R}$ est une fonction *strictement positive* $\delta : I \to]0, +\infty[$.
Si $I := [a, b]$ est un *segment*, une subdivision *pointée* $\mathcal{D}$ de $[a, b]$ est δ*-fine* si :

$$\forall (J, \xi) \in \mathcal{D}, \quad \operatorname{long}(J) \leqslant \delta(\xi).$$

Remarque. Pour simplifier, nous parlerons le plus souvent de « subdivision δ-fine » en omettant le terme « pointée ». Cela ne devrait pas prêter à confusion.

Le minimum d'une famille *finie* de réels strictement positifs est l'un de ces réels, il est strictement positif. On en déduit l'exemple ci-dessous et la proposition qui suit.

Exemple. Soient $x_0, x_1, \ldots, x_n \in [a, b]$ des points *distincts*. On définit $\delta : [a, b] \to \mathbb{R}_+$ par $\delta : x \mapsto \min\limits_{\{i \in [\![0,n]\!] \mid x_i \neq x\}} |x - x_i|$. On a, pour tout $x \in [a, b]$, $\delta(x) > 0$, donc δ est une jauge sur $[a, b]$.

Proposition 69. Si $(\delta_i)_{i \in I}$ est une famille *finie* de jauges sur $[a, b]$, $\delta := \min\limits_{i \in I} \delta_i$ est une jauge sur $[a, b]$ et toute subdivision δ-fine est alors δ_i-fine pour tout $i \in I$.

On prendra garde au fait que le résultat ci-dessus ne s'étend pas aux familles infinies.

Pour toute jauge δ sur $[a, b]$, il existe une subdivision de $[a, b]$ qui est δ-fine. La démonstration de ce résultat « de compacité » essentiel pour la théorie, appelé *lemme de Cousin*, repose sur le théorème des segments emboîtés. Nous l'avons donnée dans la partie *Fonctions réglées* du module *Intégration* de [L2] auquel nous renvoyons le lecteur pour quelques commentaires.

Proposition 70 (Lemme de Cousin).
Si δ est une jauge sur $[a, b]$, alors il existe une subdivision de $[a, b]$ qui est δ-fine.

Nous donnons ci-dessous des exemples de jauges δ permettant de *fixer certains des points marqués* des subdivisions δ-fines en des endroits *déterminés par avance*. Ces jauges permettent d'appliquer le procédé *droite-gauche* (*cf.* le point méthode de la page 248), elles seront *très souvent utilisées* par la suite pour estimer des sommes de Riemann.
Nous conseillons au lecteur d'étudier attentivement le premier exemple.

Exemples.

1. Soient $c \in [a, b]$ et $\alpha > 0$. On définit une jauge sur $[a, b]$ par :

$$\delta_c : x \in [a, b] \mapsto \begin{cases} \frac{1}{2}|x - c| & \text{si} \quad x \neq c \\ \alpha & \text{si} \quad x = c. \end{cases}$$

   Soit $\mathcal{D}$ une subdivision pointée δ_c-fine de $[a, b]$, alors c est l'un des points marqués de la subdivision pointée $\mathcal{D}$. En effet, soit $\mathcal{D} = \{(J_k, \xi_k)\}$, on a, pour (au moins) un indice i convenable, $c \in J_i$. Si $c \neq \xi_i$, $\operatorname{long} J_i \leqslant \delta_c(\xi_i) = \frac{1}{2}|\xi_i - c|$ ce qui contredit $\operatorname{long} J_i \geqslant |\xi_i - c|$ $(\xi_i, c \in J_i)$, ainsi $\xi_i = c$.

On a *deux* possibilités.

– Ou bien c est contenu dans *deux* segments adjacents J_i et J_{i+1} : c est l'extrémité x_{i+1} de J_i et l'origine de J_{i+1}.

Alors $c = \xi_i = \xi_{i+1} = x_{i+1}$ est le point marqué à la fois de J_i et J_{i+1}.

– Ou bien c est un *point intérieur* d'un segment $J_i = [x_i, x_{i+1}]$ *unique* et c est le point marqué de J_i. On peut alors obtenir une nouvelle subdivision pointée $\mathcal{D}'$ de $[a, b]$ en employant le procédé droite-gauche décrit plus haut, c'est-à-dire en *remplaçant* dans $\mathcal{D}$ le segment pointé $(J_i = [x_i, x_{i+1}], c)$ par les *deux* segments pointés $\big([x_i, c], c\big)$ et $\big([c, x_{i+1}], c\big)$. On est ainsi ramené au premier cas.

2. Plus généralement, soient $c_1, \ldots, c_p \in [a, b]$ et $\alpha_1, \ldots, \alpha_p > 0$.

   On note $\mathbf{c} = \{c_1, \ldots, c_p\}$ et l'on définit une jauge sur $[a, b]$ par :

$$\delta_{\mathbf{c}} : x \in [a, b] \mapsto \begin{cases} \frac{1}{2} \min\limits_{i \in [\![1,p]\!]} |x - c_i| & \text{si} \quad x \neq c_1, \ldots, c_p \\ \alpha_i & \text{si} \quad x = c_i,\ i \in [\![1, p]\!] \end{cases}.$$

   On montre, comme dans 1., que si $\mathcal{D}$ est une subdivision pointée $\delta_{\mathbf{c}}$-fine, alors $c_1, \ldots, c_p$ sont des points marqués de $\mathcal{D}$. Il en résulte que tout segment J_i ($i \in [\![0, m-1]\!]$) contient au plus *l'un* des $c_j \in \mathbf{c}$.

   Pour tout $k \in [\![1, p]\!]$, on a, comme dans 1., l'alternative suivante.

   – Ou bien c_k est l'un des points de subdivision de $\mathcal{D}$, il est contenu dans deux intervalles adjacents dont il est le point marqué.

   – Ou bien c_k est un *point intérieur* d'un segment $J_i = [x_i, x_{i+1}]$ *unique* et c_k est le point marqué de J_i.

   On peut alors obtenir une nouvelle subdivision pointée $\mathcal{D}'$ de $[a, b]$ en employant le procédé droite-gauche, c'est-à-dire en *remplaçant* dans $\mathcal{D}$, pour *tout* $k \in [\![1, p]\!]$, le segment pointé $(J_i = [x_i, x_{i+1}], c_k)$ par les *deux* segments pointés $\big([x_i, c], c_k\big)$ et $\big([c_k, x_{i+1}], c_k\big)$. On est ainsi ramené au premier cas.

3. Si δ' est une jauge sur $[a, b]$ et si $c_1, \ldots, c_p \in [a, b]$, on peut trouver une jauge δ, majorée par δ', telle que $c_1, \ldots, c_p \in [a, b]$ soient des points marqués pour toute subdivision pointée δ-fine (par exemple $\delta := \min(\delta', \delta_{\mathbf{c}})$). On « force » ainsi des points marqués.

Les exemples ci-dessus suggèrent l'énoncé suivant dont la preuve utilise également le procédé droite-gauche.

Lemme 71. Soient $a,\ b,\ c_1, \ldots,\ c_p \in \mathbb{R}$ $(p \geqslant 1)$ tels que $a \leqslant c_1 < \ldots < c_p \leqslant b$ et $\alpha_1, \ldots,\ \alpha_p \in \mathbb{R}_+^*$. On note $\mathbf{c} := \{c_1, \ldots,\ c_p\} \subset [a, b]$.
On définit une jauge $\delta_{\mathbf{c}}$ sur $[a, b]$ par :

$$\delta_{\mathbf{c}} : x \in [a, b] \mapsto \begin{cases} \frac{1}{2} \min\limits_{k \in [\![1,p]\!]} |x - c_k| = \frac{1}{2} d(x, \mathbf{c}) & \text{si} \quad x \notin \mathbf{c} \\ \alpha_k & \text{si} \quad x = c_k,\ k \in [\![1, p]\!] \end{cases}.$$

Cette jauge a la propriété suivante. *Pour toute* subdivision pointée $\delta_{\mathbf{c}}$-fine $\mathcal{D}$, les points $\{c_1, \ldots,\ c_p\}$ sont *des points marqués* de $\mathcal{D}$ et, en employant le procédé droite-gauche (*cf.* le point méthode de la page 248), on obtient une subdivision pointée $\mathcal{D}'$ telle que :

(i) pour *toute* fonction $f : [a, b] \to \mathbb{R}$:

$$S_{\mathcal{D}}(f) = S_{\mathcal{D}'}(f);$$

(ii) les *points de subdivision* de la subdivision $\mathcal{D}'$ sont la réunion des $\{x_i\}_{i \in [\![0,m]\!]}$ et des $\{c_k\}_{k \in [\![1,p]\!]}$;

(iii) les *points marqués* de $\mathcal{D}'$ sont d'une part ceux de $\mathcal{D}$ et d'autre part les c_k n'appartenant pas à l'ensemble $\{x_i\}_{i \in [\![0,m]\!]}$;

(iv) Supposons $a < c_1$ et $c_p < b$. Il existe une subdivision pointée δ-fine $\mathcal{D}_0$ de $[a, c_1]$, une subdivision pointée δ-fine $\mathcal{D}_k$ de $[c_p, b]$ et pour tout $k \in [\![1, p-1]\!]$ une subdivision pointée δ-fine $\mathcal{D}_k$ de $[c_k, c_{k+1}]$, telles que l'on obtienne $\mathcal{D}'$ en réunissant les $\{\mathcal{D}_k\}_{k \in [\![0,p]\!]}$. On a alors, pour toute fonction $f : [a, b] \to \mathbb{R}$:

$$S_{\mathcal{D}}(f) = \sum_{k \in [\![0,p-1]\!]} S_{\mathcal{D}_k}(f).$$

On a des variantes si $c_1 = a$ ou $c_p = b$, dont le détail est laissé au lecteur.

Notons que si une jauge δ est plus petite que $\delta_{\mathbf{c}}$, toute subdivision δ-fine est $\delta_{\mathbf{c}}$-fine. La jauge δ a donc aussi les propriétés (i), (ii), (iii), (iv). Ainsi l'on peut *remplacer* toute subdivision δ-fine par une nouvelle subdivision admettant des points de subdivision *donnés à l'avance*.
S'il existe un indice $k \in [\![1, p-1]\!]$ tel que $c_k < x < c_{k+1}$, on a :

$$d(x, \mathbf{c}) = \min_{k \in [\![1,p]\!]} |x - c_k| = \min(x - c_k, c_{k+1} - x).$$

Certains auteurs définissent $\delta_{\mathbf{c}}$ en utilisant les $\min(x - c_k, c_{k+1} - x)$.

Démonstration. D'après l'exemple 2 ci-dessus, les points de $\mathbf{c}$ sont *des points marqués* de $\mathcal{D}$, on peut donc appliquer le procédé droite-gauche en « ajoutant » les points de $\mathbf{c}$ aux points de subdivision de $\mathcal{D}$. On obtient une subdivision $\mathcal{D}'$ telle que $S_{\mathcal{D}'}(f) = S_{\mathcal{D}}(f)$. On vérifie facilement que $\mathcal{D}'$ est δ-fine. Par suite, si $a < c_1$ et $c_p < b$, $\mathcal{D}'$ induit une subdivision δ-fine sur $[a, c_1]$, $[c_p, b]$ et sur chaque $[c_k, c_{k+1}]$ $(k \in [\![1, p-1]\!])$.
Le résultat s'en déduit. ∎

Le lemme ci-dessus, très simple, semble un « exercice de style » sur la manipulation des jauges et des subdivisions. En fait, il sera par la suite un *outil essentiel* dans la preuve d'un certain nombre de résultats importants.

Point Méthode

On complète le *Point Méthode* de la page 248, dont on reprend les notations.
On a un principe de « forçage des points marqués ». En utilisant une jauge δ adéquate on peut *forcer* un certain nombre *fini* de points *donnés à l'avance* $\{c_k\}_{k\in[\![1,p]\!]}$ à être des points marqués de *toute* subdivision δ-fine $\mathcal{D}$.
En utilisant le procédé « gauche-droite », on peut alors « remplacer » la subdivision $\mathcal{D}$ par une subdivision δ-fine $\mathcal{D}'$ admettant les $\{c_k\}_{k\in[\![1,p]\!]}$ comme points de subdivision *et* comme points de marquage et telle que $S_{\mathcal{D}}(f) = S_{\mathcal{D}'}(f)$.

2.1.3 Intégrale de Henstock-Kurzweil — Définition et exemples

Dans le module *Intégration* du livre [L2] nous avons défini la notion d'intégrale de Riemann à partir des fonctions en escalier, puis nous avons montré qu'il existe une définition équivalente où l'intégrale est définie comme « limite » de sommes de Riemann (*cf.* le théorème 42 de la page 650 de [L2]). C'est cette deuxième définition que nous allons généraliser avec l'intégrale de Henstock-Kurzweil. La modification est (apparemment...) très simple, il suffit de remplacer les jauges *constantes* par des jauges *arbitraires*. Mais les conséquences de cette modification vont être considérables ; il y aura en particulier « beaucoup plus de fonctions intégrables ». On notera que le cardinal de l'ensemble des jauges constantes est celui de $\mathbb{R}$ tandis que celui des jauges arbitraires est celui des applications de $\mathbb{R}$ dans $\mathbb{R}$ (c'est-à-dire celui de $2^{\mathbb{R}}$, *cf.* l'exercice III.1.38). Il est donc *strictement* supérieur (*cf.* le théorème 12 de la page 42 du module *fondements* du livre [L1]).

Définition 17. Une fonction $f : [a,b] \to \mathbb{R}$ est dite *intégrable au sens de Henstock-Kurzweil* ou $\mathscr{HK}$-intégrable s'il existe un réel A tel que pour tout $\varepsilon > 0$, on puisse trouver une fonction jauge δ sur $[a,b]$ en sorte que si $\mathcal{D}$ est une subdivision δ-fine de $[a,b]$, alors $\left|S_{\mathcal{D}}(f) - A\right| \leqslant \varepsilon$. On appelle A *l'intégrale de Henstock-Kurzweil de* f *sur* $[a,b]$ et l'on note $A = \displaystyle\int_a^b f = \int_a^b f(x)\,\mathrm{d}x$.

Pour que cette définition de l'intégrale de f sur $[a,b]$ ait un sens, il faut vérifier l'unicité. Supposons donc qu'il existe deux réels A_1 et A_2 vérifiant les conditions précédentes. Alors, pour tout $\varepsilon > 0$, on peut trouver deux jauges δ_1 et δ_2 telles que pour toute subdivision $\mathcal{D}_1$ qui est δ_1-fine, on ait $|S_{\mathcal{D}_1}(f) - A_1| \leqslant \varepsilon$ et pour toute subdivision $\mathcal{D}_2$ qui est δ_2-fine, on ait $|S_{\mathcal{D}_2}(f) - A_2| \leqslant \varepsilon$.
La fonction $\delta := \min(\delta_1, \delta_2)$ est une jauge sur $[a,b]$ (*cf.* la proposition 69). Soit $\mathcal{D}$ une subdivision δ-fine de $[a,b]$. Alors $\mathcal{D}$ est à la fois δ_1-fine et δ_2-fine. Par conséquent :

$$|S_{\mathcal{D}}(f) - A_1| \leqslant \varepsilon \quad \text{et} \quad |S_{\mathcal{D}}(f) - A_2| \leqslant \varepsilon.$$

On en déduit que pour tout $\varepsilon > 0$, $|A_1 - A_2| \leqslant 2\varepsilon$. Donc, $A_1 = A_2$.

Exemple. Si $f : [a,b] \to \mathbb{R}$ est la fonction constante λ, alors f est $\mathscr{HK}$-intégrable et $\displaystyle\int_a^b f = \lambda(b-a)$.

Soit, en effet, une subdivision pointée $\mathcal{D} := \big\{\big([x_i, x_{i+1}], \xi_i\big)\big\}_{i\in[\![1,m-1]\!]}$, on a :

$$S_{\mathcal{D}}(f) := \sum_{i\in[\![0,m-1]\!]} f(\xi_i)(x_{i+1}-x_i) = \lambda \sum_{i\in[\![0,m-1]\!]} (x_{i+1}-x_i) = \lambda(x_m - x_0) = \lambda(b-a).$$

La jauge peut donc être choisie arbitrairement.

Exercice 20.

Soient $f,\ g : [a,b] \to \mathbb{R}$ telles que $f(x) = g(x)$, pour tout $x \in [a,b] \setminus C$, C étant une partie *finie*. Montrer que f est $\mathscr{HK}$-intégrable sur $[a,b]$ si, et seulement si, g l'est et que, s'il en est ainsi, $\displaystyle\int_a^b f = \int_a^b g$.

Solution. On peut évidemment supposer C non vide. Notons $C := \{c_k\}_{k\in[\![1,p]\!]}$. Considérons une jauge δ_0 sur $[a,b]$ définie par :

$$\delta_0(x) := \begin{cases} 1 & \text{si } x \in [a,b] \setminus C \\ \dfrac{\varepsilon}{4p} \min\left(\dfrac{1}{1+\big|f(c_k)\big|}, \dfrac{1}{1+\big|g(c_k)\big|}\right) & \text{si } x = c_k \in C. \end{cases}$$

Supposons que f est $\mathscr{HK}$-intégrable.

Soient $\varepsilon > 0$ et δ' une jauge sur $[a,b]$ telle que $\displaystyle\left|S_{\mathcal{D}}(f) - \int_a^b f\right| \leqslant \varepsilon/2$ pour toute subdivision δ'-fine $\mathcal{D} := \big((J_i, \xi_i)\big)_{i\in[\![0,m]\!]}$. On définit une jauge $\delta := \min(\delta', \delta_0)$. On note $S \subset [\![0,m]\!]$ l'ensemble des indices $i \in [\![0,m]\!]$ tels que ξ_i soit l'un des $\{c_k\}_{k\in[\![1,p]\!]}$. On note alors $c_{k_i} = \xi_i$. On a $S_{\mathcal{D}}(f) - S_{\mathcal{D}}(g) = \sum_{i\in S}\big(f(c_{k_i}) - g(c_{k_i})\big) \operatorname{long} J_i$, donc :

$$\begin{aligned} |S_{\mathcal{D}}(f) - S_{\mathcal{D}}(g)| &\leqslant \sum_{i\in S} \big|f(c_{k_i}) - g(c_{k_i})\big|\delta(c_{k_i}) \\ &\leqslant \sum_{i\in S} \big(\big|f(c_{k_i})\big| + \big|g(c_{k_i})\big|\big)\,\delta(c_{k_i}) \leqslant p\,\frac{2\varepsilon}{4p} = \frac{\varepsilon}{2}\cdot \end{aligned}$$

On en déduit :

$$\left|S_{\mathcal{D}}(g) - \int_a^b f\right| \leqslant |S_{\mathcal{D}}(g) - S_{\mathcal{D}}(f)| + \left|S_{\mathcal{D}}(f) - \int_a^b f\right| \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

d'où le résultat.

Proposition 72. Si $f : [a,b] \to \mathbb{R}$ est intégrable au sens de Riemann, alors f est intégrable au sens de Henstock-Kurzweil et l'intégrale de Riemann de f sur $[a,b]$ est égale à l'intégrale de Henstock-Kurzweil de f sur $[a,b]$.

Démonstration.
L'intégrabilité au sens de Riemann revient à ne considérer que des jauges *constantes*. ■

Il en résulte que les fonctions en escalier, continues, continues par morceaux ou monotones sur un segment sont $\mathscr{HK}$-intégrables sur ce segment. Plus généralement, les fonctions réglées sur un segment (*cf.* le module *Intégration* du livre [L2]) sont $\mathscr{HK}$-intégrables sur ce segment.

Notons qu'une jauge *continue* sur un segment est minorée par une constante positive. Si, dans la définition de l'intégrabilité, on se limitait aux jauges continues, on n'aurait donc rien de plus que l'intégrale de Riemann.

Exercice 21.

Soit δ une jauge continue sur $[a,b]$. Montrer qu'il existe une subdivision pointée δ-fine : $\mathcal{D} := \big(([x_i, x_{i+1}], \xi_i)\big)_{i\in[\![0,m-1]\!]}$ de $[a,b]$ telle que, pour tout indice $i \in [\![0, m-1]\!]$, sauf éventuellement un, l'on ait $x_{i+1} - x_i = \delta(\xi_i)$.

Solution. On note $d := \inf\limits_{x\in[a,b]} \delta(x) > 0$. On construit les x_i par récurrence croissante sur i à partir de $x_0 := a$. Si $x_i + \delta(x_i) \leqslant b$, on définit $x_{i+1} := x_i + \delta(x_i)$, sinon on définit $x_{i+1} := b$ et l'on arrête la récurrence. La récurrence *s'arrête nécessairement* car, pour $N \in \mathbb{N}^*$ assez grand, $Nd > b - a$. La subdivision $\mathcal{D} := \big(([x_i, x_{i+1}], \xi_i)\big)_{i\in[\![0,m-1]\!]}$ (pointée aux *origines* des intervalles) répond à la question. On peut faire une construction similaire par récurrence descendante en partant de $x_N := b$ et en renumérotant (si nécessaire) les indices en fin de construction pour avoir $x_0 = a$. On pointe alors $\mathcal{D}$ aux *extrémités* des intervalles.

Exercice 22.

Montrer que la fonction $f : [0,1] \to \mathbb{R}$ définie par $f(x) = 0$ si $x \in [0, 1/2[$ et $f(x) = 1$ si $x \in [1/2, 1[$ est $\mathscr{HK}$-intégrable en utilisant une jauge.

Solution. La fonction f est en escalier, elle est donc évidemment $\mathscr{R}$-intégrable et par suite $\mathscr{HK}$-intégrable. Il est toutefois intéressant de décrire des preuves *directes* n'utilisant que la définition de l'intégrabilité. C'est un cas très simple et il est très facile de prouver le résultat en utilisant une jauge constante. Nous allons détailler cette preuve, puis nous en donnerons une autre utilisant une jauge variable. Des variantes de cette deuxième preuve seront utiles dans la suite.

Première méthode : jauge constante.
Soit $\varepsilon > 0$.
Soient δ une jauge constante et $\mathcal{D} := \big(([x_i, x_{i+1}], \xi_i)\big)_{i\in[\![0,N-1]\!]}$ une subdivision pointée δ-fine de $[0,1]$. Il existe $j \in [\![0, N-1]\!]$ *unique* tel que $\frac{1}{2} \in]x_j, x_{j+1}]$. On a :

$$S_{\mathcal{D}}(f) = \sum_{i\in[\![0,N-1]\!]} f(\xi_i)(x_{i+1} - x_i) = f(\xi_j)(x_{j+1} - x_j) + 1 - x_{j+1},$$

d'où $S_{\mathcal{D}}(f) - \frac{1}{2} = f(\xi_j)(x_{j+1} - x_j) + \frac{1}{2} - x_{j+1}$. On a $f(\xi_j) = 0$ ou $f(\xi_j) = 1$, donc $S_{\mathcal{D}}(f) - \frac{1}{2} = \frac{1}{2} - x_{j+1}$ ou $S_{\mathcal{D}}(f) - \frac{1}{2} = \frac{1}{2} - x_j$. Il suffit donc de prendre pour jauge $\delta := \varepsilon$.
Deuxième méthode : jauge variable. Voici une autre preuve utilisant une jauge non constante. On reprend l'idée de l'exemple 1 de la page 250 ci-dessus, en utilisant le procédé droite-gauche. Soit $\varepsilon > 0$. On définit une jauge sur $[0, 1]$ par :

$$\delta : x \mapsto \begin{cases} \frac{1}{2}|x - \frac{1}{2}| & \text{si} \quad x \neq \frac{1}{2} \\ \varepsilon & \text{si} \quad x = \frac{1}{2}. \end{cases}$$

Soit $\mathcal{D} := \big(([x_i, x_{i+1}], \xi_i)\big)_{i \in [\![0, N-1]\!]}$ une subdivision pointée δ-fine de $[0, 1]$.
Soit $j \in [\![0, N-1]\!]$ tel que $\frac{1}{2} \in]x_j, x_{j+1}]$. Nous avons vu que l'on a nécessairement $\xi_j = \frac{1}{2}$. On a *forcé le point marqué* ξ_j du segment $[x_j, x_{j+1}]$ à être en $\frac{1}{2}$. Alors :

$$S_{\mathcal{D}}(f) = \sum_{i \in [\![0, j-1]\!]} f(\xi_i)(x_{i+1} - x_i) + f(\xi_j)\,(x_{j+1} - x_j) + \sum_{i \in [\![j+1, N-1]\!]} f(\xi_i)(x_{i+1} - x_i)$$

(la première somme étant nulle si $j = 0$).
Si $i \in [\![0, j-1]\!]$, $f(\xi_i) = 0$ et si $i \in [\![j+1, N-1]\!]$, $f(\xi_j) = 1$, donc :

$$S_{\mathcal{D}}(f) = x_{j+1} - x_j + 1 - x_{j+1} = 1 - x_j.$$

Ainsi $0 \leqslant S_{\mathcal{D}}(f) - \frac{1}{2} = \frac{1}{2} - x_j \leqslant \delta\big(\frac{1}{2}\big) = \varepsilon$.

Exemples. Pour une fonction intégrable au sens de Riemann sur un segment on peut utiliser une jauge constante, mais il peut être intéressant de considérer des jauges non constantes. En particulier en relation avec le calcul approché d'intégrales (*cf.* le module *Méthodes numériques* du livre [L2]). Pour minimiser les calculs dans l'approximation d'une intégrale avec une précision donnée, il vaut mieux, en effet, utiliser des sommes de Riemann ayant *le plus petit nombre possible* de points de subdivision.

1. Commençons par un cas très simple. Soient $0 < a < b$ et $f : x \in [a, b] \mapsto x^2$. Nous savons que cette fonction admet pour *primitive* $g : x \mapsto x^3/3$. Elle est bien sûr intégrable au sens de Riemann et $\displaystyle\int_a^b x^2 \,\mathrm{d}x = g(b) - g(a) = \frac{b^3 - a^3}{3}$. Mais nous allons « l'oublier » et faire ici une étude *directe* à partir des sommes de Riemann, *en utilisant la primitive* g (et la *dérivée* : $x \mapsto 2x$) de f, et en appliquant *la définition* de l'intégrale de Henstock-Kurzweil.
   Soit $\mathcal{D} := \big(([x_i, x_{i+1}], \xi_i)\big)_{i \in [\![0, m-1]\!]}$ une subdivision pointée de $[a, b]$. La somme de Riemann associée pour f est : $S_{\mathcal{D}}(f) := \displaystyle\sum_{i \in [\![0, m-1]\!]} \xi_i^2\,(x_{i+1} - x_i)$.
   L'idée, qui sera reprise pour d'autres exemples, est d'écrire $g(b) - g(a)$ comme une somme de Riemann pour la fonction f, associée à une subdivision $\mathcal{D}'$ ayant *la même* subdivision pointée sous-jacente $\mathbf{u} := \{x_i\}_{i \in [\![0, m-1}$ que $\mathcal{D}$. On compare ensuite $S_{\mathcal{D}}(f)$ et $S_{\mathcal{D}'}(f)$.

On a $g(b) - g(a) = \sum_{i \in [\![0,m-1]\!]} \big(g(x_{i+1}) - g(x_i)\big)$. En utilisant le *théorème des accroissements finis*, on montre que, pour tout $i \in [\![0, m-1]\!]$, il existe $c_i \in]x_i, x_{i+1}[$ tel que $g(x_{i+1}) - g(x_i) = \frac{1}{3}\big(x_{i+1}^3 - x_i^3\big) = c_i^2(x_{i+1} - x_i)$
On en déduit :

$$S_{\mathcal{D}}(f) - \frac{b^3 - a^3}{3} = \sum_{i \in [\![0,m-1]\!]} (\xi_i^2 - c_i^2)\,(x_{i+1} - x_i)$$

On a $|\xi_i^2 - c_i^2| \leqslant (\xi_i + c_i)|\xi_i - c_i| \leqslant 2x_{i+1}(x_{i+1} - x_i)$, d'où

$$\begin{aligned} \left| S_{\mathcal{D}}(f) - \frac{b^3 - a^3}{3} \right| &\leqslant \sum_{i \in [\![0,m-1]\!]} 2x_{i+1}(x_{i+1} - x_i)^2 \\ &\leqslant \sum_{i \in [\![0,m-1]\!]} 2b\,(x_{i+1} - x_i)^2. \end{aligned} \tag{18}$$

Soient $\varepsilon > 0$ et δ une jauge *constante*. Si $\mathcal{D}$ est δ-fine, on a $x_{i+1} - x_i \leqslant \delta$, pour tout $i \in [\![0, m-1]\!]$, donc :

$$\begin{aligned} \left| S_{\mathcal{D}}(f) - \frac{b^3 - a^3}{3} \right| &\leqslant \sum_{i \in [\![0,m-1]\!]} 2b\,(x_{i+1} - x_i)^2 \\ &\leqslant 2b\delta \sum_{i \in [\![0,m-1]\!]} (x_{i+1} - x_i) = 2b(b-a)\delta. \end{aligned}$$

Par suite, pour avoir $\left| S_{\mathcal{D}}(f) - \frac{b^3-a^3}{3} \right| \leqslant \varepsilon$, il suffit de choisir $\delta := \dfrac{\varepsilon}{2b(b-a)}$.
Si l'on veut une subdivision $\mathcal{D}$ en parties égales, il faut utiliser n pas, avec :

$$n := \mathrm{E}\left(\frac{b-a}{\delta}\right) = \mathrm{E}\left(\frac{2b\,(b-a)^2}{\varepsilon}\right) + 1 \in \mathbb{N}^*.$$

Considérons les trois cas particuliers suivants, avec, dans chaque cas des subdivisions en parties égales $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$:

$$\begin{aligned} a &:= 1, \quad b := 3, \quad \varepsilon_1 := 1/20 \\ a &:= 3, \quad b := 5, \quad \varepsilon_2 := 1/20 \\ a &:= 1, \quad b := 5, \quad \varepsilon_3 := 1/10. \end{aligned}$$

Il faut utiliser respectivement 481, 801 et 1601 parties, correspondant aux jauges $\delta_1 := 1/240$, $\delta_2 := 1/400$, $\delta_3 := 1/400$.
Si l'on utilise $\displaystyle\int_1^5 x^2\,\mathrm{d}x = \int_1^3 x^2\,\mathrm{d}x + \int_3^5 x^2\,\mathrm{d}x$, on constate que l'on peut obtenir $\displaystyle\int_1^5 x^2\,\mathrm{d}x$ à $1/10$ près avec 1283 points de subdivision au lieu de 1602 avec le calcul direct. Cela revient à utiliser, au lieu de la jauge constante δ_3, la jauge *constante par morceaux* égale à δ_1 sur $[1, 3]$ et à δ_2 sur $]3, 5]$ (*cf.* (18)).

Revenant à (18), on peut essayer de majorer par $\sum_{i\in[\![0,m-1]\!]} 2x_{i+1}(x_{i+1}-x_i)^2$ au lieu de la majoration par $\sum_{i\in[\![0,m-1]\!]} 2b\,(x_{i+1}-x_i)^2$ qui est « grossière ». Cela suggère d'introduire la jauge *variable* $\delta : x \mapsto \frac{\varepsilon}{2x(b-a)}$.

Si $\mathcal{D} := \big(([x_i, x_{i+1}], x_{i+1})\big)_{i\in[\![0,m-1]\!]}$ est une jauge δ-fine pointée aux *extrémités* des intervalles (il en existe d'après l'exercice 21), on a, pour tout $i \in [\![0, m-1]\!]$:

$$2x_{i+1}(x_{i+1}-x_i)^2 \leqslant 2x_{i+1}\delta(x_{i+1})(x_{i+1}-x_i) = \frac{\varepsilon}{b-a}(x_{i+1}-x_i).$$

Par suite, d'après (18) :

$$\left|S_{\mathcal{D}}(f) - \frac{b^3-a^3}{3}\right| \leqslant \frac{\varepsilon}{b-a}\sum_{i\in[\![0,m-1]\!]}(x_{i+1}-x_i) = \frac{\varepsilon}{b-a}(b-a) = \varepsilon.$$

Cette inégalité n'est pas vraie en général si $\mathcal{D}$ est δ-fine mais pointée *arbitrairement*. On peut « surmonter » le problème en utilisant une jauge de la forme $\delta : x \mapsto \frac{C\varepsilon}{2x(b-a)}$, avec un « bon » choix de $0 < C < 1$. Nous laissons la question au lecteur à titre d'exercice.

2. On reprend la méthode précédente avec la fonction $f : x \in [a,b] \mapsto x^n$ ($n \in \mathbb{N}$, $n \geqslant 3$). Cette fonction admet la primitive $g : x \in [a,b] \mapsto \frac{x^{n+1}}{n+1}$. En reprenant les calculs avec des notations similaires, on obtient (pour un $c_i \in]x_i, x_{i+1}[$) :

$$S_{\mathcal{D}}(f) - \frac{b^{n+1}-a^{n+1}}{n+1} = \sum_{i\in[\![0,m-1]\!]}(\xi_i^n - c_i^n)\,(x_{i+1}-x_i)$$

On a $f'(x) = nx^{n-1}$, donc, en appliquant à nouveau le théorème des accroissements finis, on obtient $|\xi_i^n - c_i^n| \leqslant nd_i^{n-1}|\xi_i - c_i| \leqslant nx_{i+1}^{n-1}(x_{i+1}-x_i)$ (pour un $d_i \in]\xi_i, c_i[$) (on aurait pu aussi utiliser l'identité $a^n - b^n = (a-b)(a^{n-1} + a^{n-2}b + \cdots + b^{n-1})$, comme pour $n = 2$). On en déduit :

$$\begin{aligned}\left|S_{\mathcal{D}}(f) - \frac{b^{n+1}-a^{n+1}}{n+1}\right| &\leqslant \sum_{i\in[\![0,m-1]\!]} nx_{i+1}^{n-1}(x_{i+1}-x_i)^2 \\ &\leqslant \sum_{i\in[\![0,m-1]\!]} nb^{n-1}\,(x_{i+1}-x_i)^2.\end{aligned} \tag{19}$$

Par suite, pour avoir $\left|S_{\mathcal{D}}(f) - \frac{b^{n+1}-a^{n+1}}{n+1}\right| \leqslant \varepsilon$, il suffit de choisir la jauge *constante* $\delta := \dfrac{\varepsilon}{nb^{n-1}(b-a)} = \dfrac{\varepsilon}{f'(b)(b-a)}$; on peut aussi reprendre l'analyse faite en 1. avec la jauge variable $\delta : x \mapsto \dfrac{\varepsilon}{nx^{n-1}(b-a)} = \dfrac{\varepsilon}{f'(x)(b-a)}$.

3. On peut appliquer la méthode présentée en 1. et 2. à toute fonction $f : [a,b] \to \mathbb{R}$ de classe $\mathcal{C}^1$ sur $[a,b]$ en utilisant une primitive g de f. Le lecteur pourra le vérifier à titre d'exercice. On suppose que f n'est pas constante. Soit $M := \sup\limits_{x\in[a,b]} \big|f'(x)\big|$.
   Si $\varepsilon > 0$ et si δ est la jauge constante définie par $\delta := \frac{\varepsilon}{M(b-a)}$, alors, pour toute subdivision δ-fine $\mathcal{D}$, on a $\Big|S_{\mathcal{D}}(f) - \displaystyle\int_a^b f\Big| \leqslant \varepsilon$. On peut aussi (si f' ne s'annule pas) utiliser la jauge variable $\delta : x \mapsto \frac{\varepsilon}{\big|f'(x)\big|(b-a)}$.
   Si f est *monotone*, on peut reprendre l'analyse faite plus haut.

Nous proposons maintenant un exemple simple de fonction qui n'est pas $\mathscr{R}$-intégrable sur un segment mais est $\mathscr{HK}$-intégrable sur ce segment. Notre preuve utilise l'approche introduite dans les exemples ci-dessus.

Exemple. Nous allons voir que la fonction $f : [0,1] \to \mathbb{R}$ définie par $f(x) = 1/\sqrt{x}$ si $x > 0$ et $f(0) = 0$ est $\mathscr{HK}$-intégrable sur $[0,1]$ et que $\displaystyle\int_0^1 f = 2$. Cette fonction n'est pas intégrable au sens de Riemann sur $[0,1]$ car elle n'est pas bornée.

La fonction f est intégrable au sens de Riemann, donc $\mathscr{HK}$-intégrable, sur tout *segment* $[a,1]$, $0 < a < 1$. Le problème vient de la discontinuité à l'origine et du fait que $\lim\limits_{x\to 0^+} f(x) = +\infty$. Comme dans les exemples ci-dessus on va utiliser la connaissance d'une *primitive* de f sur $f :]0,1] \to \mathbb{R}$: la dérivée de la fonction $g : x \mapsto 2\sqrt{x}$ sur $]0,1[$ est égale à f. Le point important est que $\lim\limits_{x\to 0^+} g(x) = 0$. On peut ainsi prolonger g en une fonction continue sur $[0,1]$. Cela suggère la valeur de l'intégrale cherchée : $g(1) - g(0) = 2$ et la construction des jauges en adaptant la méthode des exemples ci-dessus.

Soit $\varepsilon > 0$. Pour toute subdivision pointée $\mathcal{D} := \big(([x_k, x_{k+1}], \xi_k)\big)_{0\leqslant k\leqslant N-1}$, il s'agit de majorer $\big|S_{\mathcal{D}}(f) - 2\big|$. On a :

$$S_{\mathcal{D}}(f) - 2 = S_{\mathcal{D}}(f) - \big(g(1) - g(0)\big) = \sum_{k=0}^{N-1} \big(f(\xi_k)(x_{k+1} - x_k) - g(x_{k+1}) + g(x_k)\big).$$

La jauge doit permettre d'une part de contrôler le premier terme :

$$f(\xi_0)(x_1 - x_0) - g(x_1) + g(x_0)$$

de la sommation du deuxième membre de l'égalité ci-dessus et d'autre part la partie restante :

$$\sum_{k=1}^{N-1} \left(f(\xi_k)(x_{k+1} - x_k) - g(x_{k+1}) + g(x_k)\right).$$

– Pour contrôler le premier terme la jauge va imposer le marquage du premier segment $[x_0 = 0, x_1]$ en $\xi_0 = 0$. Il suffit pour cela de choisir δ de telle sorte que $\delta(x) < x$ pour tout $0 < x \leqslant 1$.

– Pour contrôler la partie restante, on va utiliser la dérivée $f' : x \in]0,1] \mapsto -\frac{1}{2}x^{-3/2}$ et l'analyse faite dans les exemples ci-dessus (f est de classe $\mathcal{C}^1$ sur $]0,1]$). On choisira donc $\delta :]0,1] \to \mathbb{R}$ de la forme $\delta : x \mapsto \frac{C\varepsilon}{|f'(x)|} = 2C\varepsilon x^{3/2}$, la constante $C > 0$ restant à déterminer. On va voir que $C := 2^{-5/2}$ convient.

Fixons $0 < \varepsilon < 1$ et considérons la jauge δ sur $[0,1]$ définie par :

$$\delta : x \mapsto \delta(x) = \begin{cases} \varepsilon(x/2)^{3/2} & \text{si } x > 0 \\ \varepsilon^2/16 & \text{si } x = 0. \end{cases}$$

Nous allons montrer que si $\mathcal{D} := \big\{\big([x_k, x_{k+1}], \xi_k\big)\big\}_{0 \leqslant k \leqslant N-1}$ est une subdivision pointée δ-fine de $[0,1]$, alors $\big|S_{\mathcal{D}}(f) - 2\big| \leqslant \varepsilon$.
Notons que pour $0 < x \leqslant 1$, on a $\delta(x) \leqslant (x/2)^{3/2} < x/2$.
Comme pour $x > 0$, on a $\delta(x) < x$, on a nécessairement $\xi_0 = x_0 = 0$, $x_1 \leqslant \varepsilon^2/16$ et $f(\xi_0)(x_1 - x_0) = g(x_0) = 0$. Donc :

$$\big|f(\xi_0)(x_1 - x_0) - g(x_1) + g(x_0)\big| = 2\sqrt{x_1} < \frac{\varepsilon}{2}\cdot$$

Pour $1 \leqslant k \leqslant N-1$, on reprend l'argument des exemples de la page 255 :

$$\big|f(\xi_k)(x_{k+1} - x_k) - g(x_{k+1}) + g(x_k)\big| = (x_{k+1} - x_k)\big|f(\xi_k) - f(c_k)\big|$$

avec $c_k \in]x_k, x_{k+1}[$ et $\big|f(\xi_k) - f(c_k)\big| = |\xi_k - c_k|\,\big|f'(d_k)\big|$ avec $d_k \in]\xi_k, c_k[$. On a alors $|\xi_k - c_k| \leqslant |x_{k+1} - x_k| \leqslant \delta(\xi_k)$ et $d_k \geqslant \xi_k - \delta(\xi_k) \geqslant \xi_k - \xi_k/2 = \xi_k/2$ (puisque $\delta(\xi_k) \leqslant \xi_k/2$). Donc :

$$\big|f(\xi_k) - f(c_k)\big| \leqslant \frac{\delta(\xi_k)}{2|d_k|^{3/2}} \leqslant \frac{\delta(\xi_k)}{2(\xi_k/2)^{3/2}} = \frac{\varepsilon}{2}\cdot$$

Par conséquent, en utilisant $g(x_0) = g(0) = 0$ et $g(x_N) = g(1) = 2$:

$$\begin{aligned}
\big|S_{\mathcal{D}}(f) - 2\big| &= |S_{\mathcal{D}}(f) - g(x_N) + g(x_0)| \\
&\leqslant \big|f(\xi_0)(x_1 - x_0) - g(x_1) + g(x_0)\big| \\
&\quad + \sum_{k=1}^{N-1} \big|f(\xi_k)(x_{k+1} - x_k) - g(x_{k+1}) + g(x_k)\big| \\
&\leqslant \frac{\varepsilon}{2} + \sum_{k=1}^{N-1} (x_{k+1} - x_k)\frac{\varepsilon}{2} = \frac{\varepsilon}{2} + (1 - x_1)\frac{\varepsilon}{2} \leqslant \varepsilon.
\end{aligned}$$

Dans cet exemple nous avons montré la $\mathscr{HK}$-intégrabilité de f sur $[0,1]$ en utilisant la connaissance d'une primitive g admettant une limite à droite à l'origine. Du point de vue de l'intégrale de Riemann, cela signifie que f est intégrable *au sens de Riemann généralisé* sur $]0,1]$. Plus précisément, pour tout $0 < a < 1$, $\displaystyle\int_a^1 f = g(1) - g(a) = 2 - 2\sqrt{a}$

et $\lim\limits_{a\to 0^+}\displaystyle\int_a^1 f = 2$. Ainsi l'intégrale au sens de Riemann généralisé est ègale à 2 c'est-à-dire à l'intégrale au sens de Henstock-Kurzweil. Nous verrons plus loin que cela peut être généralisé (*cf.* le corollaire 92 de la page 277).

La fonction g définie sur $]0,1]$ par $g(x) = 1/x$ n'est pas intégrable au sens de Riemann *généralisé* sur $]0,1]$. En utilisant le corollaire 92 on peut en déduire que la fonction prolongée par 0 à l'origine n'est pas $\mathscr{HK}$ -intégrable sur $[0,1]$.
Si f désigne la fonction de l'exemple ci-dessus, on a $g = f^2$. Ainsi le produit de deux fonctions $\mathscr{HK}$ -intégrables sur un segment n'est pas nécessairement $\mathscr{HK}$ -intégrable, contrairement à ce qui se passe pour l'intégrale de Riemann.

Le produit de deux fonctions $\mathscr{HK}$ -intégrables sur un segment n'est pas nécessairement $\mathscr{HK}$ -intégrable sur ce segment.

L'exercice suivant montre en particulier qu'il existe des fonctions bornées sur un segment qui sont $\mathscr{HK}$ -intégrables sans être $\mathscr{R}$-intégrables. Le caractère très simple de la preuve permet de comprendre l'efficacité de la théorie de Henstock-Kurzweil.

Exercice 23.

Montrer que si $f : [0,1] \to \mathbb{R}$ est nulle en dehors d'un ensemble *dénombrable* $Z := \{z_k\}_{k\in\mathbb{N}^*}$, alors f est $\mathscr{HK}$ -intégrable sur $[0,1]$ et $\displaystyle\int_0^1 f = 0$.

Solution. Considérons une jauge δ sur $[0,1]$ définie par :

$$\delta(x) := \begin{cases} 1 & \text{si } x \in [0,1]\setminus Z \text{ et} \\ \dfrac{\varepsilon}{2^k}\dfrac{1}{1+\left|f(z_k)\right|} & \text{si } x = z_k. \end{cases}$$

Si $\mathcal{D}$ est une subdivision δ-fine de $[0,1]$, les seuls couples $(J,\xi) \in \mathcal{D}$ qui contribuent à la somme de Riemann sont ceux pour lesquels pour lesquels $f(\xi) \neq 0$. On a alors $\xi = z_k \in Z$ et :

$$\left|f(\xi)\operatorname{long}(J)\right| \leqslant \delta(z_k)\left|f(z_k)\right| \leqslant \frac{\varepsilon}{2^k}.$$

De plus, pour chaque point z_k, il y a au plus deux intervalles J avec $(J,z_k) \in \mathcal{D}$. Par conséquent :

$$\left|S_{\mathcal{D}}(f)\right| \leqslant 2\sum_{k\geqslant 1}\frac{\varepsilon}{2^k} = 2\varepsilon.$$

Cet exercice montre que la fonction caractéristique de $\mathbb{Q}$ (qui vaut 1 quand x est rationnel et 0 sinon), dite *fonction de Dirichlet* est $\mathscr{HK}$ -intégrable sur $[0,1]$. Rappelons que cette fonction n'est pas intégrable au sens de Riemann (*cf.* le module *Intégration* du livre [L2]).

Exercice 24.

1. Soit $f : [a,b] \to \mathbb{R}_+$ une fonction $\mathscr{HK}$-intégrable sur $[a,b]$ telle que $\displaystyle\int_a^b f = 0$. Soit $0 \leqslant g \leqslant f$. Montrer que g est $\mathscr{HK}$-intégrable sur $[a,b]$ et que $\displaystyle\int_a^b g = 0$.
2. Soit $A \subset [a,b]$ tel que sa fonction caractéristique 1_A soit $\mathscr{HK}$-intégrable sur $[a,b]$ et que $\displaystyle\int_a^b 1_A = 0$. Montrer que, pour tout $B \subset A$, 1_B est $\mathscr{HK}$-intégrable sur $[a,b]$ et $\displaystyle\int_a^b 1_B = 0$.

Solution.

1. Soit $\varepsilon > 0$, il existe une fonction jauge δ sur $[a,b]$ telle que si $\mathcal{D}$ est une subdivision δ-fine de $[a,b]$, alors $0 \leqslant S_{\mathcal{D}}(f) = \left|S_{\mathcal{D}}(f) - \displaystyle\int_a^b f\right| \leqslant \varepsilon$. Puisque $0 \leqslant g \leqslant f$, on a $0 \leqslant S_{\mathcal{D}}(g) = \left|S_{\mathcal{D}}(g) - 0\right| \leqslant S_{\mathcal{D}}(f) \leqslant \varepsilon$, donc g est $\mathscr{HK}$-intégrable sur $[a,b]$ et $\displaystyle\int_a^b g = 0$.
2. On applique 1, avec $f := 1_A$ et $g := 1_B$.

La proposition ci-dessous généralise le résultat de l'exercice 23. La preuve est un peu technique et peut être omise en première lecture. Pour les notions d'ensemble négligeable et de propriété vraie presque partout, on renvoie à la partie *Caractérisation d'une fonction intégrable au sens de Riemann* du module *Intégration* du livre [L2].

Proposition 73. Soit $f : [a,b] \to \mathbb{R}$ une fonction presque partout nulle.
Alors f est $\mathscr{HK}$-intégrable sur $[a,b]$ et $\displaystyle\int_a^b f = 0$.

En particulier, si $A \subset [a,b]$ est négligeable, alors $\displaystyle\int_a^b 1_A = 0$.

Démonstration.

Quitte à modifier la valeur de f en a et b, ce qui ne change rien au problème, on peut supposer que $f(a) = f(b) = 0$. Notons $A := \{x \in [a,b] \mid f(x) \neq 0\}$. Par hypothèse, A est négligeable. Notons, pour tout $m \in \mathbb{N}^*$, $A_m := \{x \in [a,b] \mid m-1 < \left|f(x)\right| \leqslant m\}$. Les $A_m \subset A \subset]a,b[$ sont négligeables et l'on a une *partition* $A = \bigcup\limits_{m\in\mathbb{N}^*} A_m$ (on a découpé A « à la Lebesgue »).

Soit $\varepsilon > 0$, pour tout $m \in \mathbb{N}^*$ il existe une famille $(I_{m,k})_{k\in\mathbb{N}}$ d'intervalles *ouverts* de $\mathbb{R}$, contenus dans $]a,b[$, telle que $A_m \subset \bigcup\limits_{k\in\mathbb{N}} I_{m,k}$ et $\sum\limits_{k\in\mathbb{N}} \operatorname{long} I_{m,k} \leqslant \frac{\varepsilon}{m2^m}$.

Si $J =]c, d[$ est un sous-intervalle de $[a, b]$ et $x \in [a, b]$, on note $d(x, \partial J) := \min(|x - c|, |x - d|)$. Si $x \in J$ et $0 < \eta \leqslant \frac{1}{2} d(x, \partial J)$, on a $[x - \eta, x + \eta] \subset J$. Alors, si J' est un intervalle de longueur η contenant x, on a $J' \subset J$.
Si $x \in A$, il existe un unique entier $m(x) \in \mathbb{N}^*$ tel que $x \in A_{m(x)}$ et l'on note :

$$k(x) := \inf \{k \in \mathbb{N} \mid x \in I_{m(x),k}\}.$$

L'idée de la preuve est de construire une jauge δ de façon à forcer, pour toute une subdivision pointée δ-fine $\mathcal{D} := \big((J_i, \xi_i)\big)_{i \in [\![0, n-1]\!]}$, le segment J_i à être contenu dans $I_{m(\xi_i), k(\xi_i)}$ (pour tout $i \in [\![0, n-1]\!]$). On définit une jauge δ sur $[a, b]$ par :

$$\delta : x \to \begin{cases} 1 & \text{si} \quad x \notin A \\ \frac{1}{2} d\big(x, \partial I_{m(x),k(x)}\big) & \text{si} \quad x \in A \end{cases}$$

Soit $\mathcal{D} := \big((J_i, \xi_i)\big)_{i \in [\![0, n-1]\!]}$ une subdivision pointée δ-fine. Pour tout $m \in \mathbb{N}^*$, on note E_m l'ensemble des ξ_i tels que $\xi_i \in A$ et $m(\xi_i) = m$ et, pour tout $k \in \mathbb{N}$, $E_{m,k}$ l'ensemble des ξ_i tels que $\xi_i \in E_m$ et $k(\xi_i) = k$.
On a :

$$\begin{aligned} S(f, \mathcal{D}) &= \sum_{i \in [\![0, n-1]\!]} f(\xi_i) \operatorname{long} J_i = \sum_{\{i \in [\![0, n-1]\!] \mid x_i \in A\}} f(\xi_i) \operatorname{long} J_i \\ &= \sum_{m \in \mathbb{N}^*} \sum_{\{i \in [\![0, n-1]\!] \mid \xi_i \in E_m\}} f(\xi_i) \operatorname{long} J_i \\ &= \sum_{m \in \mathbb{N}^*} \sum_{k \in \mathbb{N}} \sum_{\{i \in [\![0, n-1]\!] \mid \xi_i \in E_{m,k}\}} f(\xi_i) \operatorname{long} J_i. \end{aligned}$$

Si $m(\xi_i) = m$ et $k(\xi_i) = k$, on a $J_i \subset I_{m,k}$ par construction de la jauge puisque l'on a $\operatorname{long} J_i \leqslant \frac{1}{2} d(\xi_i, \partial I_{m,k})$. Les J_i correspondants sont *d'intérieurs deux à deux disjoints* (ce sont des segments de subdivision), donc : $\displaystyle\sum_{\{i \in [\![0, n-1]\!] \mid \xi_i \in E_{m,k}\}} \operatorname{long} J_i \leqslant \operatorname{long} I_{m,k}$. Par ailleurs :

$$\big|f(\xi_i)\big| \leqslant m, \quad \text{donc} \quad \sum_{\{i \in [\![0, n-1]\!] \mid \xi_i \in E_m, k\}} \big|f(\xi_i)\big| \operatorname{long} J_i \leqslant m \operatorname{long} I_{m,k}.$$

On en déduit :

$$\big|S(f, \mathcal{D})\big| \leqslant \sum_{m \in \mathbb{N}^*} \sum_{k \in \mathbb{N}} m \operatorname{long} I_{m,k} \leqslant \sum_{m \in \mathbb{N}^*} m \frac{\varepsilon}{m 2^m} = \varepsilon \sum_{m \in \mathbb{N}^*} \frac{1}{2^m} = \varepsilon.$$

Ainsi f est $\mathscr{HK}$-intégrable sur $[a, b]$ et $\displaystyle\int_a^b f = 0$. ∎

Nous verrons plus loin que la réciproque de la proposition ci-dessus est vraie pour les fonctions *positives* (*cf.* la proposition 136 de la page 314 et le corollaire 139 de la page 316).

Définition 18. Soit $\varepsilon > 0$. Une jauge δ sur $[a, b]$ est dite *ε-adaptée* à $f : [a, b] \to \mathbb{R}$ si f est $\mathscr{HK}$-intégrable sur $[a, b]$ et si $\displaystyle\left|S_{\mathcal{D}}(f) - \int_a^b f\right| \leqslant \varepsilon$ pour toute subdivision δ-fine $\mathcal{D}$ de $[a, b]$.

Point Méthode

Si δ est une jauge ε-adaptée à f, toute jauge $\delta' \leqslant \delta$ est également ε-adaptée à f.
Par suite :

1. si $f_1 : [a,b] \to \mathbb{R}$ et $f_2 : [a,b] \to \mathbb{R}$ sont deux fonctions intégrables et si δ_1 (respectivement δ_2) est une jauge ε-adaptée à f_1 (respectivement f_2), alors, $\delta := \min(\delta_1, \delta_2)$ est une jauge ε-adaptée simultanément à f_1 et à f_2 ;
2. si δ' est une jauge ε-adaptée à f et $\delta_{\mathbf{c}}$ est la jauge définie dans le lemme 71 de la page 251, la jauge $\delta := \min(\delta', \delta_{\mathbf{c}})$ est ε-adaptée à f et « force » les points de $\mathbf{c}$ à être des points marqués de *toute* subdivision δ-fine.

Notons également que si $f : [a,b] \to \mathbb{R}$ est $\mathscr{HK}$-intégrable sur $[a,b]$, alors pour toute jauge δ sur $[a,b]$, on peut trouver une suite $(\mathcal{D}_n)$ de subdivisions δ-fines de $[a,b]$ telle que :

$$\lim_{n\to+\infty} S_{\mathcal{D}_n}(f) = \int_a^b f.$$

En effet, pour chaque entier naturel n, il suffit de se donner une jauge δ_n qui est $1/n$-adaptée à f sur $[a,b]$ puis de choisir une subdivision $\mathcal{D}_n$ qui est δ'_n-fine sur $[a,b]$ avec $\delta'_n := \min(\delta, \delta_n)$.

2.1.4 Premières propriétés de l'intégrale

Les propriétés suivantes sont analogues à celle que nous avons établies dans le cas de l'intégrale de Riemann dans le module *Intégration* du livre [L2].

Théorème 74 (Linéarité). Soient f et g deux fonctions $\mathscr{HK}$-intégrables sur $[a,b]$ et $\lambda \in \mathbb{R}$. Alors, λf et $f+g$ sont aussi $\mathscr{HK}$-intégrables et vérifient les égalités :

$$\int_a^b \lambda f = \lambda \int_a^b f \quad \text{et} \quad \int_a^b (f+g) = \int_a^b f + \int_a^b g.$$

Autrement dit l'ensemble $\mathscr{HK}([a,b])$ des fonctions $\mathscr{HK}$-intégrables sur $[a,b]$ est un *sous-espace vectoriel* de $\mathcal{F}([a,b])$ et l'application $f \in \mathscr{HK}([a,b]) \mapsto \displaystyle\int_a^b f \in \mathbb{R}$ est une *forme linéaire* sur $\mathscr{HK}([a,b])$.

Démonstration. Notons A l'intégrale de f sur $[a,b]$ et B l'intégrale de g sur $[a,b]$. Fixons $\varepsilon > 0$. Soit δ une jauge ε-adaptée simultanément à f et g. Soit $\mathcal{D}$ une subdivision δ-fine de $[a,b]$. On a $S_{\mathcal{D}}(\lambda f) = \lambda S_{\mathcal{D}}(f)$ et $S_{\mathcal{D}}(f+g) = S_{\mathcal{D}}(f) + S_{\mathcal{D}}(g)$ par définition des sommes de Riemann. Par conséquent :

$$\left|S_{\mathcal{D}}(\lambda f) - \lambda A\right| = |\lambda| \cdot \left|S_{\mathcal{D}}(f) - A\right| \leqslant \lambda\varepsilon$$

$$\text{et} \qquad \left|S_{\mathcal{D}}(f+g) - (A+B)\right| \leqslant \left|S_{\mathcal{D}}(f) - A\right| + \left|S_{\mathcal{D}}(g) - B\right| \leqslant 2\varepsilon.$$

Cela montre que λf et $f+g$ sont intégrables, que $\displaystyle\int_a^b \lambda f = \lambda A$ et que $\displaystyle\int_a^b (f+g) = A+B$.

■

Théorème 75 (Comparaison). Soient f et g deux fonctions $\mathscr{HK}$-intégrables sur $[a,b]$ telles que $f \geqslant g$. Alors $\displaystyle\int_a^b f \geqslant \int_a^b g$.

En particulier, si $f \geqslant 0$, alors $\displaystyle\int_a^b f \geqslant 0$.

Démonstration. Pour tout $\varepsilon > 0$, on peut trouver une jauge qui est ε-adaptée à f et g. Si $\mathcal{D}$ est une subdivision de $[a,b]$ qui est δ-fine, on a alors :

$$\int_a^b f \geqslant S_{\mathcal{D}}(f) - \varepsilon \geqslant S_{\mathcal{D}}(g) - \varepsilon \geqslant \left(\int_a^b g\right) - 2\varepsilon.$$

Le résultat suit en faisant tendre ε vers 0. ∎

Théorème 76 (Relation de Chasles). Soient $a < b < c$ trois nombres réels, et f une fonction intégrable sur $[a,b]$ et sur $[b,c]$. Alors, f est intégrable sur $[a,c]$ et :

$$\int_a^c f = \int_a^b f + \int_b^c f.$$

Remarque. Une réciproque sera établie dans le théorème 87 de la page 275.

Démonstration. On utilise le lemme 71 de la page 251 avec un seul point noté b (au lieu de $c\ldots$). On note δ_b la jauge définie dans le lemme.

Soient $A_1 := \displaystyle\int_a^b f$ et $A_2 := \displaystyle\int_b^c f$. Soient δ_1 une jauge sur $[a,b]$ qui est ε-adaptée à $f : [a,b] \to \mathbb{R}$ et δ_2 une jauge sur $[b,c]$ qui est ε-adaptée à $f : [b,c] \to \mathbb{R}$. On définit une jauge δ' sur $[a,c]$ par :

$$\delta' : x \mapsto \begin{cases} \delta_1(x) & \text{si } x \in [a,b[\\ \delta_2(x) & \text{si } x \in]b,c] \\ \min\big(\delta_1(b), \delta_2(b)\big) & \text{si } x = b \end{cases}.$$

On définit ensuite une jauge δ sur $[a,c]$ par $\delta := \min(\delta', \delta_b)$.

Soit $\mathcal{D}$ une subdivision δ-fine de $[a,c]$. D'après le lemme 71, b est un point marqué de $\mathcal{D}$ et, en l'ajoutant, si nécessaire, aux points de subdivision de $\mathcal{D}$, on obtient une subdivision $\mathcal{D}'$ telle que $S_{\mathcal{D}'}(f) = S_{\mathcal{D}}(f)$. On peut alors « découper » $\mathcal{D}'$ en une subdivision $\mathcal{D}_1$ de $[a,b]$ et une subdivision $\mathcal{D}_2$ de $[b,c]$ (de telle sorte que $\mathcal{D}'$ soit obtenue par union de $\mathcal{D}_1$ et $\mathcal{D}_2$). On a $S_{\mathcal{D}}(f) = S_{\mathcal{D}'}(f) = S_{\mathcal{D}_1}(f) + S_{\mathcal{D}_2}(f)$. Par construction $\mathcal{D}_1$ est une subdivision δ_1-fine de $[a,b]$ et $\mathcal{D}_2$ est une subdivision δ_2-fine de $[b,c]$, donc :

$$\big|S_{\mathcal{D}}(f) - (A_1 + A_2)\big| \leqslant \big|S_{\mathcal{D}_1}(f) - A_1\big| + \big|S_{\mathcal{D}_2}(f) - A_2\big| \leqslant 2\varepsilon.$$

Cela montre que $f : [a,c] \to \mathbb{R}$ est intégrable, que δ est une jauge 2ε-adaptée et que

$$\int_a^c f = A_1 + A_2 = \int_a^b f + \int_b^c f.$$ ∎

Il se peut que $b \leqslant a$. Afin que la relation de Chasles soit vérifiée dans ce cas, si $a = b$ on pose $\displaystyle\int_a^b f := 0$ et si $a > b$, on pose $\displaystyle\int_a^b f := -\int_b^a f$.

Soient $\tau \in \mathbb{R}$, on lui associe la translation $T : x \in \mathbb{R} \mapsto T(x) := x + \tau \in \mathbb{R}$. Si $f : [a,b] \to \mathbb{R}$, on définit sa translatée par T. C'est la fonction $T(f)$ définie sur $[a+\tau, b+\tau]$ par $T(f) : x \mapsto f(x-\tau)$, c'est-à-dire $f \circ T^{-1}$. En utilisant les définitions de la $\mathscr{HK}$-intégrabilité et de l'intégrale, on vérifie facilement le résultat suivant.

Proposition 77. Soient $f : [a,b] \to \mathbb{R}$ une fonction et $T : \mathbb{R} \to \mathbb{R}$ une translation. Alors f est $\mathscr{HK}$-intégrable sur $[a,b]$ si, et seulement si, $T(f)$ est $\mathscr{HK}$-intégrable sur $[T(a), T(b)]$. On a alors $\displaystyle\int_a^b f = \int_{T(a)}^{T(b)} T(f)$.

Le résultat qui suit permet de garantir l'intégrabilité d'une fonction sans pour autant connaître la valeur de l'intégrale. Il s'agit d'un critère de Cauchy portant sur les sommes de Riemann.

Théorème 78 (Critère de Cauchy). Une fonction $f : [a,b] \to \mathbb{R}$ est $\mathscr{HK}$-intégrable si, et seulement si, pour tout $\varepsilon > 0$, il existe une jauge δ sur $[a,b]$ telle que pour toutes subdivisions $\mathcal{D}$ et $\mathcal{D}'$ δ-fines de $[a,b]$ on ait :

$$\left|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\right| \leqslant \varepsilon.$$

Une telle jauge est ε-adaptée à f sur $[a,b]$.

Démonstration. Si f est $\mathscr{HK}$-intégrable, d'intégrale égale à A, et si δ est $\varepsilon/2$-adaptée à f, alors $\left|S_{\mathcal{D}}(f) - A\right| \leqslant \varepsilon/2$ et $\left|S_{\mathcal{D}'}(f) - A\right| \leqslant \varepsilon/2$, d'où $\left|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\right| \leqslant \varepsilon$.

Réciproquement, supposons que pour tout $n \in \mathbb{N}^*$, il existe une jauge δ_n telle que si $\mathcal{D}$ et $\mathcal{D}'$ sont des subdivisions δ_n-fines de $[a,b]$, alors $\left|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\right| \leqslant 1/n$. Quitte à remplacer δ_n par $\min(\delta_1, \delta_2, \ldots, \delta_n)$, on peut supposer la suite $(\delta_n)_{n\in\mathbb{N}^*}$ décroissante. Soit $(\mathcal{D}_n)_{n\in\mathbb{N}^*}$ une suite de subdivisions δ_n-fines. Comme la suite (δ_n) est décroissante, $\mathcal{D}_m$ est δ_n-fine pour tout $m \geqslant n$. Par conséquent, pour tout $m \geqslant n$, $\left|S_{\mathcal{D}_n}(f) - S_{\mathcal{D}_m}(f)\right| \leqslant 1/n$. La suite $\left(S_{\mathcal{D}_n}(f)\right)_{n\in\mathbb{N}^*}$ est donc une suite de Cauchy qui converge vers une limite A.

De plus, si $\mathcal{D}$ est une subdivision δ_n-fine de $[a,b]$, on a :

$$\left|S_{\mathcal{D}}(f) - A\right| \leqslant \left|S_{\mathcal{D}}(f) - S_{\mathcal{D}_n}(f)\right| + \left|S_{\mathcal{D}_n}(f) - A\right| \leqslant \frac{2}{n}\cdot$$

La fonction $f : [a,b] \to \mathbb{R}$ est donc intégrable, d'intégrale égale à A.

Soit $\varepsilon > 0$. Supposons que δ soit une jauge sur $[a,b]$ telle que pour toutes subdivisions $\mathcal{D}$ et $\mathcal{D}'$ δ-fines de $[a,b]$ on ait $\left|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\right| \leqslant \varepsilon$. Montrons que δ est ε-adaptée à f sur $[a,b]$. On reprend la construction faite ci-dessus. Quitte à remplacer, pour tout $n \in \mathbb{N}$, δ_n par $\min(\delta, \delta_n)$, on peut supposer $\delta_n \leqslant \delta$. La suite $(\mathcal{D}_n)_{n\in\mathbb{N}^*}$ est alors une

suite de subdivisions δ-*fines* de $[a,b]$ telle que la suite $(S_{\mathcal{D}_n}(f))_{n\in\mathbb{N}^*}$ converge vers A. Par suite, pour tout $\eta > 0$, il existe $m \in \mathbb{N}^*$ tel que $\left|S_{\mathcal{D}_m}(f) - \int_a^b f\right| \leqslant \eta$.

Soit $\mathcal{D}$ une subdivision δ-fine de $[a,b]$, on a :

$$\left|S_{\mathcal{D}}(f) - S_{\mathcal{D}_m}(f)\right| \leqslant \varepsilon \quad \text{et} \quad \left|S_{\mathcal{D}_m}(f) - \int_a^b f\right| \leqslant \eta,$$

donc $\left|S_{\mathcal{D}}(f) - \int_a^b f\right| \leqslant \varepsilon + \eta$, et, puisque $\eta > 0$ est arbitraire, $\left|S_{\mathcal{D}}(f) - \int_a^b f\right| \leqslant \varepsilon$. La jauge δ est donc ε-adaptée à f sur $[a,b]$. ■

La relation de Chasles peut être complétée par l'énoncé suivant, qui affirme en particulier que la restriction d'une fonction intégrable sur $[a,b]$ à un sous-segment $[c,d] \subset [a,b]$ y est intégrable.

Proposition 79. Si $f : [a,b] \to \mathbb{R}$ est intégrable sur $[a,b]$, alors f est intégrable sur $[c,d]$ pour tout segment $[c,d] \subset [a,b]$. Si δ est une jauge ε-adaptée à f sur $[a,b]$, alors δ est une jauge ε-adaptée à f sur $[c,d]$.

Démonstration. Soit δ une jauge ε-adaptée à f sur $[a,b]$. Soient Δ et Δ' deux subdivisions δ-fines de $[c,d]$. Quitte à les compléter par une subdivision δ-fine de $[a,c]$ et par une subdivision δ-fine de $[d,b]$, on obtient des subdivisions δ-fines $\mathcal{D}$ et $\mathcal{D}'$ de $[a,b]$ qui coïncident en dehors de $[c,d]$. Alors :

$$\left|S_{\Delta}(f) - S_{\Delta'}(f)\right| = \left|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\right| \leqslant \varepsilon.$$

Le critère de Cauchy est vérifié, f est intégrable sur $[c,d]$ et δ est une jauge ε-adaptée à f sur $[c,d]$. ■

Le résultat suivant fournit un critère de $\mathscr{HK}$-intégrabilité. Il est basé sur le critère de Cauchy. Certains auteurs anglo-saxons l'appellent « Squeeze Theorem ». On notera l'analogie avec la définition de la $\mathscr{R}$-intégrabilité à partir des fonctions en escalier.

Proposition 80 (Squeeze Theorem). Soit $f : [a,b]$. Les deux conditions suivantes sont équivalentes :

(i) f est $\mathscr{HK}$-intégrable.

(ii) pour tout $\varepsilon > 0$, il existe des fonctions $\mathscr{HK}$-intégrables φ et ψ, telles que :

$$\psi \leqslant f \leqslant \varphi \quad \text{et} \quad \int_a^b (\varphi - \psi) \leqslant \varepsilon.$$

Démonstration.

Montrons (i) $\Rightarrow$ (ii). Il suffit de choisir $\varphi = \psi := f$.

Montrons (ii) $\Rightarrow$ (i). Soit $\varepsilon > 0$. Par hypothèse, il existe des fonctions φ et ψ $\mathscr{HK}$-intégrables sur $[a,b]$ telles que $\psi \leqslant f \leqslant \varphi$ et $\int_a^b (\varphi - \psi) \leqslant \varepsilon/2$. Par ailleurs,

il existe une jauge δ' (respectivement δ'') telle que, pour toute subdivision δ'-fine (respectivement δ''-fine) $\mathcal{D}$ l'on ait $\left|S_{\mathcal{D}}(\varphi) - \int_a^b \varphi\right| \leqslant \varepsilon/4$ (respectivement $\left|S_{\mathcal{D}}(\psi) - \int_a^b \psi\right| \leqslant \varepsilon/4$). On définit une jauge $\delta := \min(\delta', \delta'')$. Pour toute subdivision δ-fine $\mathcal{D}$, on a ;

$$S_{\mathcal{D}}(\psi) \leqslant S_{\mathcal{D}}(f) \leqslant S_{\mathcal{D}}(\varphi), \quad \left|S_{\mathcal{D}}(\varphi) - \varphi\right| \leqslant \varepsilon/4 \quad \text{et} \quad \left|S_{\mathcal{D}}(\psi) - \psi\right| \leqslant \varepsilon/4.$$

Soient $\mathcal{D}$ et $\mathcal{D}'$ deux subdivisions δ-fines. On a :

$$\int_a^b \psi - \varepsilon/4 \leqslant S_{\mathcal{D}}(\psi) \leqslant S_{\mathcal{D}}(f) \leqslant S_{\mathcal{D}}(\varphi) \leqslant \int_a^b \varphi + \varepsilon/4$$

$$-\int_a^b \varphi - \varepsilon/4 \leqslant -S_{\mathcal{D}'}(\varphi) \leqslant -S_{\mathcal{D}'}(f) \leqslant -S_{\mathcal{D}'}(\psi) \leqslant -\int_a^b \psi + \varepsilon/4.$$

D'où, en ajoutant les deux inégalités ci-dessus :

$$-\int_a^b (\varphi - \psi) - \varepsilon/2 \leqslant S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f) \leqslant \int_a^b (\varphi - \psi) + \varepsilon/2$$

puis, en enchangeant les rôles de $\mathcal{D}$ et $\mathcal{D}'$, et en utilisant l'hypothèse :

$$\left|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\right| \leqslant \int_a^b (\varphi - \psi) + \varepsilon/2 \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon. \qquad \blacksquare$$

2.2 Le théorème fondamental de l'analyse

Dans cette partie nous complétons la partie 6.4.2 du module *Intégration* du livre [L2].
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 Leibniz et Newton :

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 éventuellement ajouter des conditions pour que l'assertion 1 (respectivement l'assertion 2) soit vraie. Cela dépend de la théorie de l'intégration choisie. Dans le module *Intégration* du livre [L2] nous avons traité le problème dans le cadre de l'intégrale de Riemann et rencontré un certain nombre de difficultés. En reprenant la question dans le cadre de l'intégrale de Henstock-Kurzweil, nous allons voir que l'assertion 1 est alors vraie *sans aucune restriction* (*cf.* ci-dessous le théorème 81). C'est en fait l'une des motivations essentielles des travaux de Denjoy, Perron, Henstock, Kurzweil (et aussi de ceux de Lebesgue même si l'assertion 1 n'est pas vraie en général dans le cadre de l'intégrale de Lebesgue...). En revanche, l'assertion 2 n'est pas vraie en général, on a seulement une assertion plus faible : *cf.* ci-dessous le théorème 86. Si l'on suppose que f est $\mathscr{HK}$-intégrable, ce résultat *ne peut pas être amélioré*, comme on peut le voir sur divers exemples. Enfin si, dans

l'assertion 1, on suppose seulement que $f = \frac{\mathrm{d}F}{\mathrm{d}x}$ *presque partout*, le résultat n'est en général plus vrai comme le montre l'exemple de *l'escalier du diable* (*cf.* page 270).

2.2.1 Intégrale d'une dérivée

Théorème 81 (Théorème fondamental).
Soit $f : [a, b] \to \mathbb{R}$ une fonction dérivable sur $[a, b]$. Alors f' est $\mathscr{HK}$-intégrable et :

$$\int_a^b f' = f(b) - f(a).$$

Démonstration. Étant donné $\varepsilon > 0$, l'hypothèse que f est dérivable implique que pour tout $x \in [a, b]$, il existe un réel $\delta(x) > 0$ tel que si $y \in [a, b]$ et si $0 < |y - x| \leqslant \delta(x)$, alors :

$$\left| \frac{f(y) - f(x)}{y - x} - f'(x) \right| \leqslant \varepsilon.$$

Par conséquent, si $y \in [a, b]$ et si $|y - x| \leqslant \delta(x)$, alors :

$$\big| f(y) - f(x) - (y - x) f'(x) \big| \leqslant \varepsilon |y - x|$$

(le cas $y = x$ étant trivial). On définit ainsi une jauge δ sur $[a, b]$.
Supposons maintenant que $\mathcal{D} := \big\{ \big([x_k, x_{k+1}], \xi_k \big) \big\}_{0 \leqslant k \leqslant N-1}$ soit une subdivision δ-fine de $[a, b]$. Alors pour tout $k \in [\![0, N-1]\!]$:

$$\begin{aligned} \big| f(x_{k+1}) - f(x_k) - (x_{k+1} - x_k) f'(\xi_k) \big| &\leqslant \big| f(x_{k+1}) - f(\xi_k) - (x_{k+1} - \xi_k) f'(\xi_k) \big| \\ &\quad + \big| f(x_k) - f(\xi_k) - (x_k - \xi_k) f'(\xi_k) \big| \\ &\leqslant (x_{k+1} - \xi_k)\, \varepsilon + (\xi_k - x_k)\, \varepsilon = (x_{k+1} - x_k)\, \varepsilon. \end{aligned}$$

En sommant ces inégalités, on obtient $\big| f(b) - f(a) - S_{\mathcal{D}}(f') \big| \leqslant (b - a)\, \varepsilon$.
Par conséquent, f' est intégrable sur $[a, b]$, la jauge δ est $(b - a) \cdot \varepsilon$ adaptée et

$$\int_a^b f' = f(b) - f(a).$$

■

Corollaire 82. Si $f : [a, b] \to \mathbb{R}$ admet une primitive F, alors f est intégrable et $\displaystyle\int_a^b f = F(b) - F(a)$.

C'est ce corollaire qui permet le calcul de beaucoup d'intégrales. On peut calculer une intégrale $\displaystyle\int_a^b f$ *si l'on connaît une primitive* de f (c'est « l'intégrale au sens de Newton »). C'était déjà le cas dans le cadre de l'intégrale de Riemann (comme nous l'avons vu sur de nombreux exemples dans les livres [L1], [L2]...), mais avec des hypothèses supplémentaires sur f.

2.2.2 Intégration par parties et changement de variable

On généralise facilement les résultats sur l'intégration par parties et le changement de variable établis pour l'intégrale de Riemann dans le module *Intégration* du livre [L2].

Proposition 83 (Intégration par parties). Soient f et g deux fonctions dérivables sur $[a,b]$. Alors, si l'une des fonctions fg' ou gf' est $\mathscr{HK}$-intégrable sur $[a,b]$, l'autre l'est aussi et l'on a :

$$\int_a^b fg' = [fg]_a^b - \int_a^b gf'.$$

Démonstration. La fonction fg est dérivable sur $[a,b]$ de dérivée $fg'+gf'$, donc, d'après le théorème fondamental (théorème 81) la fonction $fg'+gf'$ est $\mathscr{HK}$-intégrable sur $[a,b]$ et l'on a $\displaystyle\int_a^b (fg'+gf') = [fg]_a^b$. Le résultat s'en déduit. ■

Proposition 84 (Formule de changement de variable).
Soit $\varphi : [\alpha,\beta] \to \mathbb{R}$ une fonction dérivable et soit $f : \varphi\big([\alpha,\beta]\big) \to \mathbb{R}$ une fonction qui admet une primitive sur $\varphi\big([\alpha,\beta]\big)$. Alors f est $\mathscr{HK}$-intégrable sur $[\varphi(\alpha),\varphi(\beta)]$, $f\circ\varphi\cdot\varphi'$ est $\mathscr{HK}$-intégrable sur $[\alpha,\beta]$ et l'on a :

$$\int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)\,\mathrm{d}x = \int_\alpha^\beta f\circ\varphi(t)\cdot\varphi'(t)\,\mathrm{d}t.$$

Démonstration. Soit F une primitive de f. La fonction $G := F\circ\varphi$ est une primitive de $f\circ\varphi\cdot\varphi'$, donc f est $\mathscr{HK}$-intégrable sur $[\varphi(\alpha),\varphi(\beta)]$, $f\circ\varphi\cdot\varphi'$ est $\mathscr{HK}$-intégrable sur $[\alpha,\beta]$ et l'on a :

$$\int_\alpha^\beta f\circ\varphi(t)\cdot\varphi'(t)\,\mathrm{d}t = G(\beta)-G(\alpha) = F\big(\varphi(\beta)\big)-F\big(\varphi(\alpha)\big) = \int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)\,\mathrm{d}x. \quad ■$$

Notons que dans la formule de changement de variable, nous avons fait l'hypothèse que f admettait une primitive. En particulier, ce résultat est valide si f est continue sur $[a,b]$.
Il est possible de ne pas faire l'hypothèse que f admet une primitive si l'on suppose que φ est strictement croissante ou strictement décroissante. Nous énonçons le résultat mais ne le démontrerons pas (*cf.* par exemple [Bart] p. 216).

Théorème 85. Soit $\varphi : [a,b] \to [c,d]$ une application strictement monotone et dérivable. Alors, l'application $f : [c,d] \to \mathbb{R}$ est $\mathscr{HK}$-intégrable si, et seulement si, l'application $(f\circ\varphi)\cdot\varphi' : [a,b] \to \mathbb{R}$ est $\mathscr{HK}$-intégrable. On a alors :

$$\int_c^d f(x)\,\mathrm{d}x = \int_a^b f\circ\varphi(t)\cdot\big|\varphi'(t)\big|\,\mathrm{d}t.$$

2.2.3 Dérivée d'une intégrale

Théorème 86. Soit f une fonction $\mathscr{HK}$-intégrable sur $[a,b]$. Pour tout $c \in [a,b]$ l'application intégrale $F = x \in [a,b] \to \displaystyle\int_c^x f$ est *continue* et *presque partout dérivable* et sa dérivée est *presque partout* égale à f.

Démonstration. Nous prouverons plus loin la continuité de F (*cf.* la proposition 97 de la page 282), puis l'assertion concernant la dérivée (*cf.* le corollaire 101 de la page 286). ■

La fonction F n'est pas nécessairement dérivable partout, comme on le voit dans l'exemple suivant.

Exemple. Considérons la fonction $f : [-1,1] \to \mathbb{R}$ définie par $f(x) = -1$ si $x < 0$ et $f(x) = 1$ si $x \geqslant 0$. La fonction $F : [-1,1] \to \mathbb{R}$ définie par $F : x \mapsto \displaystyle\int_0^x f(t)\,\mathrm{d}t$ est la fonction valeur absolue qui n'est pas dérivable en 0.

Et, même si F est partout dérivable, sa dérivée n'est pas en général partout égale à f. Voici un exemple.

Exemple. Soient $K \subset [0,1]$ l'ensemble de Cantor et $f := 1_K$. L'ensemble de Cantor est négligeable dans $[0,1]$ (*cf.* le module *Intégration* du livre [L2]).

Pour tout $x \in [0,1]$, $K \cap [0,x] \subset K$ est négligeable, donc $\displaystyle\int_0^x f = 0$. Ainsi $F : x \mapsto \displaystyle\int_0^x f$ est la fonction nulle. Sa dérivée est nulle, elle n'est pas égale à f sur K.

2.2.4 L'escalier du diable

On peut se demander si le théorème fondamental (théorème 81) reste vrai quand l'on remplace l'hypothèse « dérivable » par « presque partout dérivable ». Nous allons voir, avec l'exemple de la fonction dite *de Cantor-Lebesgue* (due en fait à Ludwig Scheeffer, en 1885) ou *escalier du diable* que la réponse est négative. Une fonction presque partout dérivable n'est pas toujours « l'intégrale de sa dérivée ».

On va utiliser l'ensemble de Cantor $K \subset [0,1]$ (*cf.* le module *Intégration* du livre [L2]).
Rappelons quelques notations et complétons la description de K. Pour tout $n \in \mathbb{N}$ on définit un sous-ensemble fermé F_n de $[0,1]$ formé par la réunion de 2^n segments disjoints $F_{n,k}$ *de même longueur* $1/3^n$ ($k \in [1,2^n]$ croissant de gauche à droite) :

$$F_0 := [0,1],\ F_1 := [0,1/3] \cup [2/3,1],\ F_2 := [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1].$$

La longueur de F_n est $(2/3)^n$. L'ensemble de Cantor est le compact $K \subset [0,1]$ défini par $K := \displaystyle\bigcap_{n\in\mathbb{N}} F_n$.

Par ailleurs, on note $U := [0,1] \setminus K$ et, pour tout $n \in \mathbb{N}^*$, $U_n := [0,1] \setminus F_n$. Chaque U_n est réunion de $2^n - 1$ intervalles ouverts disjoints notés $U_{n,k}$ ($n \in \mathbb{N}^*$, $k \in [\![1, 2^n - 1]\!]$ croissant de gauche à droite) :

$$U_1 :=]1/3, 2/3[,\ U_2 :=]1/9, 2/9[\cup]1/3, 2/3[\cup]7/9, 8/9[.$$

On passe de U_n à U_{n+1} en *gardant* les $U_{n,k}$ renommés $U_{n+1,2k}$, puis en *ajoutant* les $U_{n+1,2k+1}$ ($k \in [\![1, 2^k - 1]\!]$). Les $U_{n+1,2k+1}$ sont tous de longueur $1/3^{k+1}$ tandis que les $U_{n+1,2k}$ sont de longueur supérieure dépendant de k.
L'idée est de construire, pour tout $n \in \mathbb{N}^*$, un « escalier » joignant $(0,0)$ à $(1,1)$ dont les $2^n - 1$ marches horizontales sont au-dessus des $U_{n,k}$ et sont régulièrement espacées en hauteur. Ainsi la marche au-dessus de $U_{n,k}$ doit être à la hauteur $k/2^n$: c'est le graphe de la fonction constante $k/2^n$ sur $U_{n,k}$.
Pour tout $n \in \mathbb{N}^*$, $n \geqslant 2$, et tout $k \in [\![2, 2^n - 2]\!]$ il existe une fonction affine unique sur $F_{n,k}$ dont le graphe joint l'extrémité de la $k^{ème}$ marche à l'origine de la $(k+1)^{ème}$. La pente de ce graphe est $(3/2)^n$.
On en déduit le résultat suivant. Pour tout $n \in \mathbb{N}^*$, il existe une *unique* fonction f_n *continue* sur $[0,1]$ et telle que :

- $f_n(0) = 0$, $f_n(1) = 1$,
- pour tout $k \in [\![1, 2^n - 1]\!]$, f_n est contante et égale à $k/2^n$ sur $U_{n,k}$,
- pour tout $k \in [\![1, 2^n]\!]$, f_n est affine sur $F_{n,k}$.

Pour tout $n \in \mathbb{N}^*$, f_n est *croissante* et, pour tout $k \in [\![1, 2^n]\!]$, la pente de f_n sur $F_{n,k}$ est indépendante de k et elle est égale à $(3/2)^n$.
Par ailleurs, on définit $f_0 : x \in [0,1] \mapsto f_0(x) := x$.
On fait les observations suivantes (*cf.* les figures III.1.2 de la page suivante).

1. Pour tout $n \in \mathbb{N}$ et tout $k \in [\![1, 2^n]\!]$, les graphes des restrictions à $F_{n,k}$ des fonctions f_n et f_{n+1} sont compris entre les deux droites horizontales d'ordonnées respectives $(k-1)/2^n$ et $k/2^n$. Il en résulte que $\sup\limits_{x \in F_{n,k}} |f_{n+1}(x) - f_n(x)| \leqslant 1/2^n$.
2. Pour tout $n \in \mathbb{N}^*$ et tout $k \in [\![1, 2^n - 1]\!]$, les restrictions à $U_{n,k}$ des fonctions f_n et f_{n+1} sont égales, donc $\sup\limits_{x \in U_{n,k}} |f_{n+1}(x) - f_n(x)| = 0$.
3. D'après 1. et 2., $\sup\limits_{x \in [0,1]} |f_{n+1}(x) - f_n(x)| \leqslant 1/2^n$.
4. On peut construire la suite $(f_n)_{n \in \mathbb{N}}$ par récurrence à partir de f_0. On passe de f_n à f_{n+1} en remplaçant $g := f_n$ sur chaque segment $[u,v] := F_{n,k}$ (où elle est affine et non constante) par la fonction continue affine par morceaux $h := f_{n+1}$ qui est égale à $g(u) := f_n(u)$ en u, à $g(v) := f_n(v)$ en v et à $\frac{g(u)+g(v)}{2} = \frac{f_n(u)+f_n(v)}{2}$ sur le segment $[\frac{2u+v}{3}, \frac{u+2v}{3}]$. On ne modifie pas f_n ailleurs.

Par construction, on a $f_{n+1} = f_n$ sur l'ouvert $U_n := [0,1] \setminus F_n$, donc, pour tout $m \geqslant n$, $f_m = f_n$ sur U_n et f_m est constante sur U_n. D'après l'observation 3. ci-dessus, pour tout $x \in [0,1]$, $\left|f_{n+1}(x) - f_n(x)\right| \leqslant 2^{-n}$. Il en résulte que la suite $(f_n)_{n \in \mathbb{N}}$ *converge uniformément* sur $[0,1]$. Notons f sa limite, c'est une fonction continue croissante et, pour tout $n \in \mathbb{N}^*$, $f = f_n$ sur U_n, donc f est constante sur U_n. Par suite f est dérivable et de dérivée nulle sur U_n, donc f est dérivable et de dérivée nulle sur $\bigcup\limits_{n \in \mathbb{N}^*} U_n$, c'est-à-dire sur $[0,1] \setminus K$. Puisque K est négligeable, f est *presque partout dérivable et de dérivée nulle*.

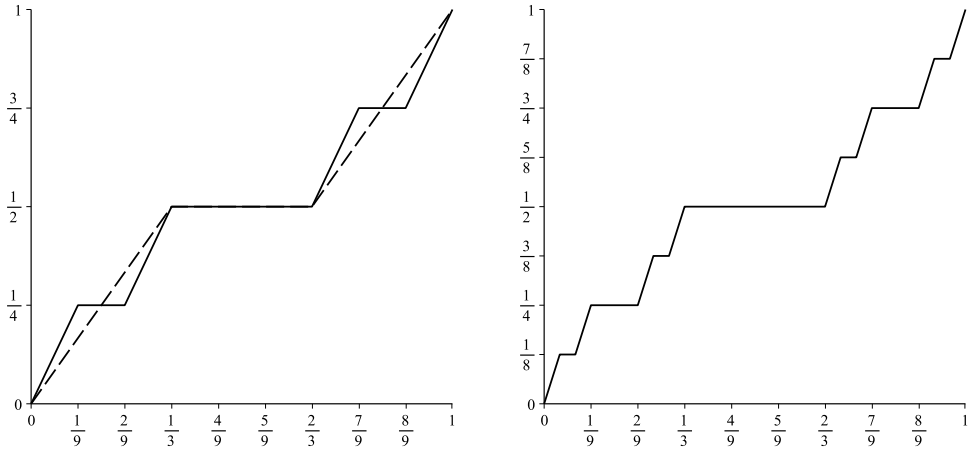


FIGURE III.1.2 – L'escalier du diable

On prolonge f' arbitrairement sur K en conservant la notation. La fonction f' ainsi définie étant presque partout nulle elle est $\mathscr{HK}$ -intégrable et l'on a $\displaystyle\int_0^1 f' = 0 \neq 1 = f(1) - f(0)$ (*cf.* proposition 73 de la page 261).

Exercice 25.

On considère la suite $(f_n)_{n\in\mathbb{N}}$ des fonctions de l'escalier du diable.
Calculer $\displaystyle\sup_{x\in[0,1]} |f_{n+1}(x) - f_n(x)|$.

Solution.

a) Soit $n \in \mathbb{N}$; la longueur de $F_{n,k}$ est indépendante de $k \in [\![1, 2^n]\!]$ (elle est égale à $1/3^n$) et la pente de la restriction de f_n à $F_{n,k}$ est indépendante de k (elle est égale à $(3/2)^n$). On en déduit, en utilisant l'observation $4.$, que, pour tous $k,\ k' \in [\![1, 2^n]\!]$, l'on passe de la restriction de $f_{n+1} - f_n$ au segment $F_{n,k}$ à sa restriction au segment $F_{n,k'}$ par une translation sur la variable (qui transforme $F_{n,k}$ en $F_{n,k'}$). Il en résulte que :

$$\forall\, k \in [\![1, 2^n]\!], \qquad \sup_{x\in F_{n,k}} |f_{n+1}(x) - f_n(x)| = \sup_{x\in F_{n,1}} |f_{n+1}(x) - f_n(x)|. \tag{20}$$

b) Pour tout $n \in \mathbb{N}$, $f_n(0) = 0$ et la restriction de f_n à $F_{n,1} = [0, 1/3^n]$ est une fonction linéaire de pente $(\frac{3}{2})^n$. Celle f_{n+1} à $F_{n+1,1}$ est donc linéaire et de pente $(\frac{3}{2})^{n+1}$. Par suite $f_{n+1}(x) = \frac{1}{2} f_n(3x)$, pour tout $x \in F_{n+1,1} = \big[0, 1/3^{n+1}\big]$. On en déduit :

$$\forall\, n \in \mathbb{N}^*,\ \forall\, x \in F_{n+1,1}, \quad f_{n+1}(x) - f_n(x) = \frac{1}{2}\big(f_n(3x) - f_{n-1}(3x)\big).$$

puis :

$$\sup_{x\in F_{n,1}} |f_{n+1}(x) - f_n(x)| = \frac{1}{2} \sup_{x\in F_{n-1,1}} |f_n(x) - f_{n-1}(x)| \tag{21}$$

c) Le graphe de la restriction de f_{n+1} à $F_{n,1} = [0, 1/3^n]$ admet une *symétrie* par rapport au point $\left(\frac{1}{2\cdot 3^n}, f_n\left(\frac{1}{2\cdot 3^n}\right)\right)$. On en déduit que le maximum m_n de la fonction $|f_{n+1} - f_n|$ sur $[0, 1/3^n]$ est atteint au point $x = 1/3^{n+1}$.
En particulier le maximum m_0 de $f_1 - f_0$ sur $[0, 1]$ est atteint en $x = 1/3$ et l'on a $m_0 = 1/2 - 1/3 = 1/6$. En utilisant (21), on en déduit, par récurrence sur n, $m_n = 2^{-n}/6$, puis, en utilisant (20) :

$$\sup_{x\in[0,1]} |f_{n+1}(x) - f_n(x)| = m_n = 2^{-n}/6.$$

2.3 Intégrale de Henstock-Kurzweil sur un intervalle quelconque

Dans cette partie, $I \subset \mathbb{R}$ est un intervalle d'intérieur non vide. Nous noterons a la borne inférieure de I et b la borne supérieure. Il se peut que l'intervalle I ne soit pas borné. Dans ce cas, on a $a = -\infty$ ou $b = +\infty$.

2.3.1 Définitions et généralités

Dans le cadre de l'intégrale de Riemann nous avons introduit, dans le module intégration du livre [L2], la notion *d'intégrale généralisée* (ou impropre) sur un intervalle quelconque comme *limite d'intégrales sur des sous-segments*. On peut reprendre cette définition, *mutatis mutandis*, dans le cadre de l'intégrale de Henstock-Kurzweil en utilisant la définition 19 ci-dessous.

Définition 19. On dit qu'une fonction $f : I \to \mathbb{R}$ est *$\mathscr{HK}$-localement intégrable sur I* si elle est $\mathscr{HK}$-intégrable sur tout sous-segment de I.

Définition 20. Soit $I \subset \mathbb{R}$ un intervalle. On dit qu'une fonction $f : I \to \mathbb{R}$ est *intégrable au sens de Henstock-Kurzweil généralisé* sur I si elle est $\mathscr{HK}$-localement intégrable sur I et si la limite $\displaystyle\lim_{c\to a^+, d\to b^-} \int_c^d$ existe.
On note alors $\displaystyle\int_I f$ la limite.

En ce qui concerne la théorie de Henstock-Kurzweil, nous allons donner une définition « directe » de l'intégrale sur un intervalle quelconque comme « *limite de sommes de Riemann* ». Le fait que l'intégrale de f sur $[a, b[$ soit la limite des intégrales de f sur $[a, c]$ quand $c \to b^-$ (c'est-à-dire que f est intégrable au sens de Henstock-Kurzweil généralisé) deviendra alors un théorème : le *théorème de Hake* et la terminologie « intégrale au sens de Henstoch-Kurzweil généralisé » deviendra marginale.
Il résultera du théorème de Hake que les fonctions intégrables au sens de Riemann généralisé sont intégrables au sens de Henstock-Kurzweil et que les deux intégrales coïncident.

Définition 21. Soit $I \subset \mathbb{R}$ un intervalle. Une fonction $f : I \to \mathbb{R}$ est intégrable sur I au sens de Henstock-Kurzweil (ou $\mathscr{HK}$-intégrable) s'il existe un réel A tel que pour tout $\varepsilon > 0$, on puisse trouver une fonction jauge $\delta : I \to \mathbb{R}$ *et* un segment $[\alpha, \beta] \subset I$ de telle sorte que si $\mathcal{D}$ est une subdivision δ-fine d'un segment contenu dans I et contenant $[\alpha, \beta]$, alors $\left|S_{\mathcal{D}}(f) - A\right| \leqslant \varepsilon$. Le réel A est l'intégrale de f sur I : $A = \displaystyle\int_I f$. On dit que $\big(\delta, [\alpha, \beta]\big)$ est un couple ε-adapté à f sur I.

Pour que cette définition de l'intégrale de f sur I ait un sens, il faut, comme dans le cas d'un segment, vérifier l'unicité. Supposons donc qu'il existe deux réels A_1 et A_2 vérifiant les conditions précédentes. Alors, pour tout $\varepsilon > 0$, on peut trouver deux jauges δ_1 et δ_2 et deux sous-segments $[\alpha_1, \beta_1] \subset I$, $[\alpha_2, \beta_2] \subset I$, tels que pour toute subdivision $\mathcal{D}_1$ qui est δ_1-fine d'un segment contenu dans I et contenant $[\alpha_1, \beta_1]$, on ait $|S_{\mathcal{D}_1}(f) - A_1| \leqslant \varepsilon$ et pour toute subdivision $\mathcal{D}_2$ qui est δ_2-fine d'un segment contenu dans I et contenant $[\alpha_2, \beta_2]$, on ait $|S_{\mathcal{D}_2}(f) - A_2| \leqslant \varepsilon$. La fonction $\delta := \min(\delta_1, \delta_2)$ est une jauge sur I et :

$$[\alpha, \beta] = [\min(\alpha_1, \alpha_2), \max(\beta_1, \beta_2)] \subset I \qquad (I \text{ est convexe}).$$

Soit $\mathcal{D}$ une subdivision δ-fine d'un segment J contenu dans I et contenant $[\alpha, \beta]$. Alors $\mathcal{D}$ est à la fois δ_1-fine et δ_2-fine et J contient $[\alpha_1, \beta_1]$ et $[\alpha_2, \beta_2]$. Par conséquent :

$$|S_{\mathcal{D}}(f) - A_1| \leqslant \varepsilon \qquad \text{et} \qquad |S_{\mathcal{D}}(f) - A_2| \leqslant \varepsilon.$$

On en déduit que pour tout $\varepsilon > 0$, $|A_1 - A_2| \leqslant 2\varepsilon$. Donc, $A_1 = A_2$.

Nous dirons qu'une jauge δ sur I est ε-adaptée à f sur I, s'il existe un segment $[\alpha, \beta] \subset I$ tel que le couple $\big(\delta, [\alpha, \beta]\big)$ soit ε-adapté à f sur I. Nous dirons qu'un segment $[\alpha, \beta]$ est ε-adapté à f sur I, s'il existe une jauge δ sur I telle que le couple $\big(\delta, [\alpha, \beta]\big)$ soit ε-adapté à f sur I.

Observons que si le couple $\big(\delta, [\alpha, \beta]\big)$ est ε-adapté à f sur I, alors pour toute jauge $\delta' \leqslant \delta$ et tout segment $[\alpha', \beta'] \subset I$ contenant $[\alpha, \beta]$, le couple $\big(\delta', [\alpha', \beta']\big)$ est également ε-adapté à f sur I. En effet, diminuer la jauge et augmenter le segment ne fait que réduire l'ensemble des subdivisions $\mathcal{D}$ à considérer.

En particulier, si $I = [a, b]$ est un segment, on peut toujours choisir $[\alpha, \beta] := [a, b]$ dans la démonstration précédente. Ainsi, dans le cas d'un segment, la définition 21 est équivalente à la définition déjà donnée.

Remarque. Supposons $I :=]a, b[$. Certains auteurs donnent une définition de la $\mathscr{HK}$-intégrabilité (un peu) différente de la nôtre et se rapprochant de celle donnée sur un segment. Partons d'un couple adapté $\big(\delta, [\alpha, \beta]\big)$. On étend la jauge δ sur $]a, b[$ à $[a, b]$ en posant $\delta(a) := \alpha - a$ et $\delta(b) := b - \beta$, puis on définit une notion de « subdivision pointée de $I :=]a, b[$ » en prenant pour « premier segment » $[a, x_0]$, pointé en a et pour « dernier segment » $[x_m, b]$ pointé en b. Enfin on prolonge f par 0 en a et b. On « recopie » ensuite les définitions de l'intégrabilité et de l'intégrale données dans le cas d'un segment. Le lecteur vérifiera à titre d'exercice que cette définition est équivalente à la nôtre.

La propriété de linéarité de l'intégrale (théorème 74 de la page 263) et le théorème de comparaison (théorème 75 de la page 264) restent valides dans le cas de l'intégration sur un intervalle quelconque. La démonstration donnée dans le cadre de l'intégration sur un segment s'adapte sans problème.

On rappelle que, selon nos conventions, si I est réduit à un point, toute fonction $f : I \to \mathbb{R}$ est intégrable sur I au sens de Henstock-Kurzweil et $\displaystyle\int_I f = 0$.

Dans les deux paragraphes suivants nous allons démontrer deux résultats fondamentaux : la relation de Chasles et le théorème de Hake.

2.3.2 La relation de Chasles

Théorème 87 (Relation de Chasles). Si $f : I \to \mathbb{R}$ est intégrable sur I, alors f est intégrable sur tout sous-intervalle $J \subset I$. De plus, si $I = I_1 \cup I_2$ est la réunion de deux intervalles adjacents, alors f est $\mathscr{HK}$ -intégrable sur I si, et seulement si, elle est $\mathscr{HK}$ -intégrable sur I_1 et sur I_2. Dans ce cas :

$$\int_I f = \int_{I_1} f + \int_{I_2} f.$$

Démonstration. La démonstration de la relation de Chasles pour des intervalles quelconques se fait essentiellement comme la démonstration pour des segments.

Moyennant deux lemmes, la démonstration se ramène au cas d'un segment.

Lemme 88. Si $I' \subset I$ sont deux intervalles ayant les mêmes extrémités, une fonction $f : I \to \mathbb{R}$ est $\mathscr{HK}$ -intégrable sur I si, et seulement si, elle est $\mathscr{HK}$ -intégrable sur I'. Dans ce cas :

$$\int_I f = \int_{I'} f.$$

Démonstration. Nous allons faire la démonstration pour $I :=]a, b]$ (avec $a \in \mathbb{R} \cup \{-\infty\}$, $b \in \mathbb{R}$) et $I' :=]a, b[$. Les autres cas se démontrent de la même manière.

Soient $f :]a, b] \to \mathbb{R}$ une fonction $\mathscr{HK}$ -intégrable sur $]a, b]$ et $\big(\delta, [\alpha, b]\big)$ un couple ε-adapté à f sur $]a, b]$. On peut supposer $\delta(b) \leqslant \frac{\varepsilon}{|f(b)|+1}$. Alors $\big(\delta, [\alpha, b - \delta(b)]\big)$ est un couple ε-adapté à f sur $]a, b[$. En effet, toute subdivision $\mathcal{D}$ d'un segment $[a', b'] \subset]a, b[$ contenant $[\alpha, b - \delta(b)]$ peut être complétée par le segment pointé $\big([b', b], b\big)$ pour obtenir $\mathcal{D}^\sharp$, une subdivision δ-fine d'un segment contenant $[\alpha, b]$ $(b - b' \leqslant \delta(b))$. Alors $S_{\mathcal{D}^\sharp}(f) = S_{\mathcal{D}}(f) + f(b)(b - b')$ et :

$$\left| S_{\mathcal{D}}(f) - \int_{]a,b]} f \right| \leqslant \delta(b) \left| f(b) \right| + \left| S_{\mathcal{D}^\sharp}(f) - \int_{]a,b]} f \right| \leqslant 2\varepsilon.$$

Par conséquent, f est intégrable sur $]a, b[$ et $\displaystyle\int_{]a,b[} f = \int_{]a,b]} f$.

Inversement, supposons que f soit intégrable sur $]a, b[$. Soit $\big(\delta, [\alpha, \beta]\big)$ un couple ε-adapté à f sur $]a, b[$. On prolonge δ en b en posant $\delta(b) := \min\,(\varepsilon/(1 + |f(b)|), b - \beta)$. L'application

$$\delta_b : x \in]a, b] \mapsto \begin{cases} \frac{1}{2}|x - b| & \text{si} \quad x < b \\ 1 & \text{si} \quad x = b \end{cases}.$$

est une jauge sur $]a, b]$. Ainsi, quitte à remplacer δ par $\min(\delta, \delta_b)$, on peut supposer que, pour tout $x \in]a, b[$, $\delta(x) \leqslant \frac{1}{2}(b - x)$.

Soit $\mathcal{D} := \big\{\big([x_k, x_{k+1}], \xi_k\big)\big\}_{0\leqslant k\leqslant N-1}$ une subdivision δ-fine d'un segment $[a', b]$ contenant $[\alpha, b]$. Alors $N \geqslant 2$ et $\mathcal{D}^\flat := \big\{\big([x_k, x_{k+1}], \xi_k\big)\big\}_{0\leqslant k\leqslant N-2}$ est une subdivision δ-fine de $[a', x_{N-1}]$. La jauge a été choisie de façon à forcer $\xi_{N-1} = b$ (*cf.* l'exemple 1 de la page 249 ou le lemme 71 de la page 251). On a donc $b - x_{N-1} \leqslant \delta(b) \leqslant b - \beta$ et le segment $[a', x_{N-1}]$ contient $[\alpha, \beta]$. Par suite $\left|S_{\mathcal{D}^\flat}(f) - \int_{]a,b[} f\right| \leqslant \varepsilon$ et :

$$\left|S_{\mathcal{D}}(f) - \int_{]a,b[} f\right| \leqslant \delta(b)\,\big|f(b)\big| + \left|S_{\mathcal{D}^\flat}(f) - \int_{]a,b[} f\right| \leqslant 2\varepsilon.$$

Par conséquent, f est intégrable sur $]a, b]$ et $\int_{]a,b]} f = \int_{]a,b[} f$. ∎

Lemme 89.
Si f est $\mathscr{HK}$-intégrable sur I, alors f est $\mathscr{HK}$-intégrable sur tout segment $[c, d] \subset I$.

Démonstration. On se donne un couple $\big(\delta, [\alpha, \beta]\big)$ qui est ε-adapté à f sur I, avec $[c, d] \subset [\alpha, \beta]$. Ensuite, on procède essentiellement comme dans la proposition 79 de la page 266.

Soient Δ et Δ' deux subdivisions δ-fines de $[c, d]$. Quitte à les compléter par une subdivision δ-fine de $[\alpha, c]$ et par une subdivision δ-fine de $[d, \beta]$, on obtient des subdivisions δ-fines $\mathcal{D}$ et $\mathcal{D}'$ de $[\alpha, \beta]$ qui coïncident en dehors de $[c, d]$. Alors :

$$\big|S_\Delta(f) - S_{\Delta'}(f)\big| = \big|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\big| \leqslant \left|S_{\mathcal{D}}(f) - \int_I f\right| + \left|\int_I f - S_{\mathcal{D}'}(f)\right| \leqslant 2\varepsilon.$$

Le critère de Cauchy est vérifié, f est intégrable sur $[c, d]$ et δ est une jauge 2ε-adaptée à f sur $[c, d]$. ∎

Les deux lemmes précédents impliquent que si $f : I \to \mathbb{R}$ est $\mathscr{HK}$-intégrable sur I, alors elle est $\mathscr{HK}$-intégrable sur tout sous-intervalle $J \subset I$.

En reprenant la démonstration de la relation de Chasles pour les segments, on montre que si $I_1 = [a, b]$ ou $]a, b]$ et $I_2 = [b, c]$ ou $[b, c[$, et si f est intégrable sur I_1 et I_2, alors f est intégrable sur $I_1 \cup I_2$ et $\int_{I_1\cup I_2} f = \int_{I_1} f + \int_{I_2} f$. On utilise le lemme 88 de la page précédente pour en déduire que cette égalité reste vraie lorsque b n'appartient pas à I_1 ou n'appartient pas à I_2.

2.3.3 Le théorème de Hake

Le théorème de Hake affirme que lorsque f admet une intégrale impropre convergente sur un intervalle I au sens de Henstock-Kurzweil, alors, elle est en fait $\mathscr{HK}$-intégrable sur I, et l'intégrale est obtenue par passage à la limite sur les bornes d'intégration (en quelque sorte « il n'y a pas d'intégrale impropre » pour la théorie de Henstock-Kurzweil). Le théorème permet souvent de *calculer l'intégrale* d'une fonction sur un intervalle quelconque, en particulier sur un intervalle non borné.

Ce théorème a été établi en 1921 par le mathématicien allemand Hake dans le cadre de l'intégrale de Perron (qui est – *a posteriori* – équivalente à celle de Henstock-Kurzweil).

Théorème 90 (Théorème de Hake). Soit $I \subset \mathbb{R}$ un intervalle d'intérieur non vide, $a := \inf(I) \in \mathbb{R} \cup \{-\infty\}$ et $b := \sup(I) \in \mathbb{R} \cup \{+\infty\}$. Alors f est $\mathscr{HK}$-intégrable sur I si, et seulement si, f vérifie les deux propriétés suivantes :

(i) f est $\mathscr{HK}$-intégrable sur tout segment $[c, d] \subset \operatorname{Int} I$;

(ii) $\displaystyle\int_c^d f$ admet une limite finie quand $c \to a^+$ et $d \to b^-$.

Cette limite est alors égale à $\displaystyle\int_I f = \int_{\operatorname{Int} I} f$.

Nous reportons la preuve du théorème de Hake en fin de paragraphe. Le lecteur pourra dans un premier temps laisser cette démonstration de côté et reprendre en 2.4, ou même seulement en 2.6.

Corollaire 91. Soit $f : [a, b[\to \mathbb{R}$ une fonction localement $\mathscr{HK}$-intégrable sur $[a, b[$. Elle est $\mathscr{HK}$-intégrable sur $[a, b[$ si, et seulement si, l'intégrale $\displaystyle\int_c^{c'} f$ tend vers 0 quand c et c' ($c,\ c' \in [a, b[$) tendent simultanément vers b^-.

Démonstration. La condition est évidemment nécessaire d'après la relation de Chasles. Montrons qu'elle suffisante. On note, pour tout $c \in [a, b[$, $F(c) := \displaystyle\int_a^c f$. On a $F(c') - F(c) = \displaystyle\int_c^{c'} f$, donc, pour tout $\varepsilon > 0$, il existe $c_0 \in [a, b[$ tel que, pour tous $c,\ c' \in [c_0, b[$, $|F(c') - F(c)| \leqslant \varepsilon$. On peut donc appliquer le critère de Cauchy pour les fonctions : $F(c)$ a une limite finie quand c tend vers b^-. En utilisant le théorème de Hake, on en déduit que f est $\mathscr{HK}$-intégrable sur $[a, b[$. ∎

Corollaire 92. Soit $f : I \to \mathbb{R}$ localement intégrable *au sens de Riemann* sur I. Les deux conditions suivantes sont équivalentes :

(i) f est intégrable sur I *au sens de Riemann généralisé* ;

(ii) f est $\mathscr{HK}$-intégrable sur I.

Si elles sont satisfaites les deux intégrales coïncident.

Le corollaire ci-dessus permet d'obtenir de nombreux exemples de fonctions $\mathscr{HK}$-intégrables (*cf.* la partie *Intégrales généralisées* du module *Intégration* du livre [L2]). Cependant, il existe « beaucoup » de fonctions qui sont $\mathscr{HK}$-intégrables mais ne sont pas intégrables au sens de Riemann généralisé. Par exemple $1_{\mathbb{Q}}$ est $\mathscr{HK}$-intégrable sur $\mathbb{R}$ et l'on a $\displaystyle\int_{\mathbb{R}} 1_{\mathbb{Q}} = 0$, mais $1_{\mathbb{Q}}$ n'est pas intégrable au sens de Riemann généralisé sur $\mathbb{R}$ (elle n'est pas localement intégrable au sens de Riemann). Plus généralement on a le résultat suivant.

Proposition 93. Soit f une fonction presque partout nulle sur un intervalle I. Alors f est $\mathscr{HK}$-intégrable sur I et $\displaystyle\int_I f = 0$.

Démonstration. On peut supposer I d'intérieur non vide. Soit $N \subset I$ un ensemble négligeable tel que $f = 0$ sur $I \setminus N$. D'après la proposition 73 de la page 261, pour tout segment J contenu dans I, on a $\displaystyle\int_J f = 0$ ($N \cap J \subset J$ est un sous-ensemble négligeable). On déduit alors le résultat du théorème de Hake. ■

Corollaire 94. Soit $f : I \to \mathbb{R}_+$ une fonction *positive*. Alors f est $\mathscr{HK}$-intégrable sur I si, et seulement s'il existe $M \geqslant 0$ tel que, pour tout segment $[c,d] \subset \operatorname{Int} I$, f soit $\mathscr{HK}$-intégrable sur $[c,d]$ et $\displaystyle\int_c^d f \leqslant M$.

De plus, s'il en est ainsi, alors $\displaystyle\int_I f = \sup_{[c,d]\subset \operatorname{Int} I} \int_c^d f = \sup_{[c,d]\subset I} \int_c^d f$.

Démonstration. Si f est intégrable sur I, elle est intégrable sur tout segment $[c,d] \subset \operatorname{Int} I$ (*cf.* la proposition 89) et $M := \displaystyle\int_I f \geqslant 0$ répond à la question.

Supposons inversement que f soit $\mathscr{HK}$-intégrable sur tout segment $[c,d] \subset \operatorname{Int} I$, et que $\displaystyle\int_c^d f \leqslant M$. On note $A := \displaystyle\sup_{[c,d]\subset \operatorname{Int} I} \int_c^d f$, alors $\displaystyle\int_c^d f$ tend vers A quand $c \to a^+$ et $d \to b^-$ et, d'après le théorème de Hake, f est intégrable sur I et l'on a $\displaystyle\int_I f = A$.

On a $\displaystyle\int_I f = \sup_{[c,d]\subset \operatorname{Int} I} \int_c^d f \leqslant \sup_{[c,d]\subset I} \int_c^d f \leqslant \int_I f$, d'où le résultat. ■

Démonstration du théorème de Hake

Commençons par montrer que si f est une fonction intégrable sur I et si $a := \inf(I) \in \mathbb{R} \cup \{-\infty\}$ et $b := \sup(I) \in \mathbb{R} \cup \{+\infty\}$, alors $\displaystyle\int_I f = \lim_{\substack{c\to a^+\\ d\to b^-}} \int_c^d f$.

Soient $\varepsilon > 0$ et $\big(\delta, [\alpha,\beta]\big)$ un couple ε-adapté à f sur I. Si $[c,d]$ contient $[\alpha,\beta]$ et si $\mathcal{D}$ est une subdivision δ-fine de $[c,d]$, on a $\left| S_{\mathcal{D}}(f) - \displaystyle\int_I f \right| \leqslant \varepsilon$. Comme l'intégrale de f sur $[c,d]$ peut être approchée par de telles sommes de Riemann $S_{\mathcal{D}}(f)$, on en déduit que pour tout segment $[c,d]$ contenu dans I et contenant $[\alpha,\beta]$, on a $\left| \displaystyle\int_c^d f - \int_I f \right| \leqslant 2\varepsilon$. Par conséquent $\displaystyle\int_I f = \lim_{\substack{c\to a^+\\ d\to b^-}} \int_c^d f$.

Montrons maintenant la réciproque. Quitte à découper l'intervalle I en deux et à utiliser la relation de Chasles, on peut supposer que $I := [a,b[$ et montrer que si $\displaystyle\lim_{d\to b^-} \int_a^d f = A \in \mathbb{R}$, alors f est in-

tégrable sur I, d'intégrale égale à A. Étant donné $\varepsilon > 0$, on choisit une suite *strictement* croissante $(d_j)_{j\in\mathbb{N}}$ de limite b de sorte que $d_0 := a$ et, pour tout $j \geqslant 1$ et tout $d \geqslant d_j$, $\left|\int_a^d f - A\right| \leqslant \frac{\varepsilon}{2^j}\cdot$

On note $\mathbf{d} := \{d_j\}_{j\in\mathbb{N}}$. On choisit ensuite pour tout $j \in \mathbb{N}$ une jauge δ_j sur $[d_j, d_{j+1}]$ qui est $\varepsilon/2^j$-adaptée à f sur ces intervalles. On en déduit une jauge δ' sur $[a, b[$ définie par :

$$\delta' : x \mapsto \begin{cases} \delta_0(a) & \text{si } x = a \\ \delta_j(x) & \text{si } x \in]d_j, d_{j+1}[\\ \min\bigl(\delta_j(d_j), \delta_{j-1}(d_j)\bigr) & \text{si } x = d_j,\ j \geqslant 1 \end{cases}.$$

On définit ensuite une jauge $\delta_{\mathbf{d}}$ sur $[a, b[$ par :

$$\delta_{\mathbf{d}} : x \mapsto \begin{cases} 1 & \text{si } x = a \\ \frac{1}{2}\min\bigl(x - d_j, d_{j+1} - x\bigr) & \text{si } x \in]d_j, d_{j+1}[\\ 1 & \text{si } x = d_j,\ j \geqslant 1. \end{cases}$$

Cette jauge est fabriquée comme dans le lemme 71 de la page 251, mais avec une *infinité* de points. Sa restriction à tout sous-segment J est une jauge $d_{\mathbf{d}\cap J}$ du type du lemme 71 ($\mathbf{d} \cap J$ est *fini*), ce qui permet d'appliquer ce lemme.

Soit δ la jauge définie sur $[a, b[$ par : $\delta := \min(\delta', \delta_{\mathbf{d}})$. Nous allons montrer que $\bigl(\delta, [a, d_1]\bigr)$ est un couple ε-adapté à f.

Pour toute subdivision δ-fine $\mathcal{D}$ d'un segment $[a, d]$ avec $d_k \leqslant d < d_{k+1}$ ($k \geqslant 1$), les points $d_0, \ldots, d_k$ sont des *points marqués de la subdivision*, on peut « ajouter » les d_j, $j \in [\![0, k]\!]$, à $\mathcal{D}$ et l'on obtient une subdivision $\mathcal{D}'$ de $[a, d]$ telle que $S_{\mathcal{D}'}(f) = S_{\mathcal{D}}(f)$ (*cf.* le lemme 71 de la page 251). Celle-ci se décompose en des subdivisions $\mathcal{D}_j$ de $[d_j, d_{j+1}]$, pour $0 \leqslant j < k$, et de $[d_k, d]$ pour $j = k$ et $S_{\mathcal{D}'}(f) = \sum_{j=0}^{k} S_{\mathcal{D}_j}(f)$. On a par conséquent :

$$\forall j < k, \quad \left|S_{\mathcal{D}_j}(f) - \int_{d_j}^{d_{j+1}} f\right| \leqslant \frac{\varepsilon}{2^j} \quad \text{et} \quad \left|S_{\mathcal{D}_k}(f) - \int_{d_k}^{d} f\right| \leqslant \frac{\varepsilon}{2^k}\cdot$$

Pour la deuxième majoration, on a utilisé la proposition 79 de la page 266.

Comme $S_{\mathcal{D}}(f) = \sum_{j=0}^{k} S_{\mathcal{D}_j}(f)$ et $d \geqslant d_j$, il vient :

$$\begin{aligned} \bigl|S_{\mathcal{D}}(f) - A\bigr| \leqslant \left|S_{\mathcal{D}}(f) - \int_a^d f\right| + \left|\int_a^d f - A\right| &\leqslant \sum_{j=0}^{k-1} \left|S_{\mathcal{D}_j}(f) - \int_{d_j}^{d_{j+1}} f\right| + \frac{\varepsilon}{2^k} \\ &\leqslant \sum_{j=0}^{k-1} \frac{\varepsilon}{2^j} + \frac{\varepsilon}{2^k} \leqslant 2\varepsilon, \end{aligned}$$

pour toute subdivision δ-fine $\mathcal{D}$ d'un segment contenant $[a, d_1]$. Ainsi $(\delta, [a, d_1]$ est un couple 2ε-adapté pour f. ∎

2.4 Le lemme de Henstock

Il est possible de raffiner les résultats qui précèdent. Le lemme suivant, dû à Ralph Henstock, est simple mais très utile. Nous l'utiliserons plus loin pour établir un certain nombre de résultats importants comme le théorème de convergence monotone.

Point Méthode

Les jauges sont un outil théorique très important et il faut apprendre à calculer avec celles-ci pour se familiariser avec leurs propriétés et comprendre les preuves des résultats fondamentaux. En revanche, dans les applications, pour montrer l'intégrabilité d'une fonction et surtout pour calculer l'intégrale, il n'est pas conseillé d'utiliser directement une jauge mais plutôt d'appliquer le théorème de Hake ou les théorèmes de passage à la limite de la partie 2.7. En particulier, on peut souvent utiliser des suites de fonctions $\mathscr{R}$-intégrables (éventuellement au sens généralisé) dont la limite n'est pas nécessairement $\mathscr{R}$-intégrable.

Définition 22. Une subdivision partielle pointée d'un intervalle $I \subset \mathbb{R}$ est une collection $\mathcal{J} := \big((J_k, \xi_k)\big)_{1 \leqslant k \leqslant N}$ où les J_k sont des segments d'intérieurs deux à deux disjoints contenus dans I et où $\xi_k \in J_k$. Le support de $\mathcal{J}$ est $\operatorname{supp}(\mathcal{J}) := \bigcup_{(J,\xi)\in\mathcal{J}} J$.

Si δ est une jauge sur I, on dit que $\mathcal{J}$ est une *subdivision partielle δ-fine* lorsque pour tout $(J, \xi) \in \mathcal{J}$, on a $\operatorname{long}(J) \leqslant \delta(\xi)$.

Pour $f : I \to \mathbb{R}$, on note $S_{\mathcal{J}}(f) := \sum_{(J,\xi)\in\mathcal{J}} f(\xi) \operatorname{long}(J)$.

Lemme 95 (Henstock). Soient $f : I \to \mathbb{R}$ une fonction $\mathscr{HK}$-intégrable sur I et $\delta : I \to \mathbb{R}$ une jauge ε-adaptée à f sur I. Si $\mathcal{J}$ est une *subdivision partielle δ-fine* de I, alors :

$$\left| S_{\mathcal{J}}(f) - \sum_{(J,\xi)\in\mathcal{J}} \int_J f \right| \leqslant \varepsilon.$$

Démonstration.

Soit $\big(\delta, [\alpha, \beta]\big)$ une paire ε-adaptée à f sur I. Quitte à augmenter $[\alpha, \beta]$, on peut supposer que $[\alpha, \beta]$ contient le support de $\mathcal{J}$.

Soit $\varepsilon' > 0$. Soit $\mathcal{J} := \big\{(J_k, \xi_k)\big\}_{k\in[\![1,N]\!]}$. Le complémentaire $[\alpha, \beta] \setminus \bigcup_{k\in[\![1,N]\!]} \operatorname{Int} J_k$ peut être décomposé en une réunion finie de segments d'intérieurs deux à deux disjoints en nombre n au plus égal à $N + 1$, notés $(R_j)_{j\in[\![1,n]\!]}$.

La fonction f est $\mathscr{HK}$-intégrable sur chaque R_j, avec $j \in [\![1, n]\!]$ (*cf.* le théorème 87 de la page 275), on peut donc trouver sur chaque R_j une jauge $\delta_j \leqslant \delta$ telle que $\left|S_{\mathcal{D}_j}(f) - \int_{R_j} f\right| \leqslant \varepsilon'$, pour toute subdivision pointée $\mathcal{D}_j$ de R_j. En prenant la réunion de la famille $\mathcal{J}$ et des familles $\mathcal{D}_j$, pour $j \in [\![1, n]\!]$, on obtient une subdivision δ-fine de $[\alpha, \beta]$. La fonction f étant $\mathscr{HK}$-intégrable sur $[\alpha, \beta]$, on en déduit :

$$\left| S_{\mathcal{J}}(f) + \sum_{j\in[\![1,n]\!]} S_{\mathcal{D}_j}(f) - \int_\alpha^\beta f \right| \leqslant \varepsilon$$

On a :

$$S_{\mathcal{J}}(f) + \sum_{j\in[\![1,n]\!]} S_{\mathcal{D}_j}(f) - \int_\alpha^\beta f = S_{\mathcal{J}}(f) - \sum_{(J,\xi)\in\mathcal{J}} \int_J f + \sum_{j\in[\![1,n]\!]} \left(S_{\mathcal{D}_j}(f) - \int_{R_j} f \right)$$

$$S_{\mathcal{J}}(f) - \sum_{(J,\xi)\in\mathcal{J}} \int_J f = S_{\mathcal{J}}(f) + \sum_{j\in[\![1,n]\!]} S_{\mathcal{D}_j}(f) - \int_\alpha^\beta f - \sum_{j\in[\![1,n]\!]} \left(S_{\mathcal{D}_j}(f) - \int_{R_j} f \right)$$

On en déduit :

$$\left| S_{\mathcal{J}}(f) - \sum_{(J,\xi)\in\mathcal{J}} \int_J f \right| \leqslant \varepsilon + n\varepsilon' \leqslant \varepsilon + (N+1)\,\varepsilon'.$$

Le choix de $\varepsilon' > 0$ est arbitraire et N est indépendant de ε', donc : $\left| S_{\mathcal{J}}(f) - \sum_{(J,\xi)\in\mathcal{J}} \int_J f \right| \leqslant \varepsilon$.

Corollaire 96. Soient $f : I \to \mathbb{R}$ une fonction $\mathscr{HK}$-intégrable sur I et $\delta : I \to \mathbb{R}$ une jauge ε-adaptée à f sur I. Si $\mathcal{J} := \{(J_k, \xi_k)\}_{1\leqslant k\leqslant N}$ est une subdivision partielle δ-fine de I, alors :

$$\sum_{k\in[\![1,N]\!]} \left| f(\xi_k) \operatorname{long} J_k - \int_{J_k} f \right| \leqslant 2\varepsilon. \tag{22}$$

$$\left| \sum_{k\in[\![1,N]\!]} \left(\left| f(\xi_k) \right| \operatorname{long} J_k - \left| \int_{J_k} f \right| \right) \right| \leqslant 2\varepsilon. \tag{23}$$

Démonstration.

(i) On partage la subdivision partielle $\mathcal{J}$ en deux subdivisions partielles $\mathcal{J}^+$ et $\mathcal{J}^-$ définies par le « signe » de $f(\xi_k) \operatorname{long} I_k - \int_{J_k} f$. On définit pour cela :

$$K^+ := \{k \in [\![1,N]\!] \mid f(\xi_k) \operatorname{long} I_k - \int_{J_k} f \geqslant 0$$

$$K^- := \{k \in [\![1,N]\!] \mid f(\xi_k) \operatorname{long} I_k - \int_{J_k} f < 0.$$

On a une partition $[\![1,N]\!] = K^+ \cup K^-$ et l'on définit deux subdivisions partielles $\mathcal{J}^+$ et $\mathcal{J}^-$ par :

$$\mathcal{J}^+ := \{(J_k, \xi_k)\}_{k\in K^+} \quad \text{et} \quad \mathcal{J}^- := \{(J_k, \xi_k)\}_{k\in K^-}.$$

En appliquant le lemme de Henstock à $\mathcal{J}^+$ et $\mathcal{J}^-$ on obtient :

$$\sum_{k\in K^+} \left(f(\xi_k) \operatorname{long} J_k - \int_{J_k} f \right) \leqslant \varepsilon \quad \text{et} \quad -\sum_{k\in K^-} \left(f(\xi_k) \operatorname{long} J_k - \int_{J_k} f \right) \leqslant \varepsilon.$$

On obtient (22) en ajoutant les deux inégalités ci-dessus.

(ii) On déduit (23) de (22) en utilisant l'inégalité triangulaire : on a, pour tous $u, v \in \mathbb{R}$, $\big||u| - |v|\big| \leqslant |u - v|$.

Proposition 97. Soit f une fonction $\mathscr{HK}$-intégrable sur I.
Pour tout $c \in I$ l'application intégrale $F : x \in I \mapsto \displaystyle\int_c^x f$ est continue.

En fait, comme nous l'avons annoncé plus haut, F admet presque partout une dérivée égale à f, nous le prouverons plus loin (*cf.* le corollaire 101 de la page 286). Pour le cas où f est $\mathscr{R}$-intégrable, *cf.* l'exercice 26 ci-dessous.

Démonstration. Nous allons montrer que F est continue en utilisant le lemme de Henstock. Il suffit évidemment de faire la démonstration quand $I := [a, b]$ est un segment. Soit $d \in I$; on va montrer que F est continue en d, c'est-à-dire que $\lim\limits_{\substack{h \to 0 \\ d+h \in I}} F(x) = F(d)$.

Soient $\varepsilon > 0$ et $\delta : I \to \mathbb{R}$ une jauge ε-adaptée à f sur I.
Soient $h \in \mathbb{R}$ tel que $|h| \leqslant \delta(d)$ et $d+h \in I$. L'unique segment pointé $\big([d, d+h], \xi := d\big)$ (respectivement $\big([d+h, d], \xi := d\big)$) est une *subdivision partielle* de I. On en déduit en appliquant le lemme de Henstock (lemme 95) que $\left|f(d)\,h - \displaystyle\int_d^{d+h} f\right| \leqslant \varepsilon$. Par suite, si $|h| \leqslant \min\left(\delta(d), \frac{\varepsilon}{1+\left|f(d)\right|}\right)$:

$$\left|\int_d^{d+h} f\right| \leqslant \left|f(d)\,h\right| + \left|-f(d)\,h + \int_d^{d+h} f\right| \leqslant \left|f(d)\,h\right| + \varepsilon \leqslant \frac{\left|f(d)\right|}{1+\left|f(d)\right|}\,\varepsilon + \varepsilon \leqslant 2\varepsilon. \quad \blacksquare$$

Exercice 26.

Soient f une fonction $\mathscr{HK}$-intégrable sur I et $c \in I$. On note $F : x \in I \mapsto \displaystyle\int_c^x f$.

1. On suppose que f est continue au point $d \in I$.
   Montrer que F est dérivable au point d et que sa dérivée est $f(d)$.
2. On suppose que I est un segment et que f est $\mathscr{R}$-intégrable sur I.
   Montrer que F est *presque partout* dérivable de dérivée f sur I.

Solution.

1. Supposons f continue en $d \in [a, b]$. Étant donné $\varepsilon > 0$, soit $\eta > 0$ tel que $\left|f(x) - f(d)\right| \leqslant \varepsilon$, si $x \in [a, b]$ et $|x - d| \leqslant \eta$. Si $x \in [a, b]$ vérifie $|x - d| \leqslant \eta$ et si $J \subset [a, b]$ est le segment d'extrémités x et d, on a, en supposant $x \geqslant d$:
   $$-\varepsilon(x-d) = \int_J (-\varepsilon) \leqslant \underbrace{\int_x^d \big(f(t) - f(d)\big)\,\mathrm{d}t}_{=F(x)-F(d)-(x-d)f(d)} \leqslant \int_J \varepsilon = \varepsilon(x-d).$$
   En raisonnant de façon similaire pour $x < d$, on obtient :
   $$\left|F(x) - F(d) - (x-d)f(d)\right| \leqslant \varepsilon|x-d|.$$
   Cela montre que F est dérivable en d et que $F'(d) = f(d)$.
   On a évidemment des résultats similaires pour les dérivées à droite ou à gauche si f est continue à droite ou à gauche en d.

2. On suppose que f est $\mathscr{R}$-intégrable sur le segment I. D'après le théorème de Lebesgue (*cf.* le théorème 78 de la page 669 du livre [L2] ou le théorème 29 de la page 193 du présent ouvrage), f est presque partout continue, d'où le résultat en utilisant 1.

2.5 Lemme de Vitali et différentiabilité des intégrales indéfinies

Nous allons montrer que l'intégrale indéfinie d'une fonction $\mathscr{HK}$-intégrable f est presque partout dérivable de dérivée presque partout égale à f, complétant ainsi la preuve du théorème 86 de la page 270. La démonstration est basée sur le lemme de Henstock et sur un lemme de recouvrement dû à Vito Vitali.

Définition 23 (Recouvrement de Vitali). Soient $[a,b]$ un segment d'intérieur non vide, $\mathcal{K}$ une famille de segments d'intérieur non vide contenus dans $[a,b]$ et $A \subset [a,b]$. On dit que $\mathcal{K}$ est un *recouvrement de Vitali* de A si, pour tout $x \in A$ et tout $\varepsilon > 0$, il existe un segment $K \in \mathcal{K}$ tel que $x \in K$ et $\operatorname{long} K < \varepsilon$.

Si $\mathcal{K}$ est un *recouvrement de Vitali* de A, tout point $x \in A$ est contenu dans *une infinité* de segments de $\mathcal{K}$, donc tout recouvrement de Vitali de $A \neq \varnothing$ est *infini*. Il peut être dénombrable ou non.

Exemple. Soit $A := [0,1]$.

Pour tout $x \in \mathbb{Q} \cap [0,1]$ et tout $n \in \mathbb{N}^*$, on note $V_{x,n} := \{y \in [0,1] \mid |y-x| \leqslant 1/n\}$. On vérifie que $\mathcal{V} := \{V_{x,n}\}_{x\in\mathbb{Q}\cap[0,1], n\in\mathbb{N}^*}$ est un recouvrement de Vitali de $[0,1]$. Ce recouvrement est *dénombrable*.

Lemme 98 (Lemme de recouvrement de Vitali). Soient $\mathcal{K}$ une famille de segments contenus dans un segment $[a,b]$ et $A \subset [a,b]$. On suppose que $\mathcal{K}$ est *un recouvrement de Vitali* de A. Alors :

a) ou bien il existe une famille *finie* $(K_k)_{k\in[\![1,m]\!]}$ de segments *deux à deux disjoints* de $\mathcal{K}$ telle que $A \subset \bigcup_{k\in[\![1,m]\!]} K_k$;

b) ou bien il existe une famille *dénombrable* $(K_k)_{k\in\mathbb{N}}$ de segments *deux à deux disjoints* de $\mathcal{K}$ et une famille dénombrable $\{(L_k)\}_{k\in\mathbb{N}}$ de segments fermés contenus dans $[a,b]$ telles que :

   (i) pour tout $k \in \mathbb{N}$, $A \subset \Big(\bigcup_{h\in[\![1,k]\!]} K_h\Big) \cup \big(\bigcup_{r>k} L_r\big)$;

   (ii) $\sum_{r\in\mathbb{N}} \operatorname{long} L_r < +\infty$;

   (iii) $A \subset \Big(\bigcup_{k\in\mathbb{N}} K_k\Big) \cup N$, avec $N = \bigcap_{k\geqslant 0} \bigcup_{r>k} L_r$ *négligeable*.

   De plus on peut choisir les $\{K_k\}$ et les $\{L_k\}$ de telle sorte que, pour tout $k \in \mathbb{N}$, les segments K_k et L_k aient le même centre et que $\operatorname{long} L_k = 5 \operatorname{long} K_k$.

L'énoncé et la preuve du lemme de Vitali sont un peu techniques et peuvent être « sautés » en première lecture. Nous donnons ci-dessous une version « simplifiée » du lemme, suffisante pour plusieurs applications, en particulier pour le théorème de dérivation ci-dessous.

Démonstration. Nous allons construire, par récurrence, une suite (finie ou non) de segments deux à deux disjoints $K_0, \ldots, K_k, \ldots$ de la façon suivante.

On définit $K_0 := \varnothing$, $F_0 := \varnothing$ et, si $K_0, \ldots K_{k-1}$ $(k \geqslant 1)$ ont été définis, on note $F_k := K_1 \cup \ldots \cup K_{k-1}$. On a l'alternative suivante.

– Ou bien $A \backslash F_k = \varnothing$, c'est-à-dire $A \subset F_k$, on est dans le cas a) et l'on arrête la récurrence.

– Ou bien $A \setminus F_k \neq \varnothing$ et l'on peut poursuivre la récurrence en construisant K_k de la façon suivante. On considère la famille $\mathcal{K}_{A,k} \subset \mathcal{K}$ des intervalles K de $\mathcal{K}$ contenant *au moins un point* $x \in A$ et tels que $K \subset [a,b] \setminus F_k$. Cette famille $\mathcal{K}_{A,k}$ est *non vide*. En effet il existe un point $c \in A \setminus F_k$, on a $\varepsilon = d(c, F_k) := \inf\limits_{\{y \in F_k\}} |y - c| > 0$ (*cf.* le paragraphe *Espaces métriques compacts* du module *Espaces vectoriels normés* du livre [L2]) et, puisque $\mathcal{K}$ est un recouvrement de Vitali de A, il existe $K \in \mathcal{K}$ tel que $c \in K$ et $\operatorname{long} K < \varepsilon$. Par suite $K \subset [a,b] \setminus F_k$. Notons μ_k le sup des longueurs des intervalles $K \in \mathcal{K}_{A,k}$ $(\mu_k \leqslant b - a)$. On choisit K_k tel que $\operatorname{long} K_k \geqslant \mu_k/2$.

Si le processus s'arrête le lemme est démontré, sinon on obtient une suite de segments $(K_k)_{k \in \mathbb{N}}$ deux à deux distincts par construction. On a $\sum\limits_{k \in \mathbb{N}} \operatorname{long} K_k \leqslant b - a$, donc :

$$\frac{1}{2} \sum_{k \in \mathbb{N}} \mu_k \leqslant \sum_{k \in \mathbb{N}} \operatorname{long} K_k \leqslant b - a < +\infty.$$

Il en résulte que μ_k tend vers 0 quand k tend vers $+\infty$.

Pour tout $k \in \mathbb{N}$, on note L_k le segment de même centre que K_k tel que $\operatorname{long} L_k = 5 \operatorname{long} K_k$. Nous allons voir que les K_k et les L_k répondent à la question.

L'idée de l'assertion (i) est la suivante. La réunion $\bigcup\limits_{k \in \mathbb{N}} K_k$ ne contient pas en général A, mais nous allons voir que *pour tout* $k \in \mathbb{N}$, on peut « gonfler » les K_r, pour $r > k$, en les remplaçant par les L_r, et qu'alors la nouvelle réunion $\Big(\bigcup\limits_{h \in [\![1,k]\!]} K_h\Big) \cup \Big(\bigcup\limits_{r > k} L_r\Big)$ contient A.

Montrons (i). Soit $c \in A \backslash F_{k+1}$. Il faut montrer qu'il existe $n > k$ tel que $c \in L_n$ (n dépendant de c). Nous avons vu qu'il existe $K \in \mathcal{K}$ tel que $c \in K$ et $K \subset [a,b] \backslash F_{k+1}$. Montrons qu'il existe $p > k$ tel que $K \cap K_p \neq \varnothing$. Sinon, l'on aurait, pour tout $q > k$, $K \in \mathcal{K}_{A,q}$, donc $\operatorname{long} K \leqslant \mu_q$, et $\operatorname{long} K = 0$, puisque μ_q tend vers 0, quand q tend vers $+\infty$. On aurait donc une contradiction puisque l'intérieur de K n'est pas vide.

Soit $n := \min\{q \in \mathbb{N} \mid K \cap K_q \neq \varnothing\}$. On a $K \cap F_n = \varnothing$ (sinon n ne serait pas minimal), donc $K \in \mathcal{K}_{A,n}$ et $\operatorname{long} K \leqslant \mu_n \leqslant 2 \operatorname{long} K_n$. Puisque K rencontre K_n, la distance de c au milieu de K_n est majorée par $\operatorname{long} K + \frac{1}{2} \operatorname{long} K_n$, donc, *a fortiori*, par $\frac{5}{2} \operatorname{long} K_n$. Par suite $c \in L_n$. On a $n \geqslant k + 1$, d'où le résultat.

Montrons (ii). On a $\sum\limits_{r \in \mathbb{N}} \operatorname{long} L_r = 5 \sum\limits_{r \in \mathbb{N}} \operatorname{long} K_r < +\infty$.

Montrons (iii). On note $N = \bigcap_{k\in\mathbb{N}} \bigcup_{r>k} L_r$. D'après (ii), $\sum_{r\in\mathbb{N}, r>k} \operatorname{long} L_r$ tend vers 0 quand n tend vers $+\infty$, donc N est négligeable.

Il reste à montrer que $A \subset \left(\bigcup_{k\in\mathbb{N}} K_k\right) \cup N$. Soit $c \in A$ n'appartenant pas à $\bigcup_{k\in\mathbb{N}} K_k$. Alors, d'après (i), pour tout $k \in \mathbb{N}$, $c \in \bigcup_{r>k} L_r$. Par suite $c \in N$. ■

On déduit immédiatement du lemme 98 le résultat suivant.

Lemme 99 (Lemme de recouvrement de Vitali, version simplifiée). Soient $\mathcal{K}$ une famille de segments contenus dans un segment $[a,b]$ et $A \subset [a,b]$. On suppose que $\mathcal{K}$ est *un recouvrement de Vitali* de A. Alors, il existe une famille *au plus dénombrable* (K_k) ($k \in \mathbb{N}$ ou $k \in [\![1,m]\!]$) de segments *deux à deux disjoints* de $\mathcal{K}$ et un sous-ensemble *négligeable* $N \subset [a,b]$ tels que $A \subset (\bigcup K_k) \cup N$.

Proposition 100 (Théorème de Lebesgue-Denjoy). Soient I un segment et $f : I \to \mathbb{R}$ une fonction $\mathscr{HK}$-intégrable sur I. Alors, pour presque tout $x \in I$, on a :

$$\lim_{\substack{\operatorname{long} J \to 0 \\ x\in J}} \frac{1}{\operatorname{long} J} \int_J f = f(x) \tag{24}$$

(la limite porte sur les segments $J \subset I$ contenant x).

Démonstration.

On va prouver ce résultat en utilisant les lemmes de Henstock et de Vitali.

On désigne par A le sous-ensemble des points $x \in I$ tels que la limite du premier membre de (24) n'existe pas ou est différente de $f(x)$. On va montrer que A est *négligeable*.

Pour tout $m \in \mathbb{N}^*$, on note A_m l'ensemble des $x \in I$ tels que, pour tout $\alpha > 0$, il existe un segment J tel que $x \in J$, $0 < \operatorname{long} J \leqslant \alpha$ et :

$$\left| f(x) - \frac{1}{\operatorname{long} J} \int_J f \right| \geqslant 1/m. \tag{25}$$

On a $A = \bigcup_{m\in\mathbb{N}^*} A_m$. Nous allons montrer que les A_m sont négligeables, il en résultera que A l'est aussi.

Soient $\varepsilon > 0$ et une jauge δ (sur I) ε-adaptée à f fixés. Pour tout $m \in \mathbb{N}^*$, on note $\mathcal{V}_m$ la famille des segments pointés (J,ξ) contenus dans I vérifiant l'inégalité (25) et tels que $\operatorname{long} J \leqslant \delta(\xi)$. Cette famille est un *recouvrement de Vitali* de A_m. En effet, si $x \in A_m$, pour tout $\varepsilon > 0$, il existe J tel que $x \in J$, $\operatorname{long} J \leqslant \min(\varepsilon/2, \delta(x))$ et que J vérifie (25), donc $(J,x) \in \mathcal{V}_m$, $\operatorname{long} J \leqslant \delta(x)$ et $\operatorname{long} J < \varepsilon$. D'après le lemme de Vitali, il existe une famille *au plus dénombrable* $(J_h, \xi_h)_{h\in H}$ de segments pointés *disjoints* de $\mathcal{V}_m$ et un ensemble négligeable N tels que $A_m \subset \bigcup_{h\in H} J_h \cup N$.

Pour tout sous-ensemble fini H' de H, $(J_h, \xi_h)_{h\in H'}$ est une *sous-subdivision δ-fine* de I. En lui appliquant le lemme de Henstock, on obtient :

$$\sum_{h\in H'} \left| f(\xi_h) - \frac{1}{\operatorname{long} J_h} \int_{J_h} f \right| \operatorname{long} J_h \leqslant 2\varepsilon.$$

En utilisant (25), on en déduit :

$$\frac{1}{m}\sum_{h\in H'}\operatorname{long} J_h \leqslant \sum_{h\in H'}\left|f(\xi_h)-\frac{1}{\operatorname{long} J_h}\int_{J_h} f\right|\operatorname{long} J_h \leqslant 2\varepsilon.$$

Ainsi $\sum_{h\in H'}\operatorname{long} J_h \leqslant 2m\varepsilon$. Ce résultat étant vrai pour tout $H' \subset H$ fini, on en déduit $\sum_{k\in H}\operatorname{long} J_h \leqslant 2m\varepsilon$. Par ailleurs on peut recouvrir l'ensemble négligeable N par une famille dénombrable de segments dont la « somme » des longueurs est au plus égale à ε. On en déduit que l'on peut recouvrir A_m par une famille dénombrable de segments dont la « somme » des longueurs est au plus égale à $(2m+1)\varepsilon$. Cela est vrai pour tout $\varepsilon > 0$, donc l'ensemble A_m est négligeable, ce qui termine la preuve. ∎

Corollaire 101. Soient $f : [a,b] \to \mathbb{R}$ une fonction $\mathscr{HK}$-intégrable et $c \in [a,b]$. L'application intégrale $F : x \in [a,b] \to \int_c^x f(t)$ est presque partout dérivable et sa dérivée est presque partout égale à f.

Démonstration. On montre que F est presque partout dérivable à droite (respectivement à gauche) en appliquant la proposition précédente aux segments $[x, x+h]$ (respectivement $[x-h, x]$), avec $h > 0$ suffisamment petit. On a :

$$\frac{1}{\operatorname{long} J}\int_J f = \frac{F(x+h)-F(x)}{h}, \quad \text{respectivement} \quad \frac{1}{\operatorname{long} J}\int_J f = \frac{F(x)-F(x-h)}{h},$$

d'où le résultat. ∎

Remarques.

1. Nous utiliserons plus loin le lemme de recouvrement de Vitali pour donner une nouvelle caractérisation des sous-ensembles négligeables basée sur l'intégrale de Henstock-Kurzweil : *cf.* le corollaire 139 de la page 316 (*cf.* aussi la proposition 137 de la page 315).
2. Il existe des généralisations du lemme de recouvrement de Vitali où l'on remplace les segments par des boules ou des pavés fermés (soumis à une restriction convenable). La preuve est similaire (*cf.* par exemple [De]).

2.6 Fonctions absolument intégrables

Comme nous l'avons remarqué plus haut, contrairement à ce qui se passe en théorie de Riemann, une fonction peut être $\mathscr{HK}$-intégrable sans que sa valeur absolue le soit. Cela conduit à introduire la notion de fonction *absolument intégrable*, c'est-à-dire de fonction qui est $\mathscr{HK}$ intégrable ainsi que sa valeur absolue. Ces fonctions sont aussi appelées *intégrables au sens de Lebesgue* (*cf.* la remarque 1. ci-dessous). Nous donnerons une caractérisation des fonctions absolument intégrables basée sur la notion de *variation d'une fonction* et nous en déduirons que l'ensemble $\tilde{L}^1(I)$ des fonctions absolument intégrables sur un intervalle réel I est un espace vectoriel.

2.6.1 Définition et premières propriétés

Reprenons l'exemple de l'exercice 18 de la page 693 du livre [L2]. On considère la fonction $f : x \mapsto (\sin x)/x$ sur $[1, +\infty[$. Les fonctions f et $|f|$ sont continues donc localement intégrables au sens de Riemann (et *a fortiori* de Henstock-Kurzweil). Nous avons vu que f est $\mathscr{R}$-intégrable au sens de Riemann généralisé, elle est donc $\mathscr{HK}$-intégrable. Par contre la fonction positive $|f|$ n'est pas intégrable au sens de Riemann généralisé. On a en effet $\lim\limits_{x\to+\infty} \displaystyle\int_a^x \frac{|\sin t|}{t}\,\mathrm{d}t = +\infty$. On en déduit, en utilisant le corollaire 92 de la page 277, que $|f|$ n'est pas $\mathscr{HK}$-intégrable.

En faisant le changement de variable $x \mapsto t = 1/x$, on obtient l'intégrale $\displaystyle\int_0^1 \frac{1}{t}\sin\frac{1}{t}\,\mathrm{d}t$. Par suite la fonction $t \mapsto \frac{1}{t}\sin\frac{1}{t}$ est $\mathscr{HK}$-intégrable sur $[0,1]$ sans que sa valeur absolue le soit. Le phénomène peut donc également se produire sur un segment. Nous allons en donner un autre exemple.

Exemple. Soit $f : [0,1] \to \mathbb{R}$ la fonction définie par $f(x) := x^2\cos(\pi/x^2)$ si $x > 0$ et $f(0) := 0$. Alors f est dérivable sur $[0,1]$ avec :

$$f'(x) = \begin{cases} 0 & \text{si } x = 0 \\ 2x\cos\left(\dfrac{\pi}{x^2}\right) + \dfrac{2\pi}{x}\sin\left(\dfrac{\pi}{x^2}\right) & \text{si } x > 0. \end{cases}$$

D'après le théorème fondamental de l'analyse 81 de la page 268, f' est $\mathscr{HK}$-intégrable sur $[0,1]$ et $\displaystyle\int_0^1 f' = -1$.

Cependant, $|f'|$ n'est pas $\mathscr{HK}$-intégrable sur $[0,1]$. En effet, posons, pour tout $k \in \mathbb{N}^*$:

$$\alpha_k := \sqrt{\frac{2}{4k+1}} \qquad \text{et} \qquad \beta_k := \sqrt{\frac{1}{2k}}\cdot$$

Les segments $[\alpha_k, \beta_k]$ sont disjoints et $|f'|$ est continue, donc intégrable, sur chacun d'entre eux. De plus, si $x \in [\alpha_k, \beta_k]$, on a :

$$2k\pi \leqslant \frac{\pi}{x^2} \leqslant 2k\pi + \frac{\pi}{2}\cdot$$

Par conséquent :

$$\int_{\alpha_k}^{\beta_k} |f'| = \int_{\alpha_k}^{\beta_k} f' = f(\beta_k) - f(\alpha_k) = \frac{1}{2k}\cdot$$

Si $|f'|$ était $\mathscr{HK}$-intégrable sur $[0,1]$, on aurait, d'après la relation de Chasles :

$$\forall n \geqslant 1, \quad \sum_{k=1}^{n} \frac{1}{2k} = \sum_{k=1}^{n} \int_{\alpha_k}^{\beta_k} |f'| \leqslant \int_0^1 |f'| < +\infty.$$

Cela est impossible car la série harmonique est divergente.

Définition 24. Une fonction $f : I \to \mathbb{R}$ est *absolument intégrable* sur I si *les deux fonctions* f et $|f|$ sont $\mathscr{HK}$-intégrables sur I. Dans ce cas on dit aussi que f est *intégrable au sens de Lebesgue* sur I.

Remarques.

1. La définition donnée ci-dessus de l'intégrabilité au sens de Lebesgue n'est pas la définition « classique » introduite par Lebesgue dans sa note de 1901. Nous montrerons plus loin que les deux définitions (*a priori* très différentes...) sont équivalentes (*cf.* le théorème 167 de la page 333).

2. Dans la définition ci-dessus il ne suffit pas de supposer que $|f|$ est $\mathscr{HK}$-intégrable sur I. En effet cela n'implique pas que f est $\mathscr{HK}$-intégrable sur I. Nous en donnons ci-dessous un exemple (exercice 28) en admettant provisoirement qu'il existe une partie $A \subset [0,1]$ telle que 1_A ne soit pas $\mathscr{HK}$-intégrable sur $[0,1]$. Nous prouverons plus loin ce dernier résultat en utilisant la notion d'ensemble mesurable et en montrant l'existence d'une partie non mesurable de $[0,1]$ (*cf.* le corollaire 152 de la page 326).

   Notons toutefois que ce qui précède repose sur le fait que l'on travaille *dans le cadre traditionnel de la théorie des ensembles* (c'est-à-dire, pour les initiés, ZFC : Zermelo-Fraenkel plus axiome du choix[4]) comme toujours dans nos livres. Il est possible d'utiliser une autre axiomatique pour laquelle, pour tout sous-ensemble borné $A \subset \mathbb{R}$, 1_A est $\mathscr{HK}$-intégrable (*cf.* la remarque de la page 326).

3. Il y a une analogie classique entre l'intégrale d'une fonction et la somme d'une série numérique. Ainsi, on peut considérer les fonctions $\mathscr{HK}$-intégrables comme analogues aux séries convergentes et les fonctions intégrables au sens de Lebesgue comme analogues aux séries absolument convergentes. On peut en fait faire mieux et interpréter rigoureusement la théorie des séries numériques convergentes comme un « sous-produit » de celle de l'intégrale de Henstock-Kurzweil (*cf.* les exercices III.1.25 et III.1.26), celle des séries absolument convergentes apparaît alors comme un « sous-produit » de celle de l'intégrale de Lebesgue.

   Les fonctions intégrables au sens de Lebesgue sont, nous le verrons dans la partie 3, « associées » à une mesure, la mesure de Lebesgue sur $\mathbb{R}$. Les séries absolument convergentes sont aussi « associées » à une mesure, la mesure de comptage sur $\mathbb{N}$ (*cf.* l'exercice III.1.40).

Exercice 27.
Soient $I \subset \mathbb{R}$ un intervalle et $f : I \to \mathbb{R}$ une fonction presque partout nulle sur I. Montrer que f est absolument intégrable sur I et que $\displaystyle\int_I |f| = 0$.

Solution. La fonction $|f|$ est presque partout nulle, donc $\mathscr{HK}$-intégrable d'intégrale nulle.

---

[4] Cf. par exemple : J. L. Krivine, *Théorie des ensembles*, Cassini, 1998.

Proposition 102. Une fonction $f : I \to \mathbb{R}$ est absolument intégrable sur I si, et seulement si, $f^+ := \max(f, 0)$ et $f^- := \max(-f, 0)$ sont intégrables sur I.

Démonstration. Si f et $|f|$ sont intégrables, alors $f^+ = \frac{1}{2}\big(|f| + f\big)$ et $f^- = \frac{1}{2}\big(|f| - f\big)$ sont intégrables. Si f^+ et f^- sont intégrables, alors : $f = f^+ - f^-$ et $|f| = f^+ + f^-$ sont intégrables. ∎

Proposition 103. Si $f : I \to \mathbb{R}$ est absolument intégrable, alors $\left|\int_I f\right| \leqslant \int_I |f|$.

Démonstration. On a $-\int_I |f| \leqslant -\int_I f^- \leqslant \int_I f = \int_I f^+ - \int_I f^- \leqslant \int_I f^+ \leqslant \int_I |f|$. ∎

Exercice 28.
On admet qu'il existe une partie $A \subset [0,1]$ telle que 1_A ne soit pas $\mathscr{HK}$-intégrable sur $[0,1]$.
On considère la fonction $f : [0,1] \to \mathbb{R}$ définie par $f := 1_A - 1_{[0,1]\setminus A}$.
Montrer que $|f|$ est $\mathscr{HK}$-intégrable sur $[0,1]$. Montrer que f n'est pas $\mathscr{HK}$-intégrable sur $[0,1]$.
Solution. On a $|f| = 1$ et 1 est évidemment $\mathscr{HK}$-intégrable sur $[0,1]$. Si f était $\mathscr{HK}$-intégrable sur $[0,1]$, elle serait absolument intégrable et, d'après la proposition ci-dessus, il en serait de même de $f^+ = 1_A$, ce qui n'est pas.

2.6.2 Caractérisation des fonctions absolument intégrables et applications

La proposition suivante donne en particulier un *critère d'intégrabilité absolue*.

Proposition 104. Soit $f : I \to \mathbb{R}$ une fonction $\mathscr{HK}$-intégrable sur I.

(i) Si $|f|$ est $\mathscr{HK}$-intégrable sur I, l'on a :

$$\int_I |f| = \sup_{\mathbf{u}} \sum_{i \in [\![0,n-1]\!]} \left|\int_{[x_i, x_{i+1}]} f\right|, \tag{26}$$

le $\sup$ portant sur toutes les subdivisions $\mathbf{u} := (x_i)_{i \in [\![0,n]\!]}$ de segments contenus dans I.

(ii) La fonction $|f|$ est $\mathscr{HK}$-intégrable sur I si, et seulement si, le membre de droite de (26) est fini.

Démonstration.
Supposons $|f|$ $\mathscr{HK}$-intégrable sur I.

Soit $\mathbf{u} := (x_i)_{i\in[\![0,n]\!]}$ une subdivision d'un segment $[x_0, x_n]$ contenu dans I ; on a, pour tout $i \in [\![0, n-1]\!]$, $\left|\int_{[x_i,x_{i+1}]} f\right| \leqslant \int_{[x_i,x_{i+1}]} |f|$, donc :

$$\sum_{i\in[\![0,n-1]\!]} \left|\int_{[x_i,x_{i+1}]} f\right| \leqslant \int_{[x_0,x_n]} |f| \leqslant \int_I |f|.$$

Par suite $S := \sup_{\mathbf{u}} \sum_{i\in[\![0,n-1]\!]} \left|\int_{[x_i,x_{i+1}]} f\right| \leqslant \int_I |f|$, donc S est fini.

Supposons inversement que le membre de droite de (26) soit fini. Nous allons montrer que $|f|$ est $\mathscr{HK}$-intégrable sur I *et* que l'on a l'égalité (26). Cette preuve, un peu technique, est basée sur le lemme de Henstock (plus précisément sur le corollaire 96 de la page 281). Le lecteur pourra la laisser de côté dans un premier temps.

1. Commençons par le cas où I est un segment : $I = [a, b]$.
   Supposons que le membre de droite de (26) soit fini et posons :
   $$S := \sup_{\mathbf{u}} \sum_{i\in[\![0,n-1]\!]} \left|\int_{[x_i,x_{i+1}]} f\right|.$$
   Soit $\varepsilon > 0$. D'après la définition de S, il existe des points $y_0 < y_1 < \cdots < y_n$ de I tels que :
   $$\sum_{i\in[\![0,n-1]\!]} \left|\int_{[y_i,y_{i+1}]} f\right| \geqslant S - \varepsilon.$$
   On peut évidemment supposer (quitte à compléter la subdivision) que $y_0 = a$ et que $y_n = b$.
   Notons $\mathbf{y} := \{y_0, \ldots, y_n\}$. On va utiliser la jauge $\delta_{\mathbf{y}}$ définie dans le lemme 71 de la page 251.
   Puisque f est $\mathscr{HK}$-intégrable sur I, il existe une jauge δ' qui est ε-adaptée à f sur I. On définit $\delta := \min(\delta', \delta_{\mathbf{y}})$.
   Soit $\mathcal{D} = \{(J_i, \xi_i)_{i\in I}\}$ une subdivision δ-fine de $[a, b]$. D'après le lemme 71, les points $\{y_1, \ldots, y_{n-1}\}$ sont des points marqués de $\mathcal{D}$ et l'on peut les « ajouter » à $\mathcal{D}$ de façon à obtenir une subdivision δ-fine $\mathcal{D}'$ de $[a, b]$ admettant les points $\{y_0, \ldots, y_n\}$ comme points de subdivision *et* comme points marqués et l'on a $S'_{\mathcal{D}}\big(|f|\big) = S_{\mathcal{D}}\big(|f|\big)$.
   Il existe, pour tout $k \in [\![0, n-1]\!]$, une subdivision δ-fine $\mathcal{D}_k$ de $[y_k, y_{k+1}]$, de telle sorte que $\mathcal{D}'$ soit la réunion des $\mathcal{D}_k$. Alors :
   $$S_{\mathcal{D}}\big(|f|\big) = S_{\mathcal{D}'}\big(|f|\big) = \sum_{k=0}^{N+1} S_{\mathcal{D}_k}\big(|f|\big). \tag{27}$$
   Notons $\mathcal{D}' := \{(J'_h, \xi_h)_{h\in H}\}$. D'après le corollaire 96 de la page 281, on a :
   $$-2\varepsilon \leqslant \sum_{h\in H} \left|\int_{J'_H} f\right| - S_{\mathcal{D}'}\big(|f|\big) \leqslant 2\varepsilon. \tag{28}$$
   Chaque J'_h est contenu dans un intervalle $[y_k, y_{k+1}]$ *unique* et chaque intervalle $[y_k, y_{k+1}]$, $k = 0, \ldots, n-1$, est la réunion des J'_h qu'il contient. On a donc, en utilisant la définition de S :
   $$S - \varepsilon \leqslant \sum_{k=0}^{n-1} \left|\int_{[y_i,y_{i+1}]} f\right| \leqslant \sum_{h\in H} \left|\int_{J'_h} f\right| \leqslant S. \tag{29}$$

De (27), (28) et (29) on déduit :

$$S - 3\varepsilon \leqslant S_{\mathcal{D}}\big(|f|\big) \leqslant S + 2\varepsilon,$$

ce qui montre que $|f|$ est intégrable sur I, d'intégrale égale à S.

2. Passons au cas général : $I \subset \mathbb{R}$ est un intervalle arbitraire. Nous allons déduire le résultat de celui que nous venons de prouver pour un segment.
Supposons que le membre de droite de (26) soit fini et posons :

$$S := \sup_{\mathbf{u}} \sum_{i \in [\![0,n-1]\!]} \Big| \int_{[x_i, x_{i+1}]} f \Big|.$$

Soit $[a', b'] \subset I$. On note $S\big([a', b']\big)$ le réel obtenu si, dans le $\sup\limits_{\mathbf{u}}$ ci-dessus, on se limite aux subdivisions $\mathbf{u}$ de $[a', b']$. On a évidemment $S\big([a', b']\big) \leqslant S$. D'après 1., on a $S\big([a', b']\big) = \displaystyle\int_{a'}^{b'} |f| \leqslant S$ et, puisque $|f| \geqslant 0$, on en déduit (*cf.* la proposition 94 de la page 278) que $|f|$ est intégrable sur I et que $\displaystyle\int_I |f| \leqslant S$. Soit $\varepsilon > 0$. Par définition de S, il existe un segment $[a', b'] \subset I$ et une subdivision $\mathbf{u} := \{x_i\}_{i \in 0,n-1}$ de $[a', b']$ tels que :

$$S - \varepsilon \leqslant \sum_{i \in [\![0,n-1]\!]} \left| \int_{[x_i, x_{i+1}]} f \right|.$$

Alors :

$$S - \varepsilon \leqslant \sum_{i \in [\![0,n-1]\!]} \left| \int_{[x_i, x_{i+1}]} f \right| \leqslant \int_{a'}^{b'} |f| \leqslant \int_I |f| \leqslant S.$$

On en déduit $\displaystyle\int_I |f| = S$. ∎

Si I est un *segment*, on peut réinterpréter la proposition ci-dessus en utilisant les notions de *variation d'une fonction* et de *fonction à variation bornée* (*cf.* le module *Intégration* du livre [L2]).
Soit $c \in I$. À $f : I \to \mathbb{R}$ $\mathscr{HK}$-intégrable sur I, on associe $F : \begin{array}{ccc} I & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \displaystyle\int_c^x f. \end{array}$

On a :

$$\sup_{\mathbf{u}} \sum_{i \in [\![0,n-1]\!]} \Big| \int_{[x_i, x_{i+1}]} f \Big| = \sup_{\mathbf{u}} \sum_{i \in [\![0,n-1]\!]} \big| F(x_{i+1}) - F(x_i) \big|.$$

Ce $\sup$ n'est autre que la *variation* de F sur I (qui est indépendante du chois de c). Ainsi la fonction $|f|$ est $\mathscr{HK}$-intégrable sur I si, et seulement si, la fonction F est à variation bornée sur I.

Corollaire 105. Soient $f : I \to \mathbb{R}$ une fonction $\mathscr{HK}$-intégrable sur un *segment* I et $c \in I$. Elle est absolument intégrable sur I si, et seulement si, la fonction $F : x \mapsto \displaystyle\int_c^x f$ est *à variation bornée* sur I.

Proposition 106. Soient f et g des fonctions $\mathscr{HK}$-intégrables sur I.

(i) Si $f \leqslant g$ et si f est *absolument intégrable*, alors g est *absolument intégrable*.

(ii) Si $f \leqslant g$ et si g est *absolument intégrable*, alors f est *absolument intégrable*.

(iii) Si $|f| \leqslant g$, alors f est absolument intégrable sur I. De plus :

$$\left|\int_I f\right| \leqslant \int_I |f| \leqslant \int_I g.$$

En particulier, toute fonction $\mathscr{HK}$-intégrable sur un segment et minorée ou majorée est absolument intégrable sur ce segment.

Démonstration.

(i) On suppose $f \leqslant g$ et f absolument intégrable. Alors $g - f \geqslant 0$ est absolument intégrable, donc :

$$\sup_{\mathbf{u}} \sum_{i\in[\![0,n-1]\!]} \left|\int_{[x_i,x_{i+1}]} g\right| \leqslant \sup_{\mathbf{u}} \sum_{i\in[\![0,n-1]\!]} \left|\int_{[x_i,x_{i+1}]} f\right| + \sup_{\mathbf{u}} \sum_{i\in[\![0,n-1]\!]} \int_{[x_i,x_{i+1}]} (g-f)$$
$$\leqslant \int_I |f| + \int_I (g-f) < +\infty.$$

On en déduit la $\mathscr{HK}$-intégrabilité de $|g|$ en appliquant la proposition 104 de la page 289.

(ii) On applique (i) à $-f$ et $-g$.

(iii) Si $|f| \leqslant g$, alors $g \geqslant 0$ est absolument intégrable et l'on a $-g \leqslant f \leqslant g$. On en déduit l'intégrabilité de $|f|$ en utilisant (i) (ou (ii)). On a vu plus haut que $\left|\int_I f\right| \leqslant \int_I |f|$ (*cf.* la proposition 103), et de $|f| \leqslant g$, l'on déduit $\int_I |f| \leqslant \int_I g$. ∎

Corollaire 107. Si $f : I \to \mathbb{R}$ est $\mathscr{HK}$-localement intégrable sur I et s'il existe une fonction g $\mathscr{HK}$-intégrable sur I telle que $|f| \leqslant g$, alors f est $\mathscr{HK}$-absolument intégrable sur I.

Démonstration. D'après la proposition 106, $|f|$ est $\mathscr{HK}$-localement intégrable. Pour tous c, $c' \in I$ on a $\left|\int_c^{c'} f\right| \leqslant \int_c^{c'} |f| \leqslant \int_c^{c'} g$, donc, d'après le corollaire 91 de la page 277, f est $\mathscr{HK}$-absolument intégrable sur I. ∎

Corollaire 108. Le sous-ensemble $\tilde{L}^1(I) \subset \mathcal{F}(I)$ des fonctions absolument intégrables (c'est-à-dire intégrables au sens de Lebesgue) est un *sous-espace vectoriel* et $f \in \tilde{L}^1(I) \mapsto \|f\|_1 := \int_I |f|$ définit une *semi-norme* sur $\tilde{L}^1(I)$, c'est-à-dire que :

$$\forall \lambda \in \mathbb{R}, \forall f, g \in \tilde{L}^1(I), \quad \|\lambda f\|_1 = |\lambda| \|f\|_1 \quad \text{et} \quad \|f+g\|_1 \leqslant \|f\|_1 + \|g\|_1.$$

Démonstration. Soient $\lambda \in \mathbb{R}$, $f, g \in \tilde{L}^1(I)$. On a $\displaystyle\int_I |\lambda f| = |\lambda| \int_I |f|$.

Par ailleurs $|f| + |g|$ est $\mathscr{HK}$-intégrable sur I et $|f+g| \leqslant |f| + |g|$, donc, d'après la proposition 106, $f+g$ est $\mathscr{HK}$-intégrable sur I et $\displaystyle\int_I |f+g| \leqslant \int_I |f| + \int_I |g|$. ■

La semi-norme $\|.\|_1$ n'est pas une norme. Le sous-espace vectoriel de $\tilde{L}^1(I)$:

$$N := \left\{ f \in \tilde{L}^1(I) \mid \int_I |f| = 0 \right\}$$

contient les fonctions presque partout nulles sur I (*cf.* l'exercice 27 ci-dessus). Nous verrons plus loin qu'en fait N *est* l'ensemble des fonctions presque partout nulles (*cf.* la proposition 136 de la page 314 et le corollaire 139 de la page 316 : c'est un résultat difficile...).

Le résultat suivant est une version du théorème fondamental pour l'intégrale de Lebesgue. Il est dû à Lebesgue (1901) et il a été l'une des motivations de l'introduction par Lebesgue de son intégrale.

Corollaire 109. Soit $f : [a,b] \to \mathbb{R}$ une fonction *dérivable* sur $[a,b]$ telle que f' soit *bornée* sur $[a,b]$. Alors f' est intégrable au sens de Lebesgue sur $[a,b]$ et :

$$\int_a^b f' = f(b) - f(a).$$

Démonstration. D'après le théorème fondamental (théorème 81 de la page 268), f' est $\mathscr{HK}$-intégrable sur $[a,b]$ et $\displaystyle\int_a^b f' = f(b) - f(a)$. Puisque f est *bornée* sur $[a,b]$, f' est intégrable au sens de Lebesgue sur $[a,b]$. ■

On prouve facilement la variante suivante du théorème de Hake pour les fonctions intégrables au sens de Lebesgue à partir de leur définition et du théorème 90 de la page 277.

Théorème 110.
Soient I un intervalle réel d'intérieur non vide, $a := \inf(I) \in \mathbb{R} \cup \{-\infty\}$, $b := \sup(I) \in \mathbb{R} \cup \{+\infty\}$ et $f : I \to \mathbb{R}$. Les deux conditions suivantes sont équivalentes :

(i) la fonction f est intégrable au sens de Lebesgue sur I ;

(ii) la fonction f est intégrable au sens de Lebesgue sur tout segment $[c,d] \subset \operatorname{Int} I$ et $\displaystyle\int_c^d f$ admet une limite finie A quand $c \to a^+$ et $d \to b^-$.

Cette limite est alors égale à $\displaystyle\int_I f = \int_{\operatorname{Int} I} f$.

Nous concluons cette partie par l'observation suivante. Il n'est généralement pas vrai que le minimum ou le maximum de deux fonctions $\mathscr{HK}$-intégrables est $\mathscr{HK}$-intégrable.

Exemple. Soient $f : \mathbb{R} \to \mathbb{R}$ la fonction identiquement nulle et $g : \mathbb{R} \to \mathbb{R}$ la fonction définie par $g : x \mapsto \dfrac{\sin(x)}{x}\cdot$ si $x \neq 0$ et $g(0) = 0$.

Alors, f et g sont $\mathscr{HK}$ -intégrables, mais $g^+ := \sup(f, g)$ n'est pas $\mathscr{HK}$ -intégrable (le lecteur le vérifiera à titre d'exercice en procédant comme dans l'exercice 18 de la page 693 du livre [L2]).

Cependant, on dispose de la proposition suivante.

Proposition 111. Si f et g sont $\mathscr{HK}$ -intégrables sur I et *minorées* (resp. *majorées*) par une fonction $\mathscr{HK}$ -intégrable sur I, alors $\inf(f, g)$ et $\sup(f, g)$ sont $\mathscr{HK}$ -intégrables sur I.

Démonstration. Supposons h $\mathscr{HK}$ -intégrable sur I avec $h \leqslant f$ et $h \leqslant g$. Alors, les fonctions positives $f - h$ et $g - h$ sont $\mathscr{HK}$ -intégrables. De plus, la fonction $f - g$ est $\mathscr{HK}$ -intégrable et :

$$|f - g| \leqslant (f - h) + (g - h).$$

D'après la proposition 106, on voit que $f - g$ est $\mathscr{HK}$ -absolument intégrable.

On termine la démonstration en utilisant les égalités :

$$\inf(f, g) = \frac{1}{2}(f - g) - \frac{1}{2}|f - g| \quad \text{et} \quad \sup(f, g) = \frac{1}{2}(f - g) + \frac{1}{2}|f - g|. \qquad \blacksquare$$

Pour le cas où les fonctions sont majorées, on se ramène au cas où elles sont minorées en passant aux fonctions opposées.

Exercice 29.

Soient $A,\ B \subset I$. On suppose que 1_A et 1_B sont $\mathscr{HK}$ -intégrables. Montrer que $1_{A\cap B}$ et $1_{A\cup B}$ sont $\mathscr{HK}$ -intégrables

Solution. Les fonctions 1_A et 1_B sont minorées par 0. On a $\min(1_A, 1_B) = 1_{A\cap B}$ et $\max(f, g) = 1_{A\cup B}$. D'où le résultat.

2.7 Les théorèmes de convergence

La plupart des fonctions que l'on rencontre dans les problèmes « élémentaires » de calcul différentiel et intégral sont intégrables au sens de Riemann. Cependant, même dans des problèmes simples, il est nécessaire d'étudier des *limites simples* de suites de telles fonctions. Une telle limite peut être « assez compliquée » et n'est pas en général intégrable au sens de Riemann (*cf.* l'exemple de la page 302). De plus, même si elle l'est, on ne peut pas toujours échanger les processus de limite et d'intégration. C'est l'une des raisons qui ont motivé la création de théories de l'intégration plus puissantes que celle de Riemann par Lebesgue et ses successeurs.

On reprend dans cette partie le problème traité à la fin du module *Intégration* du livre [L2]. Il s'agit d'étudier l'intégrabilité de la limite *simple* f d'une suite de fonctions intégrables (f_n) et, sous réserve d'existence, de comparer $\displaystyle\int f$ et $\displaystyle\lim_{n\to+\infty}\int f_n$. Nous allons le faire dans le cadre de la théorie de Henstock-Kurzweil (avec comme « sous-produits » les résultats dans celui de la théorie de Lebesgue).

Après une généralisation du théorème de convergence uniforme que nous avons établi pour l'intégrale de Riemann, nous montrerons, dans le cadre de la théorie de l'intégration de Henstock-Kurzweil, trois théorèmes de convergence plus difficiles (monotone, encadrée, dominée). Ce sont des résultats *très importants* de cette théorie.
Nous déduirons ensuite des théorèmes de convergence des résultats sur l'interversion des sommations et intégrations. Enfin, nous terminerons le module en démontrant les énoncés des théorèmes de convergence que nous avons *provisoirement admis* dans le cadre de l'intégrale de Riemann (*cf.* la partie 8.1.3 *Vers les grands théorèmes de la théorie de l'intégration* du module *Intégration* du livre [L2]).
Nous reviendrons plus loin (une dernière fois...) sur les théorèmes de convergence avec des versions un peu plus générales utilisant les notions *d'ensemble négligeable* et de *propriété vraie presque partout* (*cf.* la partie 3.1.3).
Cette partie contient assez peu d'exemples. Le lecteur pourra reprendre les nombreux exemples et exercices de la partie *Les théorèmes de convergence* du module *Intégration* du livre [L2] en utilisant les nouveaux outils introduits ici, comme nous le faisons ci-dessous avec l'exemple de la page 302.

2.7.1 Les théorèmes de convergence pour l'intégrale de Henstock-Kurzweil et l'intégrale de Lebesgue

Théorème de convergence uniforme

On généralise le théorème 117 du module *Intégration* du livre [L2]. La preuve donne aussi une nouvelle démonstration de ce théorème.

Théorème 112. Si $(f_n : [a,b] \to \mathbb{R})_{n\in\mathbb{N}}$ est une suite de fonctions intégrables au sens de Henstock-Kurzweil (respectivement au sens de Riemann) qui converge uniformément vers $f : [a,b] \to \mathbb{R}$, alors f est intégrable au sens de Henstock-Kurzweil (respectivement au sens de Riemann) et :

$$\lim_{n\to+\infty} \int_a^b f_n = \int_a^b f.$$

Démonstration. Étant donné $\varepsilon > 0$, choisissons n suffisamment grand pour que :

$$\forall x \in [a,b], \quad \big|f_n(x) - f(x)\big| \leqslant \varepsilon.$$

Si $\mathcal{D}$ est une subdivision pointée de $[a,b]$, on a alors :

$$\big|S_{\mathcal{D}}(f_n) - S_{\mathcal{D}}(f)\big| \leqslant (b-a)\varepsilon.$$

Si δ est une jauge sur $[a,b]$ qui est ε-adaptée à f_n et si $\mathcal{D}_1$ et $\mathcal{D}_2$ sont deux subdivisions δ-fines de $[a,b]$. Alors :

$$\begin{aligned}\big|S_{\mathcal{D}_1}(f) - S_{\mathcal{D}_2}(f)\big| &\leqslant \big|S_{\mathcal{D}_1}(f) - S_{\mathcal{D}_1}(f_n)\big| + \big|S_{\mathcal{D}_1}(f_n) - S_{\mathcal{D}_2}(f_n)\big| \\ &\quad + \big|S_{\mathcal{D}_2}(f_n) - S_{\mathcal{D}_2}(f)\big| \\ &\leqslant (b-a)\varepsilon + 2\varepsilon + (b-a)\varepsilon = 2(1+b-a)\varepsilon.\end{aligned}$$

Le critère de Cauchy est donc vérifié. Cela montre que f est intégrable.

De plus si n est suffisamment grand, $f_n - \varepsilon \leqslant f \leqslant f_n + \varepsilon$ et donc :

$$\left| \int_a^b f_n - \int_a^b f \right| \leqslant \int_a^b \varepsilon = (b-a)\varepsilon.$$

Cela montre que $\lim\limits_{n \to +\infty} \displaystyle\int_a^b f_n = \int_a^b f$.

Si f_n est $\mathscr{R}$-intégrable, on peut choisir la jauge δ constante, donc f est aussi $\mathscr{R}$-intégrable. ■

Théorème de convergence monotone

Le théorème de convergence monotone est « le plus important », les théorèmes de convergence dominée et de convergence encadrée s'en déduisent simplement.

Théorème 113 (Convergence monotone). Soient $I \subset \mathbb{R}$ un intervalle et $(f_n : I \to \mathbb{R})_{n \in \mathbb{N}}$ une suite croissante de fonctions $\mathscr{HK}$-intégrables sur I qui converge simplement vers une fonction $f : I \to \mathbb{R}$. Alors f est $\mathscr{HK}$-intégrable sur I si, et seulement si, la suite croissante $\displaystyle\int_I f_n$ est majorée. Dans ce cas :

$$\lim_{n \to +\infty} \int_I f_n = \int_I f.$$

De plus, si f_0 est *intégrable au sens de Lebesgue*, alors f est *intégrable au sens de Lebesgue*.

Remarque. On a bien entendu un résultat entièrement analogue pour les suites décroissantes de fonctions, obtenu en remplaçant (f_n) par $(-f_n)$.

Démonstration. Le point clé de la preuve est l'utilisation du lemme de Henstock.

1. On commence par traiter le cas où $I := [a, b]$ est un segment.

   Si f est $\mathscr{HK}$-intégrable sur $[a, b]$, on a $A_n := \displaystyle\int_a^b f_n \leqslant \int_a^b f$. La suite croissante $(A_n)_{n \in \mathbb{N}}$ est donc majorée.

   Inversement supposons la suite croissante $(A_n)_{n \in \mathbb{N}}$ majorée. Elle admet une limite A. Nous allons voir que f est $\mathscr{HK}$-intégrable et que $\displaystyle\int_I f = A$ en utilisant la définition de la $\mathscr{HK}$-intégrabilité. Soit $\varepsilon > 0$. On choisit $n_0 \in \mathbb{N}$ tel que :

   $$\forall n \geqslant n_0, \quad A - \varepsilon \leqslant A_n = \int_a^b f_n \leqslant A.$$

   Puisque la suite $(f_n)_{n \in \mathbb{N}}$ converge *simplement* vers f, on peut choisir, pour tout $x \in [a, b]$, $N(x) \in \mathbb{N}$ tel que $N(x) \geqslant n_0$ et :

   $$\forall n \geqslant N(x), \quad f(x) - \varepsilon \leqslant f_n(x) \leqslant f(x).$$

   Pour tout $n \in \mathbb{N}$, on choisit une jauge δ_n, $\varepsilon/2^n$-adaptée à f_n, ce qui permet, compte tenu de ce qui précède, de définir une jauge δ sur $[a, b]$ par $\delta : x \mapsto \delta(x) := \delta_{N(x)}(x) > 0$.

Soit $\mathcal{D} := \big(([x_i, x_{i+1}], \xi_i)\big)_{i\in[\![0,m-1]\!]}$ une subdivision δ-fine de $[a,b]$. On a :

$$\begin{aligned} S_{\mathcal{D}}(f) - A = &\sum_{i\in[\![0,m-1]\!]} \big(f(\xi_i) - f_{N(\xi_i)}(\xi_i)\big)(x_{i+1} - x_i) \\ &+ \sum_{i\in[\![0,m-1]\!]} \left(f_{N(\xi_i)}(\xi_i)(x_{i+1} - x_i) - \int_{x_i}^{x_{i+1}} f_{N(\xi_i)}\right) \\ &+ \sum_{i\in[\![0,m-1]\!]} \int_{x_i}^{x_{i+1}} f_{N(\xi_i)} - A. \end{aligned}$$

Notons :

$$\Sigma_1 := \left| \sum_{i\in[\![0,m-1]\!]} \big(f(\xi_i) - f_{N(\xi_i)}(\xi_i)\big)(x_{i+1} - x_i) \right|,$$

$$\Sigma_2 := \left| \sum_{i\in[\![0,m-1]\!]} \left(f_{N(\xi_i)}(\xi_i)(x_{i+1} - x_i) - \int_{x_i}^{x_{i+1}} f_{N(\xi_i)}\right) \right|$$

$$\Sigma_3 := \left| \sum_{i\in[\![0,m-1]\!]} \int_{x_i}^{x_{i+1}} f_{N(\xi_i)} - A \right|.$$

On a $\big|S_{\mathcal{D}}(f) - A\big| \leqslant \Sigma_1 + \Sigma_2 + \Sigma_3$. Nous allons majorer Σ_1, Σ_2 et Σ_3. D'après le choix de $N(x)$, on a $\Sigma_1 \leqslant \varepsilon \sum_{[\![0,m-1]\!]} (x_{i+1} - x_i) = \varepsilon(b-a)$. Par ailleurs, notant $p := \max_{i\in[\![0,m-1]\!]} \big(N(\xi_i)\big)$, on a, d'après la monotonie de la suite $(f_n)_{n\in\mathbb{N}}$:

$$A_{n_0} = \int_a^b f_{n_0} \leqslant \sum_{i\in[\![0,m-1]\!]} \int_{x_i}^{x_{i+1}} f_{N(\xi_i)} \leqslant \sum_{i\in[\![0,m-1]\!]} \int_{x_i}^{x_{i+1}} f_p \leqslant \int_a^b f_p = A_p \leqslant A.$$

On en déduit $\Sigma_3 \leqslant A - A_{n_0} \leqslant \varepsilon$. Il reste à majorer Σ_2. c'est la partie délicate de la preuve.

Pour tout $i \in [\![0, m-1]\!]$, $N(\xi_i) \in [\![n_0, p]\!]$. On obtient donc une *partition* de $[\![0, m-1]\!]$ en regroupant les indices i tels que $N(\xi_i) = n$ pour $n \in [\![n_0, p]\!]$ donné. Par suite :

$$\begin{aligned} &\sum_{i\in[\![0,m-1]\!]} \left(f_{N(\xi_i)}(\xi_i)(x_{i+1} - x_i) - \int_{x_i}^{x_{i+1}} f_{N(\xi_i)}\right) \\ &= \sum_{n\in[\![n_0,p]\!]} \sum_{\{i|N(\xi_i)=n\}} \left(f_n(\xi_i)(x_{i+1} - x_i) - \int_{x_i}^{x_{i+1}} f_n\right) \end{aligned}$$

Notons, pour tout $n \in [\![n_0, p]\!]$:

$$\sigma_n := \left| \sum_{\{i|N(\xi_i)=n\}} \left(f_n(\xi_i)(x_{i+1} - x_i) - \int_{x_i}^{x_{i+1}} f_n\right) \right|.$$

On a $\Sigma_3 \leqslant \sum_{n\in[\![n_0,p]\!]} \sigma_n$. Il reste à majorer les σ_n. On utilise pour cela le lemme de Henstock. Pour tout ξ_i tel que $N(\xi_i) = n$, $\delta(\xi_i) = \delta_{N(\xi)} = \delta_n$, donc la subdivision partielle correspondante $\big(([x_i, x_{i+1}], \xi_i)\big)_{i|N(\xi_i)=n}$ est δ_n fine. Par suite $\sigma_n \leqslant \varepsilon/2^n$ et $\Sigma_3 \leqslant \sum_{n\in[\![n_0,p]\!]} \varepsilon/2^n \leqslant 2\varepsilon$.

Finalement $\big|S_{\mathcal{D}}(f) - A\big| \leqslant \Sigma_1 + \Sigma_2 + \Sigma_3 \leqslant \varepsilon(3 + b - a)$. Par suite f est $\mathscr{HK}$-intégrable sur $[a, b]$.

2. Passons à la preuve dans le cas général d'un intervalle I arbitraire.

   Quitte à remplacer la suite $(f_n)_{n\in\mathbb{N}}$ par la suite $(f_n - f_0)_{n\in\mathbb{N}}$, on peut supposer que toutes les fonctions f_n sont positives. Leur limite simple f est donc positive.

   Soit $A := \sup_{n\in\mathbb{N}} \int_I f_n < +\infty$. Pour tous $[a, b] \subset I$, on a $\int_a^b f_n \leqslant \int_I f_n$, donc $\sup_{n\in\mathbb{N}} \int_a^b f_n \leqslant A < +\infty$, par suite f est $\mathscr{HK}$-intégrable sur $[a, b]$ et $\int_a^b f \leqslant A$. On en déduit, en utilisant le corollaire 94 de la page 278, que f est intégrable sur I et que $\int_I f \leqslant A$.

   Soit $\varepsilon > 0$; il existe $n \in \mathbb{N}$ tel que $A - \varepsilon/2 \leqslant \int_I f_n \leqslant A$. D'après le corollaire 94, il existe $[a, b] \subset I$ tel que $\int_I f_n - \int_a^b f_n \leqslant \varepsilon/2$, donc $A - \varepsilon \leqslant \int_a^b f_n \leqslant A$. On en déduit $A - \varepsilon \leqslant \int_I f_n \leqslant \int_I f \leqslant A$, puis $\int_I f = A$, puisque $\varepsilon > 0$ est arbitraire.

Si f_0 est *intégrable au sens de Lebesgue*, alors f est *intégrable au sens de Lebesgue*. En effet $f = f - f_0 + f_0$, $f - f_0 \geqslant 0$ et la fonction $f - f_0$ est $\mathscr{HK}$-intégrable. Par suite $f - f_0$ est intégrable au sens de Lebesgue et f est intégrable au sens de Lebesgue. ■

Si, dans le théorème ci-dessus, on suppose les f_n intégrables au sens de Riemann (éventuellement généralisé) sur I, la limite f est $\mathscr{HK}$-intégrable mais n'est pas nécessairement intégrable au sens de Riemann (même généralisé). On en donne un exemple ci-dessous.

Exemple. Soit $f : [0, 1] \to \mathbb{R}$ la restriction à $[0, 1]$ de la fonction $1_{\mathbb{Q}}$ (fonction de Dirichlet). Elle est $\mathscr{HK}$-intégrable et son intégrale est nulle, mais elle n'est pas $\mathscr{R}$-intégrable. On choisit une énumération de $\mathbb{Q} \cap [0, 1]$, c'est-à-dire une bijection $r : k \in \mathbb{N} \mapsto r_k \in \mathbb{Q} \cap [0, 1]$. Pour tout $n \in \mathbb{N}$, on note f_n la fonction caractéristique de l'ensemble *fini* $\{r_0, \ldots, r_n\} \subset [0, 1]$. C'est une fonction en escalier, donc $\mathscr{R}$-intégrable et $\int_0^1 f_n = 0$. On vérifie que, quand n tend vers $+\infty$, la suite $(f_n)_{n\in\mathbb{N}}$ converge simplement vers f.

Lemme de Fatou et théorème de convergence encadrée

Le théorème de convergence encadrée est plus fort que le classique théorème de convergence dominée de Lebesgue (*cf.* le théorème 116 ci-dessous) car il permet de traiter le cas de fonctions qui ne sont pas absolument intégrables. Nous le prouverons à l'aide du lemme de Fatou. Avant d'énoncer ce lemme, nous faisons quelques rappels.

Soit $(u_k)_{k\in\mathbb{N}}$ une suite de $\overline{\mathbb{R}} = \mathbb{R}\cup\{\pm\infty\}$. On rappelle les définitions de *limite supérieure* et de *limite inférieure* de (u_k) (voir dans le livre [L1] l'exercice type corrigé du module *Nombres réels, suites numériques* et dans, le livre [L2], la partie *Séries entières* du module *Suites et séries de fonctions*) :

$$\limsup_{k\to+\infty} u_k := \lim_{p\to+\infty} \sup_{k\geqslant p} u_k = \inf_p \sup_{k\geqslant p} u_k \quad \text{et} \quad \liminf_{k\to+\infty} u_k := \lim_{p\to+\infty} \inf_{k\geqslant p} u_k = \sup_p \inf_{k\geqslant p} u_k.$$

On a $\liminf\limits_{k\to+\infty} u_k \leqslant \limsup\limits_{k\to+\infty} u_k$. Une suite (u_k) admet une limite si et seulement si : $\limsup\limits_{k\to+\infty} u_k = \liminf\limits_{k\to+\infty} u_k$. On a alors $\limsup\limits_{k\to+\infty} u_k = \liminf\limits_{k\to+\infty} u_k = \lim\limits_{k\to+\infty} u_k$.

Soient $I \subset \mathbb{R}$ et $(f_k : I \to \overline{\mathbb{R}})_{k\in\mathbb{N}}$ une suite de fonctions. On définit deux fonctions $\limsup f_k,\ \liminf f_k : I \to \overline{\mathbb{R}}$:

$$\limsup f_k : \begin{array}{ccc} I & \longrightarrow & \overline{\mathbb{R}} \\ x & \longmapsto & \limsup\limits_{k\to+\infty} f_k(x) \end{array} \quad \text{et} \quad \liminf f_k : \begin{array}{ccc} I & \longrightarrow & \overline{\mathbb{R}} \\ x & \longmapsto & \liminf\limits_{k\to+\infty} f_k(x). \end{array}$$

Le lemme de Fatou est une conséquence du théorème de convergence monotone. Il permet d'obtenir des résultats pour les suites qui ne sont pas monotones. Il a été prouvé par le mathématicien français Pierre Fatou en 1906.

Proposition 114 (Lemme de Fatou). Soit un intervalle $I \subset \mathbb{R}$.

(i) Soient des fonctions $f_k, h : I \to \mathbb{R}$, $k \in \mathbb{N}$, $\mathscr{HK}$-intégrables sur I, telles que, pour tout $k \in \mathbb{N}$, $h \leqslant f_k$. Si $\liminf\limits_{k\to+\infty} \int_I f_k < +\infty$, la fonction $f := \liminf\limits_{k\to+\infty} f_k$ est à valeurs finies, elle est $\mathscr{HK}$-intégrable et l'on a :

$$\int_I \liminf_{k\to+\infty} f_k \leqslant \liminf_{k\to+\infty} \int_I f_k.$$

(ii) Soient des fonctions $f_k, g : I \to \mathbb{R}$, $k \in \mathbb{N}$, $\mathscr{HK}$-intégrables sur I, telles que, pour tout $k \in \mathbb{N}$, $f_k \leqslant g$. Si $\limsup\limits_{k\to+\infty} \int_I f_k > -\infty$, la fonction $f := \limsup\limits_{k\to+\infty} f_k$ est à valeurs finies, elle est $\mathscr{HK}$-intégrable et l'on a :

$$\int_I \limsup_{k\to+\infty} f_k \geqslant \limsup_{k\to+\infty} \int_I f_k.$$

Démonstration. Il suffit de montrer (i), car (ii) s'en déduit en remplaçant les fonctions par les fonctions opposées. En remplaçant f_k par $f_k - h$, on peut supposer $h = 0$ et $f_k \geqslant 0$.

La limite inférieure $f := \liminf_{k\to+\infty} f_k$ est la limite de la suite *croissante* de fonctions : $(\varphi_p := \inf_{k\geqslant p} f_k)_{p\in\mathbb{N}}$. Pour tout $p \in \mathbb{N}$, la fonction φ_p est $\mathscr{HK}$-intégrable sur I. En effet, elle est la limite simple de la suite *décroissante* de fonctions $\mathscr{HK}$-intégrables : $(\psi_{p,q} := \min_{k\in[\![p,p+q]\!]} f_k)_{q\in\mathbb{N}}$, les $\psi_{p,q}$ sont $\mathscr{HK}$-intégrables (*cf.* la proposition 111 de la page 294) d'intégrale minorée par $\int_I h$, et l'on peut donc appliquer le théorème de convergence monotone.

Pour tout $p \in \mathbb{N}$ et tout $k \geqslant p$ on a $\varphi_p \leqslant f_k$, donc $\int_I \varphi_p \leqslant \inf_{k\geqslant p} \int_I f_k$ et :

$$\int_I \varphi_p \leqslant \inf_{k\geqslant p} \int_I f_k \leqslant \liminf_{k\to+\infty} \int_I f_k < +\infty.$$

On peut donc appliquer à la suite *croissante* $(\varphi_p)_{p\in\mathbb{N}}$ le théorème de convergence monotone. Le résultat s'en déduit. ■

Théorème 115 (Convergence encadrée). Soit $(f_n : I \to \mathbb{R})_{n\in\mathbb{N}}$ une suite de fonctions $\mathscr{HK}$-intégrables sur I qui converge simplement vers une fonction $f : I \to \mathbb{R}$. Supposons qu'il existe deux fonctions $\mathscr{HK}$-intégrables $g : I \to \mathbb{R}$ et $h : I \to \mathbb{R}$ telles que pour tout n, on ait $g \leqslant f_n \leqslant h$. Alors f est $\mathscr{HK}$-intégrable et :

$$\lim_{n\to+\infty} \int_I f_n = \int_I f.$$

Démonstration. La preuve est une application facile du lemme de Fatou. On a :

$$-\infty < \int_I h \leqslant \liminf_{k\to+\infty} \int_I f_k \leqslant \limsup_{k\to+\infty} \int_I f_k \leqslant \int_I g < +\infty$$

D'après la condition d'encadrement et le lemme de Fatou, $\liminf_{k\to+\infty} f_k$ et $\limsup_{k\to+\infty} f_k$ sont finies en tout point, sont $\mathscr{HK}$-intégrables sur I et l'on a :

$$\int_I \liminf_{k\to+\infty} f_k \leqslant \liminf_{k\to+\infty} \int_I f_k \quad \text{et} \quad \int_I \limsup_{k\to+\infty} f_k \geqslant \limsup_{k\to+\infty} \int_I f_k. \tag{30}$$

Par hypothèse on a $\lim_{k\to+\infty} f_k = \liminf_{k\to+\infty} f_k = \limsup_{k\to+\infty} f_k$, donc f est $\mathscr{HK}$-intégrable sur I et l'on a $\int_I f = \int_I \liminf_{k\to+\infty} f_k = \int_I \limsup_{k\to+\infty} f_k$. En utilisant (30), on en déduit :

$$\limsup_{k\to+\infty} \int_I f_k \leqslant \int_I f \leqslant \liminf_{k\to+\infty} \int_I f_k$$

et, puisque $\liminf_{k\to+\infty} \int_I f_k \leqslant \limsup_{k\to+\infty} \int_I f_k$, on a :

$$\liminf_{k\to+\infty} \int_I f_k = \limsup_{k\to+\infty} \int_I f_k = \lim_{k\to+\infty} \int_I f_k = \int_I f.$$ ■

Théorème de convergence dominée

En corollaire du théorème de convergence encadrée, on a le théorème de convergence dominée. Ce théorème concerne seulement les fonctions intégrables au sens de Lebesgue.

Théorème 116 (Convergence dominée). Soit $(f_n : I \to \mathbb{R})_{n\in\mathbb{N}}$ une suite de fonctions intégrables *au sens de Lebesgue sur* I qui converge simplement vers une fonction $f : I \to \mathbb{R}$. Supposons qu'il existe une fonction $g : I \to \mathbb{R}$ intégrable au sens de Lebesgue telle que pour tout n, on ait $|f_n| \leqslant g$. Alors f est intégrable au sens de Lebesgue et :

$$\lim_{n\to+\infty} \int_I f_n = \int_I f.$$

Démonstration. Les suites (f_n) et $(|f_n|)$ sont encadrées par g et $-g$ qui sont toutes deux $\mathscr{HK}$-intégrables sur I, on en déduit que f et $|f|$ sont $\mathscr{HK}$-intégrables et que $\lim\limits_{n\to+\infty} \int_I f_n = \int_I f$, en utilisant le théorème de convergence encadrée. ■

On pourrait reformuler le théorème ci-dessus en supposant seulement que g et les f_n sont $\mathscr{HK}$-intégrables. C'est sans intérêt car les énoncés sont équivalents. En effet, si g et les f_n sont $\mathscr{HK}$-intégrables, g étant positive est intégrable au sens de Lebesgue et f_n également d'après la proposition 106 de la page 292 ($f_n < g$).

La condition de domination est *suffisante* pour assurer l'intégrabilité de f et l'interversion des limites et intégrales, mais elle n'est *pas nécessaire* : *cf.* les exercices III.1.28 et III.1.29.

Rappelons que tout ouvert U de $\mathbb{R}$ s'écrit de façon unique comme réunion finie ou dénombrable d'intervalles ouverts deux à deux disjoints $U = \bigcup\limits_{j\in J} I_j$.

Si J est fini, la longueur de U est $\operatorname{long} U = \sum\limits_{j\in J} \operatorname{long} I_j$, la fonction 1_U est $\mathscr{HK}$-intégrable sur $\mathbb{R}$ et $\int_{\mathbb{R}} 1_U = \operatorname{long} U$.

Si J est infini, on peut supposer que $J = \mathbb{N}$. Si la série à termes positifs $\sum \operatorname{long} I_k$ est convergente, elle est commutativement convergente et l'on peut *définir* la longueur de U par $\operatorname{long} U := \sum\limits_{k\in\mathbb{N}} \operatorname{long} I_k$. Sinon, on dira que U est de longueur infinie. Si U est borné, U est de longueur finie mais la réciproque est fausse.

Proposition 117. Soit $U \subset \mathbb{R}$ un ouvert *de longueur finie*. Alors la fonction caractéristique 1_U est $\mathscr{HK}$-intégrable sur $\mathbb{R}$ et $\int_{\mathbb{R}} 1_U = \operatorname{long} U$.

Démonstration. Si U est réunion *finie* d'intervalles ouverts deux à deux disjoints, le résultat est évident. Sinon, on écrit $U = \bigcup\limits_{n\in\mathbb{N}} I_n$ (réunion disjointe d'intervalles ouverts bornés)

et, pour tout $k \in \mathbb{N}$, on note $f_k := \sum_{n=0}^{k} 1_{I_n}$. 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 $k \in \mathbb{N}$:

$$\int_0^1 f_k = \sum_{n=0}^{k} \operatorname{long} I_n \leqslant \operatorname{long} U < +\infty$$

On peut donc appliquer le théorème de convergence monotone (théorème 113) à la suite $(f_k)_{k\in\mathbb{N}}$.

Ainsi I_U est $\mathscr{HK}$-intégrable et $\displaystyle\int_0^1 1_U = \lim_{k\to+\infty} \int_0^1 f_k = \sum_{n=0}^{+\infty} \operatorname{long} I_n = \operatorname{long} U$. ■

Corollaire 118.
Soit $K \subset \mathbb{R}$ un compact. Sa fonction caractéristique 1_K est $\mathscr{HK}$-intégrable sur $\mathbb{R}$.

Démonstration. Soit I un intervalle *ouvert* contenant K, $U := I \setminus K$ est un ouvert de $\mathbb{R}$. On a $I_K = 1_I - 1_U$ et 1_I et 1_U sont $\mathscr{HK}$-intégrables, donc 1_K l'est aussi. ■

Nous allons revenir sur un exemple étudié dans la partie 8.1.3 du module *Intégration* du livre [L2]. Nous n'avions répondu à la question posée qu'avec une hypothèse restrictive, nous pouvons maintenant traiter le cas général.

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

$$f(t,.) : x \in [0,1] \mapsto f(t,x) := \frac{\sqrt{g^2(x)+t^2} - \left|g(x)\right|}{t} \in \mathbb{R}$$

Pour tout $x \in [0,1]$ fixé, on a, en séparant en deux cas :

(i) si $g(x) \neq 0$, $0 \leqslant f(t,x) \leqslant \frac{t}{2\left|g(x)\right|}$, donc $\lim_{t\to0^+} f(t,x) = 0$;

(ii) si $g(x) = 0$, $f(t,x) = 1$, donc $\lim_{t\to0^+} f(t,x) = 1$.

Nous avons vu que si g ne s'annule pas sur $[0,1]$, la fonction J est dérivable à droite en 0 et sa dérivée est nulle.
Nous allons traiter le cas général, où g peut s'annuler. Il est beaucoup plus difficile. Notons $K \subset [0,1]$ l'ensemble des zéros de la fonction g. C'est un compact. Nous avons traité le cas où 1_K est intégrable au sens de Riemann. Nous pouvons maintenant traiter le cas général.
Soit $(t_n)_{n\in\mathbb{N}}$ une suite de réels strictement positifs tendant vers 0. Pour tout $n \in \mathbb{N}$, on pose $f_n := f(t_n,.)$. La suite $(f_n)_{n\in\mathbb{N}}$ converge *simplement* vers 1_K.

On a $\sqrt{g^2(x)+t^2} \leqslant \left|g(x)\right| + t$ donc $0 \leqslant f_n(x) \leqslant 1$. On peut donc appliquer le théorème de convergence dominée 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 dans tous les cas dérivable à droite en 0 et sa dérivée est $\displaystyle\int_0^1 1_K$.

2.7.2 Interversion des sommations et des intégrales

Des théorèmes de convergence monotone et de convergence dominée, on déduit facilement des énoncés permettant d'échanger série et intégrale.

Proposition 119. Supposons que $\sum\limits_{k\in\mathbb{N}} f_k$ soit une série de fonctions positives et $\mathscr{HK}$ -intégrables sur un intervalle I, simplement convergente de somme $F : I \to \mathbb{R}$. Si $\displaystyle\sum_{k=0}^{+\infty} \left(\int_I f_k\right) < +\infty$, alors F est $\mathscr{HK}$ -intégrable sur I et :

$$\sum_{k=0}^{+\infty} \left(\int_I f_k\right) = \int_I \left(\sum_{k=0}^{+\infty} f_k\right).$$

Démonstration. On applique le théorème de convergence monotone à la suite croissante de fonctions $F_n := \displaystyle\sum_{k=0}^{n} f_k$ qui converge vers $F := \displaystyle\sum_{k=0}^{+\infty} f_k$. ∎

Proposition 120. Supposons que $\sum\limits_{k\in\mathbb{N}} f_k$ soit une série de fonctions absolument intégrables au sens de Henstock-Kurzweil sur un intervalle I, qui converge absolument et dont la somme est notée $F : I \to \mathbb{R}$. Alors F est absolument intégrable sur I et l'on a :

$$\sum_{k=0}^{+\infty} \left(\int_I |f_k|\right) < +\infty \quad \Longrightarrow \quad \sum_{k=0}^{+\infty} \left(\int_I f_k\right) = \int_I \left(\sum_{k=0}^{+\infty} f_k\right). \tag{31}$$

Démonstration. Par hypothèse, la série $\sum |f_k|$ converge simplement. Soit $G : I \to \mathbb{R}$ sa somme. D'après la proposition précédente, si $\displaystyle\sum_{k=0}^{+\infty} \left(\int_I |f_k|\right) < +\infty$, alors G est intégrable sur I.

Pour tout $n \geqslant 0$, posons $F_n := \sum_{k=0}^{n} f_k$. Alors, F_n est $\mathscr{HK}$-intégrable sur I comme somme de fonctions $\mathscr{HK}$-intégrables et pour tout $n \geqslant 0$, $|F_n| \leqslant G$. Par suite, d'après le théorème de convergence dominée, F est absolument intégrable et l'on a l'égalité à droite de $\Rightarrow$ dans (31). ∎

Nous donnerons plus loin une version un peu plus générale de la proposition ci-dessus (*cf.* la proposition 144 de la page 320).

2.7.3 Preuves des théorèmes de convergence énoncés dans le cadre de l'intégrale de Riemann

Cette partie est consacrée aux preuves des énoncés des théorèmes de convergence énoncés et admis dans le cadre de l'intégrale de Riemann dans le module *Intégration* du livre [L2]. Son but est uniquement « d'honorer un contrat ».

On déduit facilement du théorème 113 ci-dessus la version pour l'intégrale de Riemann énoncée et admise dans le livre [L2] du théorème de convergence monotone.

Théorème 121 (Convergence monotone pour l'intégrabilité au sens de Riemann). Soient I un intervalle réel et $(f_n : I \to \mathbb{R})_{n\in\mathbb{N}}$ une suite croissante de fonctions intégrables au sens de Riemann (éventuellement généralisé) sur I qui converge simplement vers une fonction $f : I \to \mathbb{R}$ localement intégrable au sens de Riemann sur I. Alors f est intégrable au sens de Riemann (éventuellement généralisé) sur I si, et seulement si, la suite croissante $\int_I f_n$ est majorée. Dans ce cas :

$$\lim_{n\to+\infty} \int_I f_n = \int_I f.$$

Démonstration. Montrons que le théorème 113 entraîne le théorème 121.

Soit $(f_n : I \to \mathbb{R})_{n\in\mathbb{N}}$ une suite croissante de fonctions intégrables *au sens de Riemann* (éventuellement généralisé) sur I qui converge simplement vers une fonction $f : I \to \mathbb{R}$. Les fonctions f_n sont $\mathscr{HK}$-intégrables sur I. On peut donc leur appliquer le théorème 113. On en déduit que f est $\mathscr{HK}$-intégrable et que $\lim_{n\to+\infty} \int_I f_n = \int_I f$. Mais, par hypothèse, on sait que f est localement $\mathscr{R}$-intégrable, elle donc intégrable au sens de Riemann généralisé d'après le corollaire 92 de la page 277. Son intégrale au sens de Riemann et son intégrale au sens de Henstock-Kurzweil coïncident, d'où le résultat. ∎

On déduit facilement du théorème 115 ci-dessus la version pour l'intégrale de Riemann énoncée et admise dans le livre [L2] du théorème de convergence encadrée.

Théorème 122 (Convergence encadrée pour l'intégrabilité au sens de Riemann). Soit $(f_n : I \to \mathbb{R})_{n\in\mathbb{N}}$ une suite de fonctions intégrables au sens de Riemann (éventuellement généralisé) sur I qui converge simplement vers une fonction $f : I \to \mathbb{R}$ localement intégrable au sens de Riemann.
Supposons qu'il existe deux fonctions intégrables au sens de Riemann (éventuellement généralisé) $g : I \to \mathbb{R}$ et $h : I \to \mathbb{R}$ telles que pour tout n, on ait $g \leqslant f_n \leqslant h$. Alors f est intégrable au sens de Riemann (éventuellement généralisé) sur I et :

$$\lim_{n\to+\infty} \int_I f_n = \int_I f.$$

Démonstration. Montrons que le théorème 115 entraîne le théorème 122.
D'après le théorème 115, la fonction f est $\mathscr{HK}$-intégrable sur I. Par hypothèse elle est localement intégrable au sens de Riemann sur I, d'après le corollaire 92 de la page 277, elle est donc intégrable au sens de Riemann (éventuellement généralisé) sur I. ∎

Pour une version du théorème de convergence dominée dans le cadre de l'intégrale de Riemann on se reportera au livre [L2]. Ce résultat est un corollaire du théorème ci-dessus.

2.8 Fonction définie par une intégrale : continuité et dérivabilité

Dans cette partie, $X \subset \mathbb{R}$ est un intervalle 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 :

$$F(x) := \int_I f(x,t)\,\mathrm{d}t \quad \text{avec} \quad f : X \times I \to \mathbb{R} \quad \text{une fonction de deux variables.}$$

Dans le module *Intégration* du livre [L2], nous avons traité cette question dans le cadre de l'intégrale de Riemann, nous allons le faire dans celui de l'intégrale de Henstock-Kurzweil. Les résultats s'étendent facilement au cas des fonctions dans un $\mathbb{K}$-espace vectoriel normé de dimension finie (pour les définitions, *cf.* la section 2.9 ci-dessous).

Théorème 123 (Continuité).
Soit $f : X \times I \to \mathbb{R}$ une fonction de deux variables vérifiant

1. pour tout $x \in X$, la fonction $t \mapsto f(x,t)$ est $\mathscr{HK}$-intégrable sur I ;
2. pour tout $t \in I$, la fonction $x \mapsto f(x,t)$ est continue en $x_0 \in X$;
3. il existe deux fonctions $g : I \to \mathbb{R}$ et $h : I \to \mathbb{R}$ $\mathscr{HK}$-intégrables 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 $F(x) := \displaystyle\int_I f(x,t)\,\mathrm{d}t$ est continue en x_0.

Démonstration. Montrons que pour toute suite $(y_n \in X)$ 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 $f_{y_n} \to f_{x_0}$ simplement. 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 $F(y_n) = \displaystyle\int_I f_{y_n} \underset{n\to+\infty}{\longrightarrow} \int_I f_{x_0} = F(x_0)$. ■

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.

Venons-en maintenant à la dérivabilité.

Théorème 124 (Dérivabilité).
Soit $f : X \times I \to \mathbb{R}$ une fonction de deux variables telle que

1. pour tout $x \in X$, la fonction $t \mapsto f(x,t)$ est $\mathscr{HK}$-intégrable 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 $h : I \to \mathbb{R}$ $\mathscr{HK}$-intégrables sur I telles que pour tout $x \in X$ et tout $t \in I$:

$$g(t) \leqslant \frac{\partial f}{\partial x}(x,t) \leqslant h(t).$$

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, $t \mapsto \dfrac{\partial f}{\partial x}(x,t)$ est $\mathscr{HK}$-intégrable sur I et, pour tout $x \in X$:

$$F'(x) = \int_I \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t.$$

Démonstration. Soit $(y_n \in X)_{n\in\mathbb{N}}$ une suite qui converge vers $x_0 \in X$, telle que $y_n \neq x_0$, pour tout $n \in \mathbb{N}$. La condition 1 garantit que F est définie sur X : pour tout $x \in X$, la fonction $f_x := f(x, \cdot)$ est intégrable sur I et $F(x) := \displaystyle\int_I f_x$.

Pour tout $n \in \mathbb{N}$, la fonction $t \in I \mapsto \frac{f(y_n,t)-f(x_0,t)}{y_n-x_0}$ est $\mathscr{HK}$-intégrable sur I et l'on a :

$$\frac{F(y_n) - F(x)}{y_n - x} = \int_I \frac{f(y_n,t) - f(x_0,t)}{y_n - x_0}\,\mathrm{d}t,$$

ce qui suggère l'étude de la suite de fonctions $(\varphi_n)_{n\in\mathbb{N}}$ sur I définie par :

$$\varphi_n : t \mapsto \frac{f(y_n,t) - f(x_0,t)}{y_n - x_0}.$$

D'après la condition 2 :

$$\lim_{n\to+\infty} \frac{f_{y_n} - f_{x_0}}{y_n - x_0} = \frac{\partial f}{\partial x}(x_0, \cdot).$$

Autrement dit, la suite $(\varphi_n)_{n\in\mathbb{N}}$ converge simplement sur I vers $t \in I \mapsto \frac{\partial f}{\partial x}(x_0, t)$.
Par ailleurs, les φ_n, $n \in \mathbb{N}$, sont $\mathscr{HK}$-intégrables sur I et l'on a :

$$\frac{f(y_n, t) - f(x_0, t)}{y_n - x_0} = \frac{\partial f}{\partial x}\big(c_n(t), t\big) \quad \text{pour un certain point } c_n(t) \text{ compris entre } x_0 \text{ et } y_n.$$

La condition 3 implique alors l'encadrement $g \leqslant \varphi_n \leqslant h$. Le résultat annoncé découle du théorème de convergence encadrée appliqué à la suite $(\varphi_n)_{n\in\mathbb{N}}$. ■

Remarque. La dérivabilité étant une propriété locale, il suffit d'avoir des encadrements valables au voisinage de chaque point de X. Bien souvent, on majore le module.

2.9 Intégrale de Henstock-Kurzweil des fonctions à valeurs dans un espace vectoriel de dimension finie

Dans cette partie, $\mathbb{K}$ désigne $\mathbb{R}$ ou $\mathbb{C}$. Nous étudions l'intégrale de Henstock-Kurzweil 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. La démarche est similaire à celle adoptée pour l'intégrale de Riemann dans la partie du module *Intégration* du livre [L2].
Les preuves étant similaires, nous ne les reprendrons pas, laissant leur adaptation au lecteur.

2.9.1 Définitions

Définition 25. Soit $f : [a, b] \to \mathbb{C}$. On dit que f est *intégrable au sens de Henstock-Kurzweil* (ou *$\mathscr{HK}$-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 : $\displaystyle\int_a^b f := \int_a^b \operatorname{Re} f + \Big(\int_a^b \operatorname{Im} f\Big)\, i$.

Une fonction à valeurs complexes $\mathscr{HK}$-intégrable est $\mathscr{R}$-intégrable et les intégrales coïncident.

Proposition et définition 125. 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 Henstock-Kurzweil (ou $\mathscr{HK}$-intégrable) sur $[a, b]$ si ses fonctions coordonnées $f_1, \ldots, f_m$ sur une base $\mathcal{B} := (e_1, \ldots, e_m)$ de E le sont. Cette propriété est indépendante de la base ainsi que le vecteur $\Big(\displaystyle\int_a^b f_1\Big)e_1 + \ldots + \Big(\int_a^b f_m\Big)e_m \in E$. Ce vecteur est appelé intégrale de f sur $[a, b]$ et noté $\displaystyle\int_a^b f$.

2.9.2 Propriétés

Les preuves des résultats suivants sont très simples et laissées au lecteur.

Proposition 126. Soit E un $\mathbb{K}$-espace vectoriel de dimension finie, l'ensemble $\mathscr{HK}\big([a,b],E\big)$ des fonctions intégrables au sens de Henstock-Kurzweil sur $[a,b]$ est un sous-espace vectoriel de $\mathcal{F}\big([a,b],E\big)$. L'application $f \in \mathscr{HK}\big([a,b],E\big) \mapsto \displaystyle\int_a^b f \in E$ est $\mathbb{K}$-linéaire.

Proposition 127. 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 $\mathscr{HK}$-intégrable, alors $u \circ f$ est une application $\mathscr{HK}$-intégrable de $[a,b]$ dans F et l'on a $\displaystyle\int_a^b u \circ f = u\Big(\int_a^b f\Big)$.

Proposition 128. 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 $\mathscr{HK}$-intégrables sur $[a,b]$, alors $\Phi \circ (f,g) : t \in [a,b] \mapsto \Phi\big(f(t),g(t)\big) \in G$ est une application $\mathscr{HK}$-intégrable sur $[a,b]$.

Corollaire 129. 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 $\mathscr{HK}$-intégrables sur $[a,b]$. Alors $fg : [a,b] \to E$ est $\mathscr{HK}$-intégrable sur $[a,b]$.

Relation de Chasles

On étend facilement aux applications à valeurs dans un espace vectoriel de dimension finie E la proposition 76 de la page 264 en utilisant des coordonnées sur E.

Proposition 130 (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 $\mathscr{HK}$-intégrable sur $[a,b]$ il faut et il suffit que ses restrictions à $[a,c]$ et $[c,b]$ soient $\mathscr{HK}$-intégrables respectivement sur $[a,c]$ et $[c,b]$. Alors, en notant f ces restrictions on a :

$$\int_a^b f = \int_a^c f + \int_c^b f. \tag{32}$$

Si $b \leqslant a$, on généralise la définition de l'intégrale $\displaystyle\int_a^b f$ d'une application f à valeur dans E en utilisant des convention similaires à celles adoptées pour les fonctions (*cf.* page 633 du livre [L2]). Comme dans le cas des fonctions on peut alors étendre la formule de Chasles.

Proposition 131. Soient $a, b, c \in \mathbb{R}$ et f une fonction à valeurs dans E, $\mathscr{HK}$-intégrable sur le segment $[\min(a,b,c), \max(a,b,c)]$. Alors f est $\mathscr{HK}$-intégrable sur chacun des segments déterminés par deux quelconques des points a, b, c, et l'on a :

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

2.9.3 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 $\mathscr{HK}$-intégrabilité donnée plus haut (*cf.* la définition 125) 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 132. 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 Henstock-Kurzweil sur $[a,b]$.

(ii) Il existe $I \in E$ tel que, pour tout $\varepsilon > 0$, il existe une jauge δ telle que pour toute subdivision pointée $\mathcal{D}$, δ-fine, de $[a,b]$, 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 $\big\|(v_1, \ldots, v_n)\big\| := \sup\limits_{i \in [\![1,n]\!]} \{|v_i|\}$.

Montrons (i) $\Rightarrow$ (ii).

Supposons f $\mathscr{HK}$-intégrable. On va montrer (ii) avec $I := \displaystyle\int_a^b f$.

On a, en notant $f = (f_1, \ldots, f_n)$:

$$\Big\|S_{\mathcal{D}}(f) - \int_a^b f\Big\| = \sup_{i \in [\![1,n]\!]} \Big\|S_{\mathcal{D}}(f_i) - \int_a^b f_i\Big\|.$$

Puisque les fonctions f_i sont $\mathscr{HK}$-intégrables, pour tout $\varepsilon > 0$ et tout $i \in [\![1,n]\!]$, il existe une jauge δ_i telle que pour toute subdivision pointée $\mathcal{D}_i$, δ-fine, de $[a,b]$ l'on ait $\Big\|S_{\mathcal{D}}(f_i) - \displaystyle\int_a^b f_i\Big\| \leqslant \varepsilon$. La fonction $\delta := \inf\limits_{i \in [\![1,n]\!]} \delta_i$ est une jauge sur $[a,b]$ et, pour toute subdivision pointée $\mathcal{D}$, δ-fine, de $[a,b]$ on a :

$$\Big\|S_{\mathcal{D}}(f) - \int_a^b f\Big\| = \sup_{i \in [\![1,n]\!]} \Big\|S_{\mathcal{D}}(f_i) - \int_a^b f_i\Big\| \leqslant \varepsilon.$$

Montrons (ii) $\Rightarrow$ (i).

On note $I = (I_1, \ldots, I_n)$. Soit $\varepsilon > 0$, il existe une jauge δ, telle que pour toute subdivision pointée $\mathcal{D}$, δ-fine, de $[a,b]$ l'on ait $\left\|S_{\mathcal{D}}(f) - I\right\| \leqslant \varepsilon$. Alors, pour tout $i \in [\![1,n]\!]$, on a $|S_{\mathcal{D}}(f_i) - I_i| \leqslant \varepsilon$. Par suite, les fonctions f_i sont $\mathscr{HK}$-intégrables et f est $\mathscr{HK}$-intégrable. ■

Soient $(E, \|.\|)$ un $\mathbb{K}$-espace vectoriel *normé de dimension finie* et $f : [a,b] \to E$. Pour toute subdivision pointée $\mathcal{D}$ de $[a,b]$, on a $\left\|S_{\mathcal{D}}(f)\right\| \leqslant \left\|S_{\mathcal{D}}\big(\|f\|\big)\right\|$. On en déduit le résultat suivant.

Proposition 133. Soient $(E, \|.\|)$ un espace normé de dimension finie sur $\mathbb{K}$ et $f : [a,b] \to E$ une application $\mathscr{HK}$-intégrable sur $[a,b]$. *Si* la fonction $\|f\| : [a,b] \to \mathbb{R}$ est $\mathscr{HK}$-intégrable sur $[a,b]$, on a :

$$\left\|\int_a^b f\right\| \leqslant \int_a^b \|f\| \leqslant (b-a) \sup_{x\in[a,b]} \left\|f(x)\right\|. \tag{33}$$

Contrairement à ce qui se passe avec l'intégrale de Riemann, $f : [a,b] \to E$ peut être $\mathscr{HK}$-intégrable sans que $\|f\|$ le soit. Nous l'avons vu dans le cas où $n = 1$ (*cf.* l'exemple de la page 287).

3 Intégrale de Henstock-Kurzweil et mesure de Lebesgue

Dans cette section, nous *déduisons* la théorie de la *mesure de Lebesgue* sur $\mathbb{R}$ de la théorie de l'intégration de Henstock-Kurzweil. Plus précisément la théorie de Henstock-Kurzweil nous permet de construire (facilement...) une mesure, qui coïncide *a posteriori* avec la mesure « classique » (pour ceux qui la connaissent...) d'après un résultat d'unicité, et d'établir ses propriétés.

La théorie de la mesure permet, en suivant les idées originales de Lebesgue, de construire la théorie « classique » de l'intégrale de Lebesgue. Nous présentons brièvement cette théorie et nous montrons que la notion d'intégrabilité ainsi définie coïncide avec celle de Henstock-Kurzweil restreinte aux fonctions *absolument intégrables* (que nous avons étudiées dans la partie 2.6), ce qui justifie *a posteriori* notre terminologie (*cf.* la définition 24 de la page 288). Nous nous limitons au cas de la *dimension un*. Il y a des résultats similaires en dimension quelconque, les idées sont voisines mais les preuves plus techniques et nous ne traiterons pas cette question. Le lecteur intéressé pourra consulter [De].

Cette section peut être utilisée comme introduction aux théories plus générales de la mesure et de l'intégrale abstraites que nous ne développons pas dans ce livre.

La théorie de la mesure de Lebesgue donne aussi un éclairage intéressant sur la notion d'intégrabilité au sens de Henstock-Kurzweil (*cf.* par exemple ci-dessous la proposition 164).

Cette partie relève est plus difficile que ce qui précède.

Dans ce module, nous définissons, dans un premier temps, l'intégrale de Lebesgue, les parties mesurables de la droite réelle (tribu de Lebesgue) et la mesure de Lebesgue sur ces parties *à partir de l'intégrale de Henstock-Kurzweil*. Il existe une toute autre présentation, très répandue, que nous appellerons « classique », dérivant des idées de Borel et Lebesgue et de leurs successeurs (*cf.* par exemple [Ru]). Nous prouverons à la fin de notre exposition que les deux présentations sont *équivalentes*. Ainsi, tout est pour le mieux dans le meilleur des mondes. Toutefois, le lecteur *prendra garde* au cours de la lecture et tant que l'équivalence n'est pas établie *à ne pas faire de confusion* entre nos définitions et preuves et celles qu'il pourra trouver dans d'autres livres ou documents.

3.1 Mesure de Lebesgue — Ensembles négligeables

La théorie de Henstock-Kurzweil va nous permettre de « mesurer » *certains* sous-ensembles de $\mathbb{R}$. On reprend pour cela la démarche de la partie *Sous-ensembles élémentaires* du module *Intégration* du livre [L2]. Le fait d'utiliser la théorie de Henstock-Kurzweil nous permettra de surmonter les difficultés rencontrées avec celle de Riemann et d'aller beaucoup plus loin.

3.1.1 Mesure de Lebesgue

Dans le module *Intégration* du livre [L2] nous avons abordé le problème de la mesure des parties de $\mathbb{R}$ (c'est-à-dire la généralisation de la notion de longueur) sous deux angles.

1. Nous avons défini une notion de longueur pour les ouverts de $\mathbb{R}$ en utilisant le fait qu'un tel ouvert U s'écrit de façon unique comme réunion au plus dénombrable d'intervalles ouverts disjoints : $U = \bigcup_{j\in J} I_j$. Si J est fini ou si les I_i sont bornés *et* la série à termes positifs $\sum \operatorname{long} I_j$ convergente, on définit $\operatorname{long} U := \sum \operatorname{long} I_j$. Sinon on convient de $\operatorname{long} U := +\infty$ (l'idée de cette définition est due à Émile Borel).
2. Nous avons par ailleurs défini une notion de longueur pour les parties $A \subset \mathbb{R}$ bornées telles que la fonction caractéristique 1_A soit $\mathscr{R}$-intégrable par $\operatorname{long} A := \displaystyle\int_{\mathbb{R}} 1_A$.

Nous avons vu que les approches 1 et 2 ne coïncidaient pas toujours : un ouvert borné U de $\mathbb{R}$ est de longueur finie mais 1_U n'est pas nécessairement $\mathscr{R}$-intégrable.
L'intégrale de Henstock-Kurzweil va nous permettre de réconcilier les approches 1 et 2 (*cf.* ci-dessous le corollaire 135) et de définir une large classe de parties dites mesurables de $\mathbb{R}$ et leurs mesures (en acceptant la valeur $+\infty$). Nous verrons que la théorie de la mesure ainsi obtenue est *équivalente* à celle élaborée de façon très différente par Borel et Lebesgue (qui prolonge l'approche 1 ci-dessus).

Définition 26. Si A est une partie de $\mathbb{R}$, on dit qu'elle est *intégrable* si sa fonction caractéristique sur $\mathbb{R}$ est $\mathscr{HK}$-intégrable sur $\mathbb{R}$ et, si c'est le cas, on *définit* sa *mesure de Lebesgue* par $\mu(A) = \displaystyle\int_{\mathbb{R}} 1_A$.

Pour simplifier, on notera $\displaystyle\int_{\mathbb{R}} 1_A = \int 1_A$.

La notion de partie intégrable et la mesure de Lebesgue sont *invariantes par translation*. Si $A \subset \mathbb{R}$ est intégrable, pour tout $x \in \mathbb{R}$, $x + A$ est intégrable et $\mu(x + A) = \mu(A)$ (*cf.* la proposition 77 de la page 265).

Remarque. Les fonctions 1_A étant *positives*, elles sont $\mathscr{HK}$ -intégrables si, et seulement si, elles sont intégrables au sens de Lebesgue (au sens de la définition 24 de la page 288).

Proposition 134. Les parties intégrables de $\mathbb{R}$ et la mesure de Lebesgue μ possèdent les propriétés suivantes.

(i) Si $A, B \subset \mathbb{R}$ sont des parties intégrables, les parties $A \cap B$, $A \cup B$ et $A \setminus B$ sont intégrables.

(ii) Si $A, B \subset \mathbb{R}$ sont des parties intégrables et si $A \subset B$, $\mu(A) \leqslant \mu(B)$.

(iii) Si $(A_k)_{k\in\mathbb{N}}$ est une *suite croissante* de parties intégrables, $\bigcup\limits_{k\in\mathbb{N}} A_k$ est intégrable si, et seulement si, la suite numérique $(\mu(A_k))_{k\in\mathbb{N}}$ est *bornée*. S'il en est ainsi : $\mu\big(\bigcup\limits_{k\in\mathbb{N}} A_k\big) = \lim\limits_{k\to+\infty} \mu(A_k)$

(iv) Si $(A_k)_{k\in\mathbb{N}}$ est une *suite décroissante* de parties intégrables, $\bigcap\limits_{k\in\mathbb{N}} A_k$ est intégrable et $\mu\left(\bigcap\limits_{k\in\mathbb{N}} A_k\right) = \lim\limits_{k\to+\infty} \mu(A_k)$

(v) Si $(A_k)_{k\in\mathbb{N}}$ est une famille de parties intégrables deux à deux disjointes, $\bigcup\limits_{k\in\mathbb{N}} A_k$ est intégrable si, et seulement si, la série $\sum \mu(A_k)$ converge. S'il en est ainsi : $\mu\left(\bigcup\limits_{k\in\mathbb{N}} A_k\right) = \sum\limits_{k\in\mathbb{N}} \mu(A_k)$.

La propriété (v) est appelée σ*-additivité*.

Démonstration.

(i) On a :

$$1_{A\cap B} = \min(1_A, 1_B), \quad 1_{A\cup B} = \max(1_A, 1_B), \quad 1_{A\setminus B} = 1_A - 1_{A\cap B}.$$

Le résultat s'en déduit (*cf.* l'exercice 29 de la page 294).

(ii) Si $A \subset B$, $1_A \leqslant 1_B$, donc $\mu(A) = \displaystyle\int 1_A \leqslant \int 1_B = \mu(B)$.

(iii) L'assertion résulte du *théorème de convergence monotone* appliqué à la suite *croissante* $(1_{A_k})_{k\in\mathbb{N}}$.

(iv) L'assertion résulte du théorème de convergence monotone appliqué à la suite *décroissante* $(1_{A_k})_{k\in\mathbb{N}}$.

(v) Notons, pour tout $n \in \mathbb{N}$, $B_k := A_0 \cup \ldots \cup A_k$. D'après (i), B_k est une partie intégrable. On a $1_{B_k} = 1_{A_0} + \cdots + 1_{A_k}$, d'où $\mu(B_k) = \mu(A_0) + \cdots + \mu(A_k)$. La série $\sum \mu(A_k)$ est bornée si, et seulement si, la suite $(\mu(B_k)$ converge, on déduit alors le résultat (v) de (iii). ∎

De l'assertion (v) de la proposition précédente, on déduit le résultat suivant.

Corollaire 135. Soit $U \subset \mathbb{R}$ un ouvert et $U = \bigcup_{j \in J} I_j$, les I_j étant des intervalles ouverts *bornés* disjoints et J étant au plus dénombrable. Les deux conditions suivantes sont équivalentes.

(i) La « série » $\sum_{j \in J} \operatorname{long} I_j$ est convergente.

(ii) L'ouvert U est intégrable.

S'il en est ainsi, on a $\mu(U) = \operatorname{long} U = \sum_{j \in J} \operatorname{long} I_j$.

En particulier tout ouvert borné est intégrable.

3.1.2 Ensembles négligeables

Nous donnons ci-dessous la définition « naturelle » des ensembles (respectivement fonctions) négligeables pour la théorie de Henstock-Kurzweil. Ce sont les ensembles (respectivement fonctions) « invisibles » pour cette théorie. Il y a ici *une difficulté technique* car les ensembles $\mathscr{HK}$-négligeables ainsi définis sont *a priori* différents des ensembles négligeables définis dans le module *Intégration* du livre [L2] (*cf.* également plus haut, page 187). Nous verrons ci-dessous qu'en fait les deux définitions coïncident (*cf.* le corollaire 139 de la page 316). Ce n'est pas du tout évident, la preuve utilise le lemme de Vitali.

Définition 27. Soit $I \subset \mathbb{R}$ un intervalle. On dit :

- qu'une fonction $f : I \to \mathbb{R}$ est $\mathscr{HK}$*-négligeable* si $|f|$ est $\mathscr{HK}$-intégrable sur I et si $\displaystyle\int_I |f| = 0$;
- qu'une partie $A \subset \mathbb{R}$ est $\mathscr{HK}$*-négligeable* si sa fonction caractéristique 1_A est négligeable, c'est-à-dire si 1_A est $\mathscr{HK}$-intégrable et $\mu(A) = 0$.

Une fonction caractéristique est toujours positive, donc une partie $A \subset \mathbb{R}$ est $\mathscr{HK}$-*négligeable* si et seulement si 1_A est intégrable au sens de Lebesgue et si son intégrale est nulle. Certains auteurs parlent donc de parties négligeables au sens de Lebesgue ou de parties négligeables pour la mesure de Lebesgue (puisque $\mu(A) = 0$) au lieu de $\mathscr{HK}$-négligeables.

La proposition suivante donne quelques propriétés des fonctions $\mathscr{HK}$-négligeables et des parties $\mathscr{HK}$-négligeables.

Proposition 136. Soit $I \subset \mathbb{R}$ un intervalle d'intérieur non vide.

(i) Soient $f, g : I \to \mathbb{R}$ telles que $|g| \leqslant f$. Si f est négligeable, g est négligeable.

(ii) Une fonction $f : I \to \mathbb{R}$ est négligeable si, et seulement si, f^+ et f^- le sont.

(iii) Si $f : I \to \mathbb{R}$ est négligeable, f est $\mathscr{HK}$-intégrable et $\int_I f = 0$.

(iv) Tout sous-ensemble d'un ensemble $\mathscr{HK}$-négligeable est $\mathscr{HK}$-négligeable.

(v) Une réunion au plus dénombrable de parties $\mathscr{HK}$-négligeables est $\mathscr{HK}$-négligeable.

(vi) Une fonction $f : I \to \mathbb{R}$ est négligeable si, et seulement si, $\{x \in I \mid f(x) \neq 0\}$ est $\mathscr{HK}$-négligeable.

(vii) Soit $m \in \mathbb{N}^*$. Une fonction $f : I \to \mathbb{R}$ est négligeable si, et seulement si, la fonction f^m est négligeable.

(viii) Soient $f : I \to \mathbb{R}$ une fonction négligeable et $g : I \to \mathbb{R}$ une fonction. Alors fg est négligeable.

Démonstration.

(i) Soient $f, g : I \to \mathbb{R}$ telles que $|g| \leqslant f$. On reprend la méthode utilisée dans l'exercice 24 de la page 261.

Soient $\varepsilon > 0$ et $\big(\delta, [\alpha, \beta]\big)$ un couple ε-adapté à f sur I. Si $\mathcal{D}$ est une subdivision δ-fine d'un segment $[a, b] \subset I$ contenant $[\alpha, \beta]$, on a $0 \leqslant S_{\mathcal{D}}(f) = \big|S_{\mathcal{D}}(f) - \displaystyle\int_I f\big| \leqslant \varepsilon$.

Puisque $0 \leqslant |g| \leqslant f$, on a :

$$0 \leqslant \big|S_{\mathcal{D}}(|g|) - 0\big| \leqslant S_{\mathcal{D}}(|g|) \leqslant S_{\mathcal{D}}(f) \leqslant \varepsilon,$$

donc $|g|$ est $\mathscr{HK}$-intégrable sur I et $\displaystyle\int_I |g| = 0$.

(ii) On a $0 \leqslant f^+ \leqslant |f|$ et $0 \leqslant f^- \leqslant |f|$. Ainsi, d'après (i), si f est négligeable f^+ et f^- le sont. Inversement on utilise $|f| = f^+ + f^-$.

(iii) On suppose f négligeable. On a $f = f^+ - f^-$, donc, d'après (ii) :

$$\int_I f = \int_I f^+ - \int_I f^- = 0.$$

(iv) Si $B \subset A \subset \mathbb{R}$, $0 \leqslant 1_B \leqslant 1_A$. D'après (i), si 1_A est négligeable, 1_B l'est aussi.

(v) Soit $A := A_1 \cup \ldots \cup A_m$, $1_A \leqslant 1_{A_1} + \cdots + 1_{A_m}$, donc, si $A_1, \ldots, A_m$ sont $\mathscr{HK}$-négligeables, A l'est aussi.

Soit $A := \displaystyle\bigcup_{k \in \mathbb{N}} A_k$. On note, pour tout $n \in \mathbb{N}$, $B_n := \displaystyle\bigcup_{k \in [\![0,n]\!]} A_k$.

On a $A := \bigcup_{n\in\mathbb{N}} B_n$ et les B_n sont $\mathscr{HK}$-négligeables. On peut appliquer le théorème de convergence monotone à la suite de fonctions $(1_{B_n})_{n\in\mathbb{N}}$ qui converge simplement vers 1_A sur $\mathbb{R}$. On en déduit que 1_A est $\mathscr{HK}$-intégrable sur $\mathbb{R}$ et que $\mu(A) = \int_{\mathbb{R}} 1_A = \lim_{n\to+\infty} \int_{\mathbb{R}} 1_{B_n} = 0$.

(vi) Posons $A := \{x \in I \mid f(x) \neq 0\}$. Pour tout $k \in \mathbb{N}$, on note $g_k := \min(1_A, k|f|)$ et $h_k := \min(k1_A, |f|)$. On vérifie que la suite de fonctions positives $(g_k)_{k\in\mathbb{N}}$ est croissante et converge simplement vers 1_A sur $\mathbb{R}$ tandis que la suite de fonctions positives $(h_k)_{k\in\mathbb{N}}$ est croissante et converge simplement vers $|f|$ sur $\mathbb{R}$

Si f est négligeable, les g_k sont négligeables d'après (i) puisque $0 \leqslant g_k \leqslant k|f|$. On peut appliquer le théorème de convergence monotone à la suite $(g_k)_{k\in\mathbb{N}}$. On en déduit que 1_A est $\mathscr{HK}$-intégrable sur $\mathbb{R}$ et que $\mu(A) = \int_{\mathbb{R}} 1_A = \lim_{k\to+\infty} \int_{\mathbb{R}} g_k = 0$.

Inversement si A est $\mathscr{HK}$-négligeable, les fonctions h_k sont négligeables d'après (i) puisque $0 \leqslant g_k \leqslant k1_A$. On peut appliquer le théorème de convergence monotone à la suite $(h_k)_{k\in\mathbb{N}}$.

On en déduit que $|f|$ est $\mathscr{HK}$-intégrable sur $\mathbb{R}$ et que $\int_{\mathbb{R}} |f| = \lim_{k\to+\infty} \int_{\mathbb{R}} h_k = 0$.

(vii) On a $A := \{x \in I \mid f(x) \neq 0\} = \{x \in I \mid f^m(x) \neq 0\}$, d'où le résultat en utilisant (vi).

(viii) On a $\{x \in I \mid g(x)f(x) \neq 0\} \subset \{x \in I \mid f(x) \neq 0\} = A$, d'où le résultat en utilisant (iv) et (vi). ■

Le lemme de Vitali permet d'interpréter la mesure des ensembles intégrables.

Proposition 137. Soit $A \subset \mathbb{R}$ une partie intégrable. Alors pour tout $\varepsilon > 0$, il existe un *ouvert intégrable* U contenant A et une famille dénombrable d'intervalles ouverts $(U_k)_{k\in\mathbb{N}}$ tels que :

$$U = \bigcup_{k\in\mathbb{N}} U_k \quad \text{et} \quad \mu(A) \leqslant \mu(U) \leqslant \sum_{k\in\mathbb{N}} \operatorname{long}(U_k) \leqslant \mu(A) + \varepsilon.$$

Démonstration. On traite d'abord le cas où A est *borné*. On suppose $A \subset I$, I étant un segment de $\mathbb{R}$. Soient $\varepsilon > 0$ et δ une jauge ε-adaptée à 1_A fixés. On considère la famille de segments pointés $\mathcal{V} := \big((J,x) \mid J \subset I,\ x \in A \text{ et } \operatorname{long} J \leqslant \delta(x)\big)$. C'est (avec l'abus de langage évident) un recouvrement de Vitali de A, donc, d'après le lemme de Vitali, il existe une famille au plus dénombrable $((J_k, x_k))_{k\in K}$ de segments pointés deux à deux disjoints « de » $\mathcal{V}$ et un sous-ensemble négligeable $N \subset I$ tels que $A \subset \bigcup_{k\in K} J_k \cup N$ et, pour tout $k \in \mathbb{N}$, $\operatorname{long} J_k \leqslant \delta(x_k)$.

Pour tout sous-ensemble *fini* $K' \subset K$, $((J_k, x_k))_{k\in K'}$ est une subdivision partielle δ-fine.

On peut donc lui appliquer le lemme de Henstock, d'où :

$$\sum_{k\in K'} \operatorname{long} J_k = \sum_{k\in K'} 1_A(x_k) \operatorname{long} J_k \leqslant \int_I 1_A + \varepsilon/4 = \mu(A) + \varepsilon/4.$$

On en déduit $\sum\limits_{k\in K} \operatorname{long} J_k \leqslant \mu(A) + \varepsilon/4$.

Supposons K infini (le cas fini, plus simple, est laissé au lecteur). On peut alors supposer que $K = \mathbb{N}$. Pour tout $k \in \mathbb{N}$, il existe un intervalle ouvert V_k contenant J_k et tel que $\operatorname{long} V_k \leqslant \operatorname{long} J_k + 2^{-k-3}\varepsilon$. Par ailleurs N peut être recouvert par une famille d'intervalles ouverts $\{W_h\}_{h\in\mathbb{N}}$ telle que $\sum\limits_{h\in\mathbb{N}} \operatorname{long} W_h \leqslant \varepsilon/2$. On a $A \subset U := \bigcup\limits_{k\in\mathbb{N}} V_k \cup \bigcup\limits_{h\in\mathbb{N}} W_h$, avec :

$$\begin{aligned}
\mu(U) &\leqslant \sum_{k\in\mathbb{N}} \operatorname{long} V_k + \sum_{h\in\mathbb{N}} \operatorname{long} W_h \\
&\leqslant \sum_{k\in\mathbb{N}} \operatorname{long} J_k + \sum_{k\in\mathbb{N}} 2^{-k-3}\varepsilon + \sum_{h\in\mathbb{N}} \operatorname{long} W_h \\
&\leqslant \mu(A) + \varepsilon/4 + \varepsilon/4 + \varepsilon/2 = \mu(A) + \varepsilon.
\end{aligned}$$

On a donc prouvé le résultat quand A est borné.

Supposons maintenant que A ne soit pas borné. On écrit $\mathbb{R} = \bigcup\limits_{m\in\mathbb{Z}} \big(m + [0,1[\big)$ et, pour tout $m \in \mathbb{Z}$, on note $A_m := A \cap (m + [0,1[)$. Les A_m sont deux à deux disjoints donc, par σ-additivité (*cf.* la proposition 134 (v)) : $\mu(A) = \sum\limits_{m\in\mathbb{Z}} \mu(A_m)$.

En utilisant ce qui précède, on obtient, pour tout $m \in \mathbb{Z}$, un *ouvert intégrable* U_m contenant A_m et une famille dénombrable d'intervalles ouverts $\{U_{m,k}\}_{k\in\mathbb{N}}$ tels que :

$$U_m = \bigcup_{k\in\mathbb{N}} U_{m,k} \quad \text{et} \quad \mu(A_m) \leqslant \mu(U_m) \leqslant \sum_{k\in\mathbb{N}} \operatorname{long} U_{m,k} \leqslant \mu(A_m) + 2^{-|m|-2}\varepsilon.$$

Notons $U := \bigcup\limits_{m\in\mathbb{Z}} U_m = \bigcup\limits_{(m,k)\in\mathbb{Z}\times\mathbb{N}} U_{m,k}$. On a :

$$\mu(A) \leqslant \mu(U) \leqslant \sum_{(m,k)\in\mathbb{Z}\times\mathbb{N}} \operatorname{long} U_{m,k} \leqslant \sum_{m\in\mathbb{Z}} \mu(A_m) + \sum_{m\in\mathbb{Z}} 2^{-|m|-2}\varepsilon \leqslant \mu(A) + \varepsilon. \quad \blacksquare$$

Corollaire 138. Soit $A \subset \mathbb{R}$ une partie intégrable. Alors $\mu(A) = \inf\limits_{A\subset U} \operatorname{long}(U)$, la borne inférieure portant sur les *ouverts intégrables* contenant A.

Nous reviendrons plus loin sur le corollaire ci-dessus avec la notion de *mesure extérieure* d'une partie (*cf.* la partie 3.2.3).

Corollaire 139. Soit $A \subset \mathbb{R}$. Les conditions suivantes sont équivalentes :

(i) A est $\mathscr{HK}$-négligeable

(ii) A est négligeable.

Démonstration. Montrons (i) $\Rightarrow$ (ii). Supposons que A soit $\mathscr{HK}$ -négligeable.

Soit $\varepsilon > 0$. D'après la proposition 137 il existe un ouvert U contenant A et une famille dénombrable d'intervalles ouverts $\{U_k\}_{k\in\mathbb{N}}$ tels que :

$$A \subset U = \bigcup_{k\in\mathbb{N}} U_k \quad \text{et} \quad \mu(A) \leqslant \sum_{k\in\mathbb{N}} \operatorname{long}(U_k) \leqslant \varepsilon,$$

donc A est négligeable.

Montrons (ii) $\Rightarrow$ (i). On obtient le résultat en appliquant la proposition 93 de la page 278. En voici une autre preuve.

Supposons A négligeable. Soit $m \in \mathbb{N}^*$; d'après la proposition 137 il existe un ouvert intégrable U_m contenant A et une famille dénombrable d'intervalles ouverts $\{U_{m,k}\}_{k\in\mathbb{N}}$ tels que :

$$A \subset U_m = \bigcup_{k\in\mathbb{N}} U_{m,k} \quad \text{et} \quad \sum_{k\in\mathbb{N}} \operatorname{long}(U_{m,k}) \leqslant 1/m,$$

Les U_m, $m \in \mathbb{N}^*$ sont intégrables, donc $N := \bigcap_{m\in\mathbb{N}^*} U_m$ est intégrable.

On a, pour tout $m \in \mathbb{N}^*$, $\mu(N) \leqslant \mu(U_m) \leqslant 1/m$, donc $\mu(N) = 0$ et N est $\mathscr{HK}$ -négligeable. Puisque $A \subset N$, on en déduit que A est $\mathscr{HK}$ -négligeable. ∎

Compte tenu du corollaire ci-dessus nous abandonnerons, à partir de maintenant, la terminologie « $\mathscr{HK}$ -négligeable » et ne parlerons dans la suite que d'ensembles négligeables.

Définition 28. Une *fonction étagée* est une combinaison linéaire finie, à coefficients réels, de fonctions caractéristiques de sous-ensembles *intégrables*.

La notion de fonction étagée, introduite par Lebesgue, généralise celle de fonction en escalier (une fonction étagée ne prend qu'un *nombe fini* de valeurs). Toute fonction étagée est intégrable au sens de Lebesgue. Si $f := \lambda_1 1_{A_1} + \cdots + \lambda_m 1_{A_m}$, où les A_i sont cubables, on a $\int f = \lambda_1\mu(A_1) + \cdot + \lambda_m\mu(A_m)$. Dans la théorie classique on *définit* l'intégrale de Lebesgue d'une fonction en escalier par cette expression.

3.1.3 Retour sur les théorèmes de convergence

La notion d'ensemble négligeable (réinterprétée ci-dessus) permet de donner facilement des versions plus fortes des théorèmes de convergence, très utiles pour les applications.

Théorème 140 (Convergence monotone, version forte). Soient I un intervalle réel et $(f_k : I \to \mathbb{R})_{k\in\mathbb{N}}$ une suite de fonctions $\mathscr{HK}$-intégrables croissante presque partout sur I. Alors :

(i) S'il existe une fonction $f : I \to \mathbb{R}$ $\mathscr{HK}$-intégrable sur I telle que la suite (f_k) converge presque partout vers f sur I :

$$\lim_{k\to+\infty} \int_I f_k = \int_I f < +\infty.$$

(ii) Inversement si $\lim\limits_{k\to+\infty} \int_I f_k < +\infty$, l'ensemble A des réels x tels que $\lim\limits_{k\to+\infty} f_k(x) = +\infty$ est négligeable. De plus la fonction f définie par

$$f : x \mapsto \begin{cases} \lim\limits_{k\to+\infty} f_k(x) & \text{si} \quad x \notin A \\ 0 & \text{si} \quad x \in A \end{cases}$$

est $\mathscr{HK}$-intégrable et satisfait (i).

Démonstration. Les résultats se déduisent simplement du théorème de convergence monotone établi précédemment (*cf.* le théorème 113 de la page 296).

(i) On désigne par B l'ensemble des points de I tels que :
- ou bien la suite $\big(f_k(x)\big)$ n'est pas croissante ;
- ou bien la suite $\big(f_k(x)\big)$ est croissante mais non bornée ;
- ou bien la suite $\big(f_k(x)\big)$ est croissante et bornée mais sa limite est différente de $f(x)$. Cet ensemble est la réunion de trois ensembles négligeables, il est donc négligeable. On modifie f et les f_k sur B en remplaçant leurs valeurs par 0. On obtient $f = \lim\limits_{k\to+\infty} f_k$ et la suite (f_k) est croissante. On obtient alors le résultat en appliquant le théorème 113.

(ii) Quitte à modifier les f_k en remplaçant leur valeur par 0 sur un ensemble de mesure nulle (ce qui ne change pas $\int_I f_k$), on peut supposer que la suite (f_k) est croissante. En remplaçant $f_k - f_0$, on peut supposer que $f_0 = 0$ et que $f_k \geqslant 0$, pour tout $k \in \mathbb{N}$. Si $\lim\limits_{k\to+\infty} \int_I f_k < +\infty$, il existe $M > 0$ tel que $0 \leqslant \int_I f_k \leqslant M$, pour tout $k \in \mathbb{N}$.
Montrons que A est négligeable.
Soient $n \in \mathbb{N}$ et $m \in \mathbb{N}^*$, on note $A_{n,m} := \{x \in I \mid f_n(x) > m\}$. On vérifie $1_{A_{n,m}} = \lim\limits_{p\to+\infty} \min(1, p(f_n - m)^+)$, la suite du second membre de l'égalité étant *croissante*. On a, par définition : $1_{A_{n,m}} \leqslant \frac{1}{m} f_n$, donc, d'après le théorème de convergence monotone « usuel » (théorème 113) $1_{A_{n,m}}$ est $\mathscr{HK}$-intégrable sur I et :

$$\mu(A_{n,m}) = \int_I 1_{A_{n,m}} \leqslant \frac{1}{m} \int_I f_n \leqslant \frac{M}{m}.$$

Notons, pour tout $m \in \mathbb{N}^*$, $A_m := \{x \in I \mid \lim_{n\to+\infty} f_n(x) > m\}$, $(A_{n,m})_{n\in\mathbb{N}}$ est une suite *croissante* de parties intégrables dont la réunion est A_m. Puisque, pour tout, $n \in \mathbb{N}$, $\mu(A_{n,m}) \leqslant M/m$, A_m est intégrable et $\mu(A_m) \leqslant M/m$. On a $A = \bigcap_{n\in\mathbb{N}^*} A_m$, donc A est négligeable.

On peut modifier les f_n en supposant qu'elle s'annulent sur A et l'on est ramené au théorème de convergence monotone « usuel ». On en déduit que f est $\mathscr{HK}$ -intégrable et satisfait (i). ∎

Convention. Dans la suite on dira qu'une fonction $f : I \to \overline{\mathbb{R}}$ définie sur un intervalle $I \subset \mathbb{R}$, prenant éventuellement les valeurs $\pm\infty$ est $\mathscr{HK}$ -intégrable sur I s'il existe une fonction $g : I \to \mathbb{R}$ égale presque partout à f et $\mathscr{HK}$ -intégrable sur I. On *définit* alors $\int_I f := \int_I g$, cette intégrale étant indépendante du choix de g.

Proposition 141 (Lemme de Fatou, version forte). Soit I un intervalle de $\mathbb{R}$.

(i) Soient des fonctions $f_k, h : I \to \mathbb{R}$, $k \in \mathbb{N}$, $\mathscr{HK}$ -intégrables sur I, telles que, pour tout $k \in \mathbb{N}$, $h \leqslant f_k$ presque partout. Si $\liminf_{k\to+\infty} \int_I f_k < +\infty$, la fonction $f := \liminf_{k\to+\infty} f_k$ est finie presque partout, $\mathscr{HK}$ -intégrable et l'on a :

$$\int_I \liminf_{k\to+\infty} f_k \leqslant \liminf_{k\to+\infty} \int_I f_k.$$

(ii) Soient des fonctions $f_k, g : I \to \mathbb{R}$, $k \in \mathbb{N}$, $\mathscr{HK}$ -intégrables sur I, telles que, pour tout $k \in \mathbb{N}$, $f_k \leqslant g$ presque partout. Si $\limsup_{k\to+\infty} \int_I f_k > -\infty$, la fonction $f := \limsup_{k\to+\infty} f_k$ est finie presque partout, $\mathscr{HK}$ -intégrable et l'on a :

$$\int_I \limsup_{k\to+\infty} f_k \geqslant \limsup_{k\to+\infty} \int_I f_k.$$

Démonstration. La preuve est une variante de celle de celle du lemme 114 de la page 299. Nous laissons les détails au lecteur. ∎

Théorème 142 (Convergence encadrée, version forte). Soient I un intervalle réel et $(f_k : I \to \mathbb{R})_{k\in\mathbb{N}}$ une suite de fonctions $\mathscr{HK}$ -intégrables. On suppose qu'il existe des fonctions $g, h : I \to \mathbb{R}$ $\mathscr{HK}$ -intégrables sur I telles que $h \leqslant f_k \leqslant g$ presque partout. On suppose qu'il existe une fonction $f : I \to \mathbb{R}$ telle que la suite (f_k) converge presque partout vers f. Alors f est $\mathscr{HK}$ -intégrable sur I et :

$$\lim_{n\to+\infty} \int_I f_n = \int_I f.$$

Démonstration. La preuve est une variante de celle du théorème 115 de la page 300. Nous laissons les détails au lecteur. ∎

Théorème 143 (Convergence dominée, version forte). Soient I un intervalle réel et $(f_k : I \to \mathbb{R})_{k\in\mathbb{N}}$ une suite de fonctions $\mathscr{HK}$-intégrables. On suppose qu'il existe une fonction *positive* $g : I \to \mathbb{R}_+$, $\mathscr{HK}$-intégrable sur I, telle que, pour tout $k \in \mathbb{N}$, $|f_k| \leqslant g$ presque partout. On suppose qu'il existe une fonction $f : I \to \mathbb{R}$ telle que la suite (f_k) converge presque partout vers f. Alors les fonctions f_k et f sont absolument intégrables sur I et :

$$\lim_{k\to+\infty} \int_I f_k = \int_I f.$$

Démonstration. Par hypothèse il existe, pour tout $k \in \mathbb{N}$, un ensemble de mesure nulle $A_k \subset I$ tel que, $f_k \leqslant g$ sur $I \setminus A_k$. On définit $g_k : I \to \mathbb{R}$ par :

$$g_k : x \to \begin{cases} g(x) & \text{si} \quad x \notin A_k \\ |f_k|(x) & \text{si} \quad x \in A_k. \end{cases}$$

La fonction g_k est $\mathscr{HK}$-intégrable sur I et l'on a $f_k \leqslant g_k$, donc, d'après la proposition 106 de la page 292, $|f_k|$ est $\mathscr{HK}$-intégrable et f_k est absolument intégrable. On a, pour tout $k \in \mathbb{N}$, $-g \leqslant f_k \leqslant g$ et $0 \leqslant |f_k| \leqslant g$ presque partout, on peut donc appliquer aux suites (f_k) et $(|f_k|)$ le théorème de convergence encadrée (théorème 142). Ainsi f et $|f|$ sont $\mathscr{HK}$-intégrables (donc f est absolument intégrable) et l'on a :

$$\lim_{k\to+\infty} \int_I f_k = \int_I f.$$

■

Le résultat suivant améliore la proposition 120 de la page 303 et son homologue du livre [L2] (proposition 125 de la page 713 dans le cas des fonctions scalaires).

Proposition 144. Soit $\sum f_k$ une série de fonctions *absolument intégrables* au sens de Henstock-Kurzweil sur un intervalle I. On suppose que $\displaystyle\sum_{k=0}^{+\infty} \left(\int_I |f_k| \right) < +\infty$. Alors la série $\sum f_k$ converge *absolument* presque partout sur I vers une fonction F qui est *absolument intégrable* au sens de Henstock-Kurzweil sur I et l'on a :

$$\sum_{k=0}^{+\infty} \left(\int_I f_k \right) = \int_I \left(\sum_{k=0}^{+\infty} f_k \right).$$

Démonstration. Notons, pour tout $n \in \mathbb{N}$, $F_n := \sum_{k=0}^{n} f_k$ et $G_n := \sum_{k=0}^{n} |f_k|$. Par hypothèse les fonctions f_k et $|f_k|$ sont $\mathscr{HK}$-intégrables, donc les fonctions F_n et G_n sont absolument intégrables. La suite (G_n) est croissante et $\displaystyle\lim_{n\to+\infty} \int_I G_n = \sum_{k=0}^{+\infty} \left(\int_I |f_k| \right) < +\infty$, on peut donc appliquer le théorème de convergence monotone (théorème 140). On en déduit que l'ensemble A des réels x tels que $\displaystyle\lim_{n\to+\infty} G_n(x) = +\infty$ est négligeable. Ainsi la

série $\sum f_k$ converge *absolument* sur $I \setminus A$. De plus, si l'on définit $G : I \to \mathbb{R}$ par :

$$G : x \longmapsto \begin{cases} \lim\limits_{n\to+\infty} G_n(x) & \text{si} \quad x \notin A \\ 0 & \text{si} \quad x \in A \end{cases},$$

on a $\displaystyle\int_I G = \lim_{n\to+\infty} \int_I G_n = \sum_{k=0}^{+\infty} \left(\int_I |f_k| \right) < +\infty$ et $|F_n| \leqslant G_n$ presque partout. En appliquant à la suite (F_n) le théorème de convergence dominée (théorème 143), on en déduit que la série $\sum f_k$ converge *absolument* presque partout sur I vers une fonction F qui est absolument intégrable et que $\displaystyle\sum_{k=0}^{+\infty} \left(\int_I f_k \right) = \int_I \left(\sum_{k=0}^{+\infty} f_k \right)$. ■

Remarque. On peut remplacer dans les deux énoncés ci-dessus *absolument intégrable au sens de Henstock-Kurzweil* par *intégrable au sens de Lebesgue* qui est équivalent.

3.2 Ensembles et fonctions mesurables

3.2.1 Ensembles mesurables

Définition 29.
Une partie $A \subset \mathbb{R}$ est dite *mesurable* si, pour tout *segment* I de $\mathbb{R}$, $A \cap I$ est intégrable.

On prouve facilement les résultats suivants.

Théorème 145.

(i) Toute réunion *au plus dénombrable* de parties mesurables est mesurable.

(ii) Toute intersection *au plus dénombrable* de parties mesurables est mesurable.

(iii) Si $A \subset \mathbb{R}$ est mesurable, son complémentaire $\mathbb{R} \setminus A$ est mesurable.

(iv) Les ouverts et les fermés de $\mathbb{R}$ sont mesurables.

(v) Les parties intégrables, en particulier les parties négligeables, sont mesurables.

On étend la mesure de Lebesgue aux parties mesurables en posant :

$$\mu(A) := \lim_{k\to+\infty} \mu\big(A \cap [-k, k]\big) \in \overline{\mathbb{R}}_+ = [0, +\infty]\,.$$

En utilisant la proposition 134 de la page 312, on vérifie que les parties intégrables sont les parties mesurables de mesure finie (on aurait pu choisir dans l'extension ci-dessus n'importe quelle suite exhaustive de segments de $\mathbb{R}$ sans changer le résultat). Les parties négligeables sont les parties mesurables de mesure nulle.
La notion de partie intégrable et la mesure de Lebesgue sont *invariantes par translation*.
Si $(A_k)_{k\in\mathbb{N}}$ est une suite de parties mesurables deux à deux disjointes, alors on a la relation $\mu\big(\bigcup\limits_{k\in\mathbb{N}} A_k\big) = \sum\limits_{k\in\mathbb{N}} \mu(A_k)$, avec la convention $\sum\limits_{k\in\mathbb{N}} \mu(A_k) := +\infty$ si l'un des $\mu(A_k)$ est égal à $+\infty$ ou si la série est divergente. On dit alors que la mesure de Lebesgue est σ*-additive*.

Les notions de famille des parties mesurables et de mesure de Lebesgue admettent des généralisations *abstraites* avec celles de *tribu* et de *mesure*. Ces concepts sont très importants en analyse et en probabilités (*cf.* page 667) mais leur étude dépasse le cadre de nos ouvrages, nous nous limiterons donc à quelques définitions. Ces définitions sont tout à fait naturelles compte tenu des résultats ci-dessus.

Définition 30. On appelle *tribu* (ou σ-algèbre) $\mathcal{T}$ sur un ensemble Ω une famille de parties de Ω telle que :

(i) $\varnothing \in \mathcal{T}$;

(ii) $\mathcal{T}$ est stable par passage au complémentaire : si $A \in \mathcal{T}$, alors $\Omega \setminus A \in \mathcal{T}$;

(iii) $\mathcal{T}$ est stable par réunions *dénombrables* : si, pour tout $k \in \mathbb{N}$, $A_k \in \mathcal{T}$, alors $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{T}$.

Exemples.

1. L'ensemble des parties de Ω est une tribu.
2. D'après le théorème 145, la famille des parties mesurables de $\mathbb{R}$ est une tribu sur $\mathbb{R}$. On l'appelle *tribu de Lebesgue*.

Proposition et définition 146. Soient Ω un ensemble et $\mathcal{C}$ une famille de parties de Ω. Il existe une plus petite tribu sur Ω contenant $\mathcal{C}$. On l'appelle tribu engendrée par $\mathcal{C}$. C'est l'intersection de toutes les tribus contenant $\mathcal{C}$.

Démonstration. La famille des tribus contenant $\mathcal{C}$ n'est pas vide (elle contient l'ensemble des parties). L'intersection de toutes les tribus contenant $\mathcal{C}$ est évidemment une tribu, elle répond à la question. ■

Définition 31. La tribu sur $\mathbb{R}$ engendrée par la famille des *ouverts* est appelée *tribu de Borel* de $\mathbb{R}$, ses éléments sont appelés *sous-ensembles boréliens* de $\mathbb{R}$.

La tribu de Borel est contenue dans la tribu de Lebesgue. On peut montrer que l'inclusion est *stricte*. C'est l'objet de l'exercice III.1.38, qui permet aussi de comprendre à quel point la tribu de Borel est « compliquée ».

La terminologie « tribu *engendrée* par une famille de parties » peut entraîner des confusions. Par exemple, le lecteur optimiste pourrait croire que l'on peut construire par récurrence la tribu de Borel à partir de la famille des ouverts et des fermés en itérant le processus qui permet de passer d'une famille de parties $\mathcal{F}$ à celle des parties formées d'unions dénombrables et d'intersections dénombrables de parties de $\mathcal{F}$. Ce n'est pas le cas (*cf.* l'exercice III.1.38).

La notion de tribu permet de définir celle de mesure (« abstraite »).

Définition 32. Soit $\mathcal{T}$ une tribu sur un ensemble Ω. Une *mesure* est une application $\mu : \mathcal{T} \to \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$ telle que

(i) $\mu(\varnothing) = 0$;

(ii) si $(A_k)_{k\in\mathbb{N}}$ est une suite d'éléments deux à deux disjoints de $\mathcal{T}$, alors :

$$\mu\Big(\bigcup_{k\in\mathbb{N}} A_k\Big) = \sum_{k\in\mathbb{N}} \mu(A_k).$$

La propriété (ii) est appelée *σ-additivité*.

La mesure de Lebesgue est une mesure sur la tribu de Lebesgue. Elle induit une mesure sur la tribu de Borel. Les ensembles de mesure nulle de la tribu de Lebesgue sont les ensembles négligeables, par suite, tout sous-ensemble d'un ensemble de mesure nulle est mesurable et de mesure nulle. On dit que la mesure de Lebesgue sur la tribu de Lebesgue est complète. La mesure induite sur la tribu de Borel n'est pas complète.

3.2.2 Caractérisation des ensembles mesurables

Nous allons voir que, quitte à « cacher sous le tapis » les ensembles négligeables (ce qui est sans importance du point de vue de la théorie de l'intégration), on peut, au lieu de travailler avec les ensembles mesurables, travailler avec les réunions dénombrables de fermés (appelées $\mathcal{F}_\sigma$) qui sont « beaucoup plus simples ».

Définition 33. Un $\mathcal{F}_\sigma$ de $\mathbb{R}$ est une *réunion au plus dénombrable* de fermés de $\mathbb{R}$. Un $\mathcal{G}_\delta$ de $\mathbb{R}$ est une *intersection au plus dénombrable* d'ouverts de $\mathbb{R}$.

On vérifie que tout $\mathcal{F}_\sigma$ de $\mathbb{R}$ est une réunion au plus dénombrable *croissante* de fermés de $\mathbb{R}$ et que tout $\mathcal{G}_\delta$ de $\mathbb{R}$ est une intersection au plus dénombrable *décroissante* d'ouverts de $\mathbb{R}$. Autrement dit, tout $\mathcal{F}_\sigma$ (respectivement $\mathcal{G}_\delta$) s'écrit $F = \bigcap_{n\in\mathbb{N}} F_n$ (respectivement $U = \bigcup_{n\in\mathbb{N}} U_n$), où $(F_n)_{n\in\mathbb{N}}$ (respectivement $(U_n)_{n\in\mathbb{N}}$) est une suite *croissante* (respectivement *décroissante*) de fermés (respectivement d'ouverts). On a alors :

$$\mu(F) = \lim_{n\to+\infty} \mu(F_n) \qquad \text{et} \qquad \mu(U) = \lim_{n\to+\infty} \mu(U_n). \tag{34}$$

Le résultat suivant donne une description des *ensembles mesurables* de $\mathbb{R}$ à partir des *ensembles boréliens* (et plus précisément des $\mathcal{F}_\sigma$ ou des $\mathcal{G}_\delta$) et des *ensembles négligeables*.

Proposition 147. Soit $A \subset \mathbb{R}$. Les conditions suivantes sont équivalentes :

(i) A est mesurable ;

(ii) pour tout $\varepsilon > 0$, il existe un fermé F et un ouvert U tels que $F \subset A \subset U$ et que $\mu(U \setminus F) \leqslant \varepsilon$;

(iii) il existe une *réunion au plus dénombrable* de fermés F' et une *intersection au plus dénombrable* d'ouverts U' telles que $F' \subset A \subset U'$ et $\mu(U' \setminus F') = 0$;

(iv) A peut s'écrire $A = F' \cup N$, où F' est une *réunion au plus dénombrable* de fermés (c'est-à-dire un $\mathcal{F}_\sigma$) et N est une partie *négligeable* ;

(v) A peut s'écrire $A = B \cup N$, où B est une partie *borélienne* et N est une partie *négligeable*.

Intuitivement, si A est une partie mesurable, on peut « glisser des fermés » dans A et « l'envelopper » avec des ouverts. Cela permet « d'approcher » A du point de vue de la mesure, puisque les fermés et les ouverts sont mesurables.

Démonstration.

- Montrons (i) $\Rightarrow$ (ii). On suppose d'abord que A est *borné* : $A \subset I$, où I est un segment. Les parties A et $A' := I \setminus A$ sont intégrables, donc, d'après la proposition 137 de la page 315, pour tout $\varepsilon > 0$, il existe des ouverts U, U' tels que $A \subset U$, $A' \subset U'$, $\mu(U \setminus A) \leqslant \varepsilon/2$ et $\mu(U' \setminus A') \leqslant \varepsilon/2$. En notant $F := I - U'$, on a $F \subset I \setminus A' = A \subset U$ et :
$$\mu(U \setminus F) = \mu(U \setminus A) + \mu(A \setminus F) \leqslant \mu(U \setminus A) + \mu(U' \setminus A') \leqslant \varepsilon.$$
Supposons que A ne soit pas borné. On écrit $\mathbb{R} = \bigcup_{k\in\mathbb{Z}} \big(k+[0,1]\big)$ et, pour tout $k \in \mathbb{Z}$, on note $A_k := A \cap \big(k + [0,1]\big)$.
Soit $\varepsilon > 0$. D'après ce qui précède, pour tout $k \in \mathbb{Z}$, il existe un fermé $F_k \subset k+[0,1]$ et un ouvert $U_k \subset \mathbb{R}$ tels que $F_k \subset A_k \subset U_k$ et $\mu(U_k \setminus F_k) \leqslant 2^{-|k|-2}\varepsilon$.
On vérifie que $F := \bigcup_{k\in\mathbb{Z}} F_k$ et un fermé de $\mathbb{R}$. Notant $U := \bigcup_{k\in\mathbb{Z}} U_k$, on a $F \subset A \subset U$ et $U \setminus F \subset \bigcup_{k\in\mathbb{Z}} (U_k \setminus F_k)$. Donc $\mu(U \setminus F) \leqslant \sum_{k\in\mathbb{Z}} \mu(U_k \setminus F_k) \leqslant \varepsilon/2 + \sum_{k\in\mathbb{N}^*} 2^{-k-1}\varepsilon = \varepsilon$.
- Montrons (ii) $\Rightarrow$ (iii). D'après (ii), pour tout $k \in \mathbb{N}^*$, il existe un fermé F_k et un ouvert U_k tels que $F_k \subset A \subset U_k$ et $\mu(U_k \setminus F_k) \leqslant 1/k$.
On pose $F' := \bigcup_{k\in\mathbb{N}^*} F_k$ et $U' := \bigcap_{k\in\mathbb{N}^*} U_k$. On a $F' \subset A \subset U'$ et, pour tout $k \in \mathbb{N}^*$, $U' \setminus F' \subset U_k \setminus F_k$, donc $\mu(U' \setminus F') \leqslant \mu(U_k \setminus F_k) \leqslant 1/k$. Ainsi $\mu(U' \setminus F') = 0$.
- Montrons (iii) $\Rightarrow$ (iv). On a par hypothèse $F' \subset A \subset U'$, avec $U' \setminus F'$ négligeable, donc $A = F' \cup (A \setminus F)$. On a $N := A \setminus F \subset U' \setminus F'$, donc N est négligeable.
- On a (iv) $\Rightarrow$ (v), puisqu'une réunion au plus dénombrable de fermés est un borélien.
- Montrons (v) $\Rightarrow$ (i). On a $A = B \cup N$. L'ensemble B est borélien donc il est mesurable et l'ensemble N est négligeable donc il est également mesurable. Par suite la réunion $B \cup N$ est mesurable. ∎

On peut montrer qu'il y existe des parties mesurables au sens de Lebesgue de $\mathbb{R}$ qui ne sont pas des boréliens (*cf.* l'exercice III.1.38).

Corollaire 148. L'ensemble de parties obtenu en réunissant les $\mathcal{F}_\sigma$ de $\mathbb{R}$ et l'ensemble des parties négligeables de $\mathbb{R}$ est une tribu. C'est la tribu de Lebesgue de $\mathbb{R}$.

Remarque. D'après le corollaire ci-dessus, il est possible, au lieu de procéder comme nous l'avons fait à partir de la théorie de Henstock-Kurzweil, de *définir* la famille de parties formant la tribu de Lebesgue en connaissant *seulement* la topologie (c'est-à-dire les ouverts de $\mathbb{R}$) et les ensembles de mesure nulle (que l'on connaît si

l'on a la notion de longueur d'un intervalle). Cela permet de voir que notre approche et la présentation « classique » donnent *la même famille de parties*, c'est-à-dire *la même tribu de Lebesgue*. Notons que, pour *montrer que l'on a bien une tribu* (et pas seulement une famille de parties) nous avons utilisé la théorie de Henstock-Kurzweil. Dans l'exposition classique on utilise une tout autre méthode.

Corollaire 149.
Il existe une *unique* mesure ν sur la tribu de Lebesgue telle que $\nu(]a,b[) := b-a$ pour tout intervalle ouvert borné $]a,b[$. C'est la mesure de Lebesgue ($\nu = \mu$).

Démonstration. Les mesures μ et ν sont égales pour les segments, donc pour les réunions dénombrables de segments et, par suite, pour les ensembles négligeables. Ces mesures sont égales pour les fermés donc pour les $\mathcal{F}_\sigma$ (*cf.* (34)). On en déduit que $\mu = \nu$ en utilisant le corollaire 148 ci-dessus. ■

Ainsi, la mesure sur la tribu de Lebesgue construite classiquement à partir des idées de Borel et Lebesgue (*cf.* par exemple [Ru]) et celle que nous avons définie en utilisant l'intégrale de Henstock-Kurzweil sont égales. Cela justifie *a posteriori* notre terminologie. Le corollaire 149 sera utilisé en probabilités (*cf.* page 675).

Point Méthode

Toute partie mesurable $A \subset \mathbb{R}$ est réunion d'une partie borélienne B (que l'on peut supposer être un $\mathcal{F}_\sigma$) et d'une partie négligeable N. Pour étudier une propriété presque partout sur A, on peut donc remplacer A par B et travailler sur une partie borélienne (et même sur un $\mathcal{F}_\sigma$).

3.2.3 Existence d'un ensemble non mesurable

Soit $A \subset \mathbb{R}$. S'il existe un ouvert *de longueur finie* contenant A (en particulier si A est borné), on définit la *mesure extérieure* $\mu^+(A)$ de A comme la borne inférieure des longueurs des ouverts contenant A.

Proposition 150. Soit $A \subset \mathbb{R}$.

(i) Soit $T : \mathbb{R} \mapsto \mathbb{R}$ une *translation*. Alors $\mu^+\big(T(A)\big) = \mu^+(A)$.

(ii) Si la partie A est intégrable, $\mu(A) = \mu^+(A)$.

(iii) Si $A \subset [a,b]$ $(a < b)$, A est intégrable si, et seulement si :

$$\mu^+(A) + \mu^+\big([a,b] \setminus A\big) = b - a.$$

Démonstration.

(i) Pour tout ouvert $U \subset \mathbb{R}$ de longueur finie, on a $\operatorname{long} T(U) = \operatorname{long} U$. Le résultat s'en déduit.

(ii) Cela résulte du corollaire 138 de la page 316.

(iii) Soit $A \subset [a,b]$. Si A est intégrable $[a,b] \setminus A$ est intégrable et l'on a :

$$\mu^+(A) + \mu^+\big([a,b] \setminus A\big) = b - a.$$

Inversement supposons que $\mu^+(A) + \mu^+\big([a,b] \setminus A\big) = b - a$. Montrons d'abord que, pour tout $\varepsilon > 0$, il existe un fermé $F \subset A$ tel que $\mu^+(A) - \mu(F) \leqslant \varepsilon/2$.
Il existe un ouvert V de $\mathbb{R}$ contenant $[a,b] \setminus A$ tel que $\mu(V) - \mu^+\big([a,b] \setminus A\big) \leqslant \varepsilon/2$.
Notons $V' := V \cap [a,b]$, on a $[a,b] \setminus A \subset V'$ et $\mu(V') - \mu^+\big([a,b] \setminus A\big) \leqslant \varepsilon/2$.
Soit $F := [a,b] \setminus V'$, on a $\mu(F) + \mu(V') = b - a$, donc :

$$-\mu(F) \leqslant a - b + \mu^+\big([a,b] \setminus A\big) + \varepsilon/2 \quad \text{et} \quad -\mu(F) \leqslant a - b + b - a - \mu^+(A) + \varepsilon/2,$$

c'est-à-dire $\mu^+(A) - \mu(F) \leqslant \varepsilon/2$.
Il existe un ouvert U de $\mathbb{R}$ tel que $A \subset U$ et $\mu(U) - \mu^+(A) \leqslant \varepsilon/2$.
Ainsi $F \subset A \subset U$, avec $\mu(U) - \mu(F) \leqslant \varepsilon/2 + \varepsilon/2 = \varepsilon$.
Par suite, d'après le critère (ii) de la proposition 147, la partie A est intégrable. ■

Remarque. Lebesgue utilisait la propriété (ii) pour *définir* les sous-ensembles mesurables de $[a,b]$.

Proposition 151. Il existe une famille dénombrable $(A_n)_{n\in\mathbb{N}}$ de sous-ensembles *disjoints* de $[0,1]$ telle que :

$$\mu^+\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \neq \sum_{n\in\mathbb{N}} \mu^+(A_n). \tag{35}$$

La preuve de cette proposition fait l'objet de l'exercice III.1.30. Elle utilise l'axiome du choix et n'est donc pas constructive. L'idée est une variante de celle utilisée par Giuseppe Vitali pour prouver l'existence d'un sous-ensemble non mesurable de $\mathbb{R}$ en 1905.

Corollaire 152. Il existe un sous-ensemble *non mesurable* de $[0,1]$.

Nous avons utilisé plus haut le corollaire ci-dessus pour montrer qu'il existait une fonction non $\mathscr{HK}$-intégrable dont la valeur absolue est $\mathscr{HK}$-intégrable (*cf.* la remarque de la page 288 et l'exercice 28).

Démonstration. Considérons la famille $(A_n)_{n\in\mathbb{N}}$ de la proposition ci-dessus. Si tous les A_n étaient mesurables, $\bigcup_{n\in\mathbb{N}} A_n$ le serait aussi et l'on aurait :

$$\mu^+\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n) = \sum_{n\in\mathbb{N}} \mu^+(A_n),$$

ce qui contredirait (35). On en déduit que l'un (au moins) des A_n n'est pas mesurable. ■

Remarque. Dans sa (courte) note originale de 1901, Lebesgue passe très rapidement sur les questions d'ensembles mesurables considérées comme « allant de soi ». Dans cet esprit, pour comprendre la notion « classique » d'intégrale de Lebesgue (*cf.* ci-dessous la partie 3.2.5), nous conseillons au lecteur de penser dans un premier temps

« comme si tous les sous-ensembles de $\mathbb{R}$ étaient mesurables » (et en conséquence toutes les fonctions mesurables). Cela permet de s'affranchir de problèmes techniques un peu pénibles et de comprendre l'essentiel de l'idée de Lebesgue (le mathématicien français J. Deny disait « la non mesurabilité est un luxe de mathématicien »). On peut donner une forme rigoureuse à cette démarche en utilisant un résultat de logique de Robert Martin Solovay (1966) : on peut remplacer l'axiomatique habituelle de la théorie des ensembles (ZFC, c'est-à-dire Zermelo-Fraenkel + axiome du choix) par une autre où il n'y a plus d'axiome du choix général mais seulement l'axiome du choix *dénombrable* et où, par contre, *tous* les sous-ensembles de $\mathbb{R}$ sont mesurables. Cette démarche n'est pas nécessairement « déraisonnable », en effet, accepter l'axiome du choix traditionnel permet des « paradoxes » très surprenants comme celui de Banach-Tarski[5]. Toutefois, il est *aussi* possible d'accepter seulement l'axiome du choix *dénombrable* et l'existence d'un ensemble non mesurable de $\mathbb{R}$! C'est un exemple *d'indécidabilité*.

3.2.4 Fonctions mesurables

On vérifie facilement le résultat suivant.

Lemme 153. Soient $f : \mathbb{R} \to \overline{\mathbb{R}}$ une fonction et $\mathcal{T}$ une *tribu* de $\mathbb{R}$ alors l'ensemble de parties $\mathcal{T}_1 := \{A \subset \mathbb{R} \mid f^{-1}(A) \in \mathcal{T}\}$ de $\mathbb{R}$ est une *tribu* de $\mathbb{R}$.

Définition 34. Soient $I \subset \mathbb{R}$ un intervalle. Une fonction $f : I \to \overline{\mathbb{R}}$ est dite *mesurable* si, pour tout $c \in \mathbb{R}$, $f^{-1}\big(]c, +\infty]\big) = \{x \in \overline{\mathbb{R}} \mid f(x) > c\}$ est mesurable.

Proposition et définition 154. Soient $I \subset \mathbb{R}$ un intervalle et $f : I \to \overline{\mathbb{R}}$ une fonction. Les propriétés suivantes sont équivalentes.

(i) Pour tout $c \in \overline{\mathbb{R}}$, $f^{-1}\big([-\infty, c[\big)$ est mesurable.

(ii) Pour tout $c \in \overline{\mathbb{R}}$, $f^{-1}\big([-\infty, c]\big)$ est mesurable.

(iii) Pour tout $c \in \overline{\mathbb{R}}$, $f^{-1}\big(]c, +\infty]\big)$ est mesurable.

(iv) Pour tout $c \in \overline{\mathbb{R}}$, $f^{-1}\big([c, +\infty]\big)$ est mesurable.

(v) L'image réciproque de tout intervalle $J \subset \overline{\mathbb{R}}$ est mesurable.

(vi) L'image réciproque de tout intervalle *ouvert* (respectivement *fermé*) $J \subset \overline{\mathbb{R}}$ est mesurable.

(vii) L'image réciproque de tout *ouvert* (respectivement fermé) de $\overline{\mathbb{R}}$ est mesurable.

Démonstration. Les preuves sont élémentaires et demandent seulement un peu de patience... Nous les laissons au lecteur. ∎

[5] Banach et Tarski ont montré que l'on peut découper une sphère en cinq morceaux de telle sorte qu'en les rassemblant convenablement on obtienne une sphère de rayon double, *cf.* S. Wagon *The Banach-Tarski-paradox*.

Corollaire 155. Soit $f : I \to \overline{\mathbb{R}}$ une fonction *mesurable*, alors l'image réciproque par f de tout *borélien* $A \subset \mathbb{R}$ est mesurable.

Démonstration. Soit $\mathcal{A}$ l'ensemble des parties A de $\mathbb{R}$ telles que $f^{-1}(A)$ soit une partie mesurable. D'après le lemme 153, $\mathcal{A}$ est une tribu. Elle contient tous les intervalles $\{]c, +\infty[\}_{c\in\mathbb{R}}$ et, par suite, la tribu de Borel, d'où le résultat. ■

Proposition 156. Soient $I \subset \mathbb{R}$ un intervalle et $f, g : I \to \overline{\mathbb{R}}$ deux fonctions *mesurables*. Alors :

(i) la fonction f est mesurable si $-f$ l'est ;

(ii) les fonctions $\max(f, g)$ et $\min(f, g)$ sont *mesurables* ;

(iii) soit $(f_i)_{i\in I}$ une famille *au plus dénombrable* de fonctions mesurables, alors les fonctions $\sup\limits_{i\in I} f_i$ et $\inf\limits_{i\in I} f_i$ sont mesurables.

(iv) les fonctions f^+, f^- et $|f|$ sont *mesurables*.

Démonstration.

(i) Pour tout $a \in \overline{\mathbb{R}}$, $f < a$ si, et seulement si, $-f > -a$.

(ii) Notons $h := \max(f, g)$. Pour tout $a \in \overline{\mathbb{R}}$, l'ensemble $\{x \in I \mid h(x) > a\}$ est la réunion des ensembles $\{x \in I \mid f(x) > a\}$ et $\{x \in I \mid g(x) > a\}$. Donc si f et g sont mesurables, h l'est aussi.
Pour $\min(f, g)$, on utilise (i) et $\min(f, g) = -\max(-f, -g)$.

(iii) Notons $h = \sup\limits_{i\in I} f_i$. L'ensemble $\{x \in I \mid h(x) > a\}$ est la réunion (au plus dénombrable) $\bigcup\limits_{i\in I} \{x \in I \mid f_i(x) > a\}$.
Les ensembles $\{x \in I \mid f_i(x) > a\}$ sont mesurables, donc $\{x \in I \mid h(x) > a\}$ l'est.
Pour $\inf\limits_{i\in I} f_i$, on utilise $\inf\limits_{i\in I} f_i = -\sup\limits_{i\in I}(-f_i)$

(iv) Le résultat se déduit de (i) et (ii) :

$$f^+ = \max(f, 0) \qquad f^- = \max(-f, 0) \quad \text{et} \quad |f| = \max(f, -f).$$ ■

Proposition 157. Soient $(f_1, \ldots, f_m)$ une famille de fonctions *mesurables* définies sur un intervalle $I \subset \mathbb{R}$, à valeurs dans $\mathbb{R}$ et $g : \mathbb{R}^m \to \mathbb{R}$ une application *continue*. Alors $G := g \circ (f_1, \ldots, f_m)$ est *mesurable*.

Démonstration. Soit $V \subset \mathbb{R}$ un ouvert. L'ensemble $g^{-1}(V)$ est un ouvert de $\mathbb{R}^m$ donc $g^{-1}(V) = \bigcup\limits_{k\in\mathbb{N}} U_k$, les $U_k := U_k^1 \times \cdots \times U_k^m$ étant des pavés ouverts de $\mathbb{R}^m$ (c'est une variante facile de la proposition 186 de la page 347). Il en résulte que :

$$G^{-1}(V) = \bigcup_{k\in\mathbb{N}} (f_1, \ldots, f_m)^{-1}(U_k) = \bigcup_{k\in\mathbb{N}} \left(f_1^{-1}(U_k^1) \cap \ldots \cap f_m^{-1}(U_k^m)\right)$$

est mesurable. ■

On déduit immédiatement le corollaire suivant de la proposition ci-dessus.

Corollaire 158. Une combinaison linéaire de fonctions mesurables à valeurs dans $\mathbb{R}$ est mesurable. Le produit de deux fonctions mesurables à valeurs dans $\mathbb{R}$ est mesurable. L'ensemble des fonctions mesurables sur un intervalle $I \subset \mathbb{R}$, à valeurs dans $\mathbb{R}$, est une sous-algèbre de $\mathcal{F}(I)$.

Proposition 159.

1. Toute fonction continue est mesurable.
2. Une fonction $f : I \to \mathbb{R}$ est mesurable si, et seulement si, sa restriction à tout sous-segment de I est mesurable.
3. Toute fonction continue par morceaux est mesurable.

Démonstration. Soit $f : I \to \mathbb{R}$.

1. Si f est continue, pour tout intervalle *ouvert* $J \subset \mathbb{R}$, $f^{-1}(J)$ est un ouvert de I, donc mesurable.
2. Soit $J \subset \mathbb{R}$ un intervalle. Alors, d'après les définitions, $f^{-1}(J)$ est mesurable si, et seulement si, pour tout segment $K \subset I$, $f^{-1}(J) \cap K$ et mesurable.
3. Il suffit de traiter le cas d'une fonction continue par morceaux sur un segment. On se ramène ensuite, en utilisant l'addition des fonctions, au cas d'une fonction $g : \mathbb{R} \to \mathbb{R}$ continue sur un intervalle borné et nulle en dehors. On vérifie facilement qu'une telle fonction est mesurable. ∎

Proposition 160. Soient $I \subset \mathbb{R}$ un intervalle et $(f_k : I \to \overline{\mathbb{R}})_{k\in\mathbb{N}}$ une suite de fonctions *mesurables*. Alors $\limsup f_k$ et $\liminf f_k$ sont mesurables. Si la suite admet une limite simple f presque partout celle-ci est mesurable.

Démonstration. Notons $f := \limsup f_k$. Par définition $f(x) < c$ si, et seulement s'il existe $m \in \mathbb{N}^*$ et $N \in \mathbb{N}$ tels que pour tout $k \geqslant N$, $f_k(x) \leqslant c - 1/m$, c'est-à-dire :

$$f^{-1}\big([-\infty, c[\big) = \bigcup_{m\in\mathbb{N}^*} \bigcup_{N\in\mathbb{N}} \bigcap_{k\geqslant N} f_k^{-1}\big([-\infty, c - 1/m[\big).$$

Les f_k étant mesurables, les ensembles $f_k^{-1}\big([-\infty, c - 1/m[\big)$ sont mesurables, donc l'ensemble $f^{-1}\big([-\infty, c[\big)$ est mesurable et f est mesurable.

Pour montrer que $\liminf f_k$ est mesurable, on utilise le résultat pour $\limsup$ en remplaçant f par $-f$: $\liminf f_k = -\limsup(-f_k)$.

Si $(f)_{k\in\mathbb{N}}$ admet une limite f presque partout, on a $f = \limsup f_k$ presque partout, donc f est mesurable. ∎

Proposition 161. Soit $I \subset \mathbb{R}$ un segment.
Toute fonction $f : I \to \mathbb{R}$ $\mathscr{HK}$-intégrable sur I est *mesurable*.

Démonstration. La fonction $(x, y) \in I \times I \to F(x, y) := \displaystyle\int_x^y f(t)\,\mathrm{d}t$ est *continue*. Soit en effet $a \in I$; les fonctions $x \mapsto \displaystyle\int_x^a f(t)\,\mathrm{d}t$ et $y \mapsto \displaystyle\int_a^y f(t)\,\mathrm{d}t$ sont continues (*cf.* la proposition 97 de la page 282). On prolonge F en une fonction sur $\mathbb{R}^2$ notée $\tilde{F}$ nulle en dehors de $I \times I$. D'après la proposition 101 de la page 286, $f(x) = \lim\limits_{h\to 0} \frac{1}{h}\tilde{F}(x, x+h)$ pour presque tout $x \in I$. Posons, pour tout $k \in \mathbb{N}^*$ et tout $x \in I$, $f_k(x) := k\tilde{F}(x, x + 1/k)$. Les fonctions f_k, $k \in \mathbb{N}^*$, sont continues par morceaux, donc mesurables, et f est presque partout égale à la limite de la suite $(f_k)_{k\in\mathbb{N}^*}$, donc, d'après la proposition 160 ci-dessus, f est mesurable. ∎

La proposition suivante dérive de l'idée originale de Lebesgue : diviser l'axe des ordonnées au lieu de l'axe des abcisses comme pour l'intégrale de Riemann.

Proposition 162. Soit $I \subset \mathbb{R}$ un intervalle. Toute fonction mesurable $f : I \to \overline{\mathbb{R}}$ peut s'écrire $f = \lim\limits_{k\to\infty} f_k$ où les f_k sont des fonctions *étagées* (*cf.* page 317).
Si de plus $f \geqslant 0$, on peut choisir la suite (f_k) croissante et telle que, pour tout $k \in \mathbb{N}$, on ait $0 \leqslant f_k \leqslant f$.

Démonstration. On suppose d'abord que I est un *segment*.
On note, pour tout $k \in \mathbb{N}$, $N_k := k2^{k+1} + 2$. On va définir, par des réels $\{\lambda_{k,h}\}_{h\in[\![1,N_k]\!]}$, une partition $\mathcal{P}$ de $\overline{\mathbb{R}}$ et en déduire une *partition mesurable* $(A_{k,h})_{h\in[\![1,N_k]\!]}$ de I par $f^{-1}(\mathcal{P})$.
Pour définir $\mathcal{P}$, on écrit d'abord $\overline{\mathbb{R}} = [-\infty, -k[\cup[-k, k[\cup[k, +\infty]$, puis on coupe $[-k, k[$ en parties égales de longueur 2^{-k} :

$$\begin{cases} \lambda_{k,h} := -k, & A_{k,h} := f^{-1}\big([-\infty, -k[\big) & \text{si} \quad h = 0 \\ \lambda_{k,h} := (h-1)\,2^{-k} - k, & A_{k,h} := f^{-1}\big(\big[\lambda_{k,h}, \lambda_{k,h} + 2^{-k}\big[\big) & \text{si} \quad h = 1, 2, \ldots, k\,2^{k+1} \\ \lambda_{k,h} := k, & A_{k,h} := f^{-1}\big([k, +\infty]\big) & \text{si} \quad h = k\,2^{k+1} + 1 \end{cases}.$$

On définit, pour tout $k \in \mathbb{N}$, une *fonction* $f_k := \sum\limits_{k\in[\![0,N_k]\!]} \lambda_{k,h} 1_{A_{k,h}}$. Les $A_{k,h}$ étant *intégrables* ($A_{k,h}$ est mesurable et $A_{k,h} \subset I$), f_k est une *fonction étagée*.
Si $\big|f(x)\big| < k$, $\big|f(x) - f_k(x)\big| \leqslant 2^{-k}$. Si x est fixé, $\big|f(x)\big| < k$ pour $k \in \mathbb{N}$ assez grand, donc la suite $(f_k(x)))$ converge vers $f(x)$.
Si I n'est pas borné, on remplace les $A_{k,h}$ par les $B_{k,h} := A_{k,h} \cap [-k, k]$.
Si de plus $f \geqslant 0$, on vérifie que la suite (f_k) est croissante et telle que, pour tout $k \in \mathbb{N}$, on ait $0 \leqslant f_k \leqslant f$. ∎

Corollaire 163. Une combinaison linéaire à coefficients *positifs* de fonctions mesurables à valeurs dans $\mathbb{R} \cup \{+\infty\}$ (respectivement $\mathbb{R} \cup \{-\infty\}$) est mesurable.

Démonstration. Ce résultat étend celui du corollaire 158 aux fonctions à valeurs dans $\mathbb{R} \cup \{+\infty\}$ (respectivement $\mathbb{R} \cup \{-\infty\}$). Pour le prouver, il suffit de remarquer qu'une telle combinaison a toujours un sens et que toute combinaison linéaire de fonctions étagées est une fonction étagée. On prendra garde au fait que $+\infty - \infty$ n'a pas de sens et que, par suite, la somme de deux fonctions mesurables n'est pas toujours définie. ■

Les ensembles négligeables sont mesurables, par suite, pour vérifier qu'une fonction est mesurable, on peut la modifier arbitrairement sur un ensemble négligeable.
La proposition suivante permet de préciser les relations entre mesurabilté et $\mathscr{HK}$-intégrabilité.

Proposition 164. Soient $f,\ g,\ h : I \to \mathbb{R}$ telles que $h \leqslant f \leqslant g$.
Si g et h sont $\mathscr{HK}$*-intégrables* et si f est *mesurable*, alors f est $\mathscr{HK}$*-intégrable*.

Démonstration. En remplaçant si nécessaire f par $f - h$ et g par $g - h$ on peut supposer que $0 \leqslant f \leqslant g$ avec g $\mathscr{HK}$-intégrable (donc absolument intégrable puisqu'elle est positive). D'après la proposition 162 ci-dessus, $f = \lim f_k$, les fonctions f_k étant étagées et positives et la suite croissante. Une fonction étagée est $\mathscr{HK}$ intégrable et la suite de fonctions positives (f_k) est majorée par g qui est absolument intégrable. Donc f est $\mathscr{HK}$-intégrable d'après le théorème de convergence dominée. ■

Corollaire 165. Une fonction *mesurable* et *bornée* sur un *segment* I est intégrable au sens de Lebesgue sur I.

Démonstration. Soit $f : I \to \mathbb{R}$ *mesurable* et *bornée*.
Il existe $M > 0$ tel que $-M \leqslant f \leqslant M$. D'après la proposition ci-dessus f est $\mathscr{HK}$-intégrable sur I.
La fonction $|f|$ est mesurable (*cf.* la proposition 156) et $0 \leqslant |f| \leqslant M$. D'après la proposition ci-dessus $|f|$ est $\mathscr{HK}$-intégrable sur I. Ainsi f est absolument intégrable, c'est-à-dire intégrable au sens de Lebesgue, sur I. ■

> ***Remarque***. Compte tenu des propriétés des fonctions mesurables décrites plus haut, on peut dire, en termes « imprécis », que toute fonction que l'on peut décrire « explicitement » est mesurable. Ainsi, en pratique, toutes les fonctions bornées « que l'on rencontre » sont intégrables au sens de Lebesgue.

Le résultat suivant précise, pour les suites de fonctions mesurables, la relation entre convergence simple et convergence uniforme.
Soit $E \subset \mathbb{R}$. On dira qu'une fonction $f : E \to \mathbb{R}$ est mesurable si la fonction obtenue en la prolongeant par 0 sur $\mathbb{R} \setminus E$ est mesurable.

Proposition 166 (Théorème d'Egorov).
Soient $E \subset \mathbb{R}$ un sous-ensemble *intégrable* et $(f_k : I \to \mathbb{R})_{k\in\mathbb{N}}$ une suite de fonctions *mesurables* sur E convergeant presque partout vers une fonction $f : E \to \mathbb{R}$. Alors, pour tout réel $\varepsilon > 0$, il existe un sous-ensemble $A \subset E$ intégrable tel que $\mu(E) - \varepsilon \leqslant \mu(A) \leqslant \mu(E)$ et sur lequel la convergence de la suite est *uniforme*, c'est-à-dire $\lim\limits_{k\to+\infty} \sup\limits_{x\in A} \big|f_k(x) - f(x)\big| = 0$.

Démonstration. Notons $B \subset E$ l'ensemble des points où la suite converge simplement. On a, par hypothèse, $\mu(B) = \mu(E)$. Pour tous $p \in \mathbb{N}^*$ et $n \in \mathbb{N}$, on note :

$$B_{p,n} = \{x \in B \mid \sup_{k\geqslant n} \big|f_k(x) - f(x)\big| \leqslant 1/p\}.$$

Les $B_{p,n}$ sont mesurables et, pour tout $p \in \mathbb{N}^*$, la suite $(B_{p,n})_{n\in\mathbb{N}}$ est croissante et vérifie $\bigcup\limits_{n\in\mathbb{N}} B_{p,n} = B$. Ainsi $\lim\limits_{n\to+\infty} 1_{B_{p,n}} = 1_B$ et, d'après le théorème de convergence monotone, $\lim\limits_{n\to+\infty} \mu(B_{p,n}) = \mu(B) = \mu(E)$.

Par suite, pour tout $p \in \mathbb{N}^*$, il existe $n_p \in \mathbb{N}$ tel que $\mu(B_{p,n_p}) > \mu(E) - \varepsilon/2^p$.

Soit $A := \bigcap\limits_{p\in\mathbb{N}^*} B_{p,n_p}$, c'est un ensemble mesurable et l'on a :

$$\mu(E \setminus A) \leqslant \sum_{p\in\mathbb{N}^*} \mu(E \setminus B_{n,n_p}) \leqslant \sum_{p\in\mathbb{N}^*} \varepsilon/2^p = \varepsilon.$$

De $\mu(A) + \mu(E \setminus A) = \mu(E)$, on déduit $\mu(E) - \varepsilon \leqslant \mu(A)$.

Par ailleurs, si $x \in A$, pour tout $p \in \mathbb{N}^*$ on a, pour tout $k \geqslant n_p$, $\big|f_k(x) - f(x)\big| \leqslant 1/p$, ce qui prouve la convergence uniforme sur A. ∎

3.2.5 Caractérisation des fonctions intégrables au sens de Lebesgue

Comparaison avec la théorie de Lebesgue

Nous allons montrer que l'intégration « au sens de Lebesgue » définie plus haut dans la partie 2.6 à l'aide de la théorie de Henstock-Kurzweil (*cf.* la définition 24 de la page 288) et l'intégration définie originellement par Lebesgue à partir de « sa » mesure (que l'on trouve aujourd'hui, à quelques variations près, dans de très nombreux ouvrages, par exemple, [Ru]) coïncident.

Ce résultat peut paraître surprenant compte tenu de la grande différence apparente des définitions.

Définition 35. Soit $f : \mathbb{R} \to [0, +\infty]$ une fonction *mesurable* positive.

On définit son intégrale $\displaystyle\int_I f \, \mathrm{d}\mu$ pour la mesure de Lebesgue μ sur I comme la borne supérieure (dans $\mathbb{R}_+ \cup \{+\infty\}$) des intégrales des *fonctions étagées* majorées par f.

Si $A \subset \mathbb{R}$ est mesurable, on définit $\displaystyle\int_A f \, \mathrm{d}\mu := \int_{\mathbb{R}} f 1_A \, \mathrm{d}\mu$.

La fonction f est dite $\mathscr{L}$-intégrable sur I si $\displaystyle\int_I f \, \mathrm{d}\mu < +\infty$.

Il n'y a pas d'ambiguïté sur la notion de fonction étagée car elle découle immédiatement de la connaissance de la mesure. L'intégrale d'une fonction étagée est la même dans l'approche $\mathscr{HK}$ et dans l'approche classique comme nous l'avons vu plus haut.
Il résulte de la définition que si f et g sont positives et mesurables, avec $f \leqslant g$, et si g est $\mathcal{L}$-intégrable, alors f l'est.

La proposition suivante montre que les théories de Lebesgue (classique) et de Kurzweil-Henstock *coïncident* pour les fonctions *positives* partout finies.

Proposition 167. Soient I un intervalle réel et $f : I \to \mathbb{R}_+$. Les trois propriétés suivantes sont équivalentes :

(i) f est $\mathscr{L}$-intégrable sur I ;

(ii) f est $\mathscr{HK}$-intégrable sur I ;

(iii) f est intégrable au sens de Lebesgue sur I.

Si elles sont satisfaites, les intégrales $\displaystyle\int_I f \,\mathrm{d}\mu$, au sens de Lebesgue, et $\displaystyle\int_I f$, au sens de Kurzweil-Henstock, coïncident.

Démonstration.
Montrons (i) $\Rightarrow$ (ii). Supposons f $\mathscr{L}$-intégrable sur I. Elle est en particulier mesurable, donc, d'après la proposition 162, il existe une suite croissante de fonctions étagées positives $(f_k)_{k\in\mathbb{N}}$ convergeant vers f. On a, pour tout $k \in \mathbb{N}$, $\displaystyle\int_I f_k \leqslant \int_I f \,\mathrm{d}\mu < +\infty$.
D'après le théorème de convergence dominée, f est $\mathscr{HK}$-intégrable et :

$$\int_I f = \lim_{k\to+\infty} \int_I f_k \leqslant \int_I f \,\mathrm{d}\mu.$$

L'assertion (ii) $\Rightarrow$ (iii) est triviale puisque $f \geqslant 0$.
Montrons (iii) $\Rightarrow$ (i). Supposons f intégrable au sens de Lebesgue. D'après la proposition 161 de la page 330, elle est mesurable.

Par ailleurs, si g est une fonction étagée sur I telle que $g \leqslant f$, on a $\displaystyle\int_I g \leqslant \int_I f$, donc $\displaystyle\int_I f \,\mathrm{d}\mu \leqslant \int_I f < +\infty$ et f est $\mathscr{L}$-intégrable.
Si les assertions sont satisfaites, on a :

$$\int_I f \leqslant \int_I f \,\mathrm{d}\mu \quad \text{et} \quad \int_I f \,\mathrm{d}\mu \leqslant \int_I f, \qquad \text{donc} \qquad \int_I f \,\mathrm{d}\mu = \int_I f. \qquad \blacksquare$$

Proposition et définition 168. Soient I un intervalle réel et $f : I \to \overline{\mathbb{R}}$ une fonction *mesurable*. On dit que f est $\mathscr{L}$-intégrable sur I si $|f|$ l'est, c'est-à-dire si $\displaystyle\int_I |f| \,\mathrm{d}\mu < +\infty$. S'il en est ainsi les fonctions positives f^+ et f^- sont $\mathscr{L}$-intégrables sur I et l'on *définit* l'intégrale de Lebesgue de f par $\displaystyle\int_I f \,\mathrm{d}\mu := \int_I f^+ \,\mathrm{d}\mu - \int_I f^- \,\mathrm{d}\mu$.

Démonstration. Si $|f|$ est mesurable, f^+ et f^- le sont aussi.
On a $0 \leqslant f^+ \leqslant |f|$ et $0 \leqslant f^- \leqslant |f|$, donc :

$$\int_I f^+ \,\mathrm{d}\mu \leqslant \int_I |f| \,\mathrm{d}\mu < +\infty \qquad \text{et} \qquad \int_I f^- \,\mathrm{d}\mu \leqslant \int_I |f| \,\mathrm{d}\mu < +\infty. \qquad \blacksquare$$

Le théorème suivant montre que les théories de Lebesgue et de Kurzweil-Henstock *coïncident* pour les fonctions absolument intégrables partout finies. L'équivalence (ii) $\Leftrightarrow$ (iii) figure ici « pour mémoire » puisque ce n'est qu'une question de définitions.

Théorème 169. Soient I un intervalle réel et $f : I \to \mathbb{R}$. Les trois propriétés suivantes sont équivalentes :

(i) f est $\mathscr{L}$-intégrable sur I ;

(ii) f est absolument intégrable sur I au sens de Henstock-Kurzweil, c'est-à-dire f et $|f|$ sont $\mathscr{HK}$-intégrables sur I ;

(iii) f est intégrable au sens de Lebesgue sur I.

Si elles sont satisfaites, les intégrales $\displaystyle\int_I f \,\mathrm{d}\mu$, au sens de Lebesgue, et $\displaystyle\int_I f$, au sens de Kurzweil-Henstock, coïncident.

Démonstration. On applique la proposition 167 aux fonctions *positives* f^+ et f^-. $\blacksquare$

Finalement, on peut considérer la théorie de Lebesgue sur $\mathbb{R}$ comme un « sous-produit » de celle de Kurzweil-Henstock. Il y a des résultats analogues pour les théories sur $\mathbb{R}^n$ mais leur étude sort du cadre de notre ouvrage (*cf.* par exemple [De]).

3.3 Espaces L^1 et L^2

3.3.1 Espaces L^1

Rappelons et complétons le corollaire 108 de la page 292 (l'assertion (ii) résulte immédiatement des définitions).

Proposition 170. Soit $I \subset \mathbb{R}$ un intervalle.

(i) Le sous-ensemble $\tilde{L}^1(I) \subset \mathcal{F}(I, \mathbb{R})$ des fonctions intégrables au sens de Lebesgue sur I est un sous-espace vectoriel de $\mathcal{F}(I, \mathbb{R})$ et
$$\begin{array}{rcl} \tilde{L}^1(I) & \longrightarrow & \mathbb{R}_+ \\ f & \longmapsto & \|f\|_1 := \displaystyle\int_I |f| \end{array}$$
est une *semi-norme* sur $\tilde{L}^1(I)$.

(ii) Le sous-ensemble $N(I) \subset \tilde{L}^1(I)$ des fonctions négligeables sur I, c'est-à-dire des fonctions f telles que $\|f\|_1 = 0$, est un sous-espace vectoriel de $\tilde{L}^1(I)$.

Si $I \subset \mathbb{R}$ est un intervalle, on note $L^1(I) := \tilde{L}^1(I)/N(I)$ l'espace vectoriel quotient. D'après la proposition 136 de la page 314 et le corollaire 139, les classes d'équivalence sont formées de fonctions égales presque partout sur I. Nous nous conformerons à la tradition

et noterons ces classes comme des fonctions avec la convention implicite que l'on peut modifier arbitrairement une fonction sur un ensemble négligeable (ce qui ne change pas son intégrale). Si $f \in \tilde{L}^1(I)$, on a $f \in N(I)$ si, et seulement si, $\displaystyle\int_I |f| = 0$, donc $\|.\|_1$ définit (par passage au quotient) une *norme* sur $L^1(I)$. Si une fonction sur I prend la valeur $\pm\infty$ sur un ensemble A de mesure nulle et est $\mathcal{L}$-intégrable, on peut lui associer un élément de $L^1(I)$ en la modifiant sur A.
Si I est d'intérieur non vide, le théorème 110 de la page 293 permet d'identifier $L^1(I)$ et $L^1(\operatorname{Int} I)$.

Théorème 171 (Théorème de Fischer-Riesz). Soit I un intervalle réel.

(i) Soit $(f_n : I \to \mathbb{R})_{n\in\mathbb{N}}$ une suite de fonctions intégrables au sens de Lebesgue telle que $\sum\limits_{n\in\mathbb{N}} \|f_n\|_1 < +\infty$. Alors la série $\sum\limits_{n\in\mathbb{N}} f_n$ converge dans $L^1(I)$ et la convergence a lieu ponctuellement (et absolument) presque partout sur I.

(ii) L'espace $\big(L^1(I), \|.\|_1\big)$ est un espace de Banach, de plus de toute suite de Cauchy de cet espace on peut « extraire » une suite convergeant ponctuellement presque partout sur I.

(iii) L'application $f \in L^1(I) \mapsto \displaystyle\int_I f$ est une forme linéaire continue sur $L^1(I)$.

Démonstration.

(i) C'est une conséquence immédiate de la proposition 144.

(ii) Cette assertion se déduit de la précédente. Soit $(h_k)_{k\in\mathbb{N}}$ une suite de Cauchy de $L^1(I)$. Pour tout $m \in \mathbb{N}$, il existe $k_m \in \mathbb{N}$ tel que, pour tous $p, q \geqslant k_m$ l'on ait $\big\|h_p - h_q\big\|_1 \leqslant 2^{-m}$. Quitte à remplacer, par récurrence, k_m par $1+\max(k_0, \ldots, k_m)$, on peut supposer que la suite $(k_m)_{m\in\mathbb{N}}$ est *strictement croissante*.
Considérons la suite $(h_{k_m})_{m\in\mathbb{N}}$ extraite de $(h_k)_{k\in\mathbb{N}}$.
Pour tout $m \in \mathbb{N}$, on a $\big\|h_{k_{m+1}} - h_{k_m}\big\|_1 \leqslant 2^{-m}$. Donc la série $\sum(h_{k_{m+1}} - h_{k_m})$ converge *normalement* dans $L^1(I)$ vers $S \in L^1(I)$, ce qui prouve que la sous-suite $(h_{k_m})_{m\in\mathbb{N}}$ converge pour la norme $\|.\|_1$ et aussi presque partout vers $S + h_{k_0}$.
On vérifie que la suite initiale $(h_k)_{k\in\mathbb{N}}$ converge vers $S + h_{k_0}$ pour la norme $\|.\|_1$.

(iii) On a $\Big|\displaystyle\int_I f - \int_I g\Big| \leqslant \int_I |f - g| = \|f - g\|_1$, d'où le résultat. ∎

Une suite de Cauchy $(f_k)_{k\in\mathbb{N}}$ de $\big(L^1(I), \|.\|_1\big)$ converge pour la norme $\|.\|_1$ vers $f \in \big(L^1(I), \|.\|_1\big)$ mais ne converge pas, en général, presque partout vers f. Il est même possible que $(f_k)_{k\in\mathbb{N}}$ ne converge *en aucun point* de I (*cf.* l'exercice III.1.32 de la page 369).

Définition 36. Soient $I \subset \mathbb{R}$ un intervalle et $f : I \to \mathbb{R}$. On appelle *support* de f et l'on note $\operatorname{supp} f$ l'adhérence dans I de l'ensemble des points où f est non nulle : $\operatorname{supp} f := \overline{\{x \in I \mid f(x) \neq 0\}}$.
Si $\operatorname{supp} f$ est un compact, on dit que f est à support compact.

Soit $I \subset \mathbb{R}$ un intervalle de $\mathbb{R}$. On note $\mathcal{C}_c(I)) \subset \mathcal{C}(I)$ l'ensemble des fonctions continues à support compact sur I. On vérifie que c'est un sous-espace vectoriel de $\mathcal{C}(I)$.
La preuve du résultat suivant fait l'objet de l'exercice III.1.34.

Proposition 172. Soit $I \subset \mathbb{R}$ un intervalle. L'espace $\mathcal{C}_c(I)$ des fonctions continues à support compact sur I est dense dans $L^1(I)$.

3.3.2 Fonctions de carré sommable — Espaces de Hilbert L^2 — Séries de Fourier

Soit $I \subset \mathbb{R}$ un intervalle. Si $f,\ g \in \mathcal{C}_c(I),\ fg \in \mathcal{C}_c(I)$. On définit une forme bilinéaire symétrique :

$$\mathcal{C}_c(I) \times \mathcal{C}_c(I) \to \mathbb{R} \quad \text{par} \quad (f,g) \mapsto (\, f \mid g\,) := \int_I fg$$

(la fonction continue à support compact fg est absolument intégrable sur I au sens de Riemann généralisé).

Si $f \in \mathcal{C}_c(I)$ et si $(\, f \mid f\,) = \displaystyle\int_I f^2 = 0$, alors $f = 0$, donc $(\,.\mid .\,)$ est un *produit scalaire*. On note $\|.\|_2$ la norme associée : $\|f\|_2^2 := (\, f \mid f\,)$. L'espace normé $\big(\mathcal{C}_c(I), \|.\|_2\big)$ est un espace préhilbertien réel (*cf.* le module *Espaces vectoriels normés* du livre [L2]). Cet espace n'est pas complet. En utilisant l'intégrale de Lebesgue, nous allons définir un espace de Hilbert $L^2(I)$ « contenant » $\mathcal{C}_c(I)$ (on verra que $\mathcal{C}_c(I)$ est en fait *dense* dans $L^2(I)$).

Proposition et définition 173. Soit $I \subset \mathbb{R}$ un intervalle.

(i) Le sous-ensemble $\tilde{L}^2(I) \subset \mathcal{F},\mathbb{R}(I)$ des fonctions f mesurables sur I telles que f^2 soit $\mathscr{HK}$-intégrable sur I en est un sous-espace vectoriel, appelé *espace des fonctions de carré sommable* sur I.

(ii) Si $f,\ g \in \tilde{L}^2(I)$, alors $fg \in \tilde{L}^1(I)$ et l'application :

$$(f,g) \in \tilde{L}^2(I) \times \tilde{L}^2(I) \to (\, f \mid g\,) := \int_I fg \in \mathbb{R},$$

définit une forme bilinéaire sur $\tilde{L}^2(I)$, dont la forme quadratique associée est *positive*.

(iii) L'application $f \mapsto \|f\|_2 := \Big(\displaystyle\int_I f^2\Big)^{1/2}$ définit une *semi-norme* sur $\tilde{L}^2(I)$, c'est-à-dire que :

$$\forall \lambda \in \mathbb{R}, \forall f,g \in \tilde{L}^2(I), \quad \|\lambda f\|_2 = |\lambda|\|f\|_2 \quad \text{et} \quad \|f+g\|_2 \leqslant \|f\|_2 + \|g\|_2.$$

(iv) Les fonctions négligeables forment un sous-espace vectoriel $N(I)$ de $\tilde{L}^2(I)$.

Démonstration.

(i) Soient $\lambda \in \mathbb{R}$ et $f,\ g \in \tilde{L}^2(I)$. On a évidemment $\lambda f \in \tilde{L}^2(I)$. On a :

$$(f+g)^2 \leqslant \big(|f|+|g|\big)^2 \leqslant 2\,(f^2+g^2).$$

Ainsi $(f+g)^2$ est majorée par une fonction $\mathscr{HK}$ -intégrable sur I.

Par ailleurs $f+g$ est mesurable, donc $(f+g)^2$ est $\mathscr{HK}$ -intégrable sur I, c'est-à-dire que $f+g \in \tilde{L}^2(I)$.

(ii) On a $(f+g)^2 - f^2 - g^2 = 2fg$. Puisque $(f+g)^2,\ f^2,\ g^2 \in \tilde{L}^1(I),\ fg \in \tilde{L}^1(I)$. L'application : $(f,g) \in \tilde{L}^2(I) \times \tilde{L}^2(I) \to (\,f \mid g\,) := \displaystyle\int_I fg \in \mathbb{R}$, définit une forme bilinéaire positive sur $\tilde{L}^2(I)$, dont la forme quadratique associée est *positive*..

(iii) Soient $\lambda \in \mathbb{R}$ et $f,\ g \in \tilde{L}^2(I)$. On a $\|\lambda f\|_2 = |\lambda|\|f\|_2$. Par ailleurs, en appliquant l'inégalité de Cauchy-Schwarz (*cf.* le module *Espaces préhilbertiens* du livre [L2]), on obtient :

$$\Big|\int_I fg\Big| \leqslant \Big(\int_I f^2\Big)^{1/2}\Big(\int_I g^2\Big)^{1/2} = \|f\|_2\,\|g\|_2,$$

d'où $\|f+g\|_2^2 \leqslant \big(\|f\|_2 + \|g\|_2\big)^2$.

(iv) D'après la proposition 136 de la page 314, la fonction f appartient à $N(I)$ si, et seulement si, $f^2 \in N(I)$.

Donc $N(I) \subset \tilde{L}^2(I)$. ■

Si $I \subset \mathbb{R}$ est un intervalle, on note $L^2(I) := \tilde{L}^2(I)/N(I)$ l'espace vectoriel quotient. Si $f \in \tilde{L}^2(I)$, on a $f \in N(I)$ si, et seulement si, $\|f\|_2 = 0$, donc $\|.\|_2$ définit (par passage au quotient) une *norme* sur $L^2(I)$.

Si I est d'intérieur non vide, le théorème 110 de la page 293 permet d'identifier $L^2(I)$ et $L^2(\operatorname{Int} I)$.

Théorème 174. Soit $I \subset \mathbb{R}$ un intervalle.

(i) Soit $(f_k : I \to \mathbb{R})_{k\in\mathbb{N}}$ une suite de fonctions intégrables au sens de Lebesgue telle que $\sum\limits_{k\in\mathbb{N}} \|f_k\|_2 < +\infty$. Alors la série $\sum\limits_{k\in\mathbb{N}} f_k$ converge dans $L^2(I)$ et la convergence a lieu ponctuellement presque partout sur I.

(ii) L'espace $\big(L^2(I), \|.\|_2\big)$ est un espace de Banach. De plus, de toute suite de Cauchy de cet espace on peut extraire une suite convergeant ponctuellement presque partout sur I.

(iii) Si $f,\ g \in L^2(I)$, alors $fg \in L^1(I)$ et l'application :

$$(f,g) \in L^2(I) \times L^2(I) \to (\,f \mid g\,) := \int_I fg \in \mathbb{R},$$

définit un *produit scalaire* sur $L^2(I)$ tel que $(\,f \mid f\,) = \|f\|_2^2$.

Ainsi $\big(L^2(I), \|.\|_2\big)$ est un *espace de Hilbert*.

Démonstration.

(i) Notons, pour tout $k \in \mathbb{N}$, $G_k := |f_0| + \cdots + |f_k|$. Pour tout $x \in I$ la suite $\big(G_k(x)\big)_{k\in\mathbb{N}}$ est croissante. On note $G(x) \in \overline{\mathbb{R}}_+ := \mathbb{R} \cup \{+\infty\}$ sa limite.

On a $G_k \in L^2(I)$, $\Big(G_k^2 = \big(|f_0| + \cdots + |f_k|\big)^2\Big)_{k\in\mathbb{N}}$ est une suite croissante de fonctions intégrables au sens de Lebesgue et l'on a, pour tout $k \in \mathbb{N}$:

$$\|G_k^2\|_1 = \|G_k\|_2^2 \leqslant \Big(\sum_{n=0}^{k} \|f_n\|_2\Big)^2 \leqslant \Big(\sum_{n\in\mathbb{N}} \|f_n\|_2\Big)^2$$

D'après le théorème de convergence monotone (théorème 140 de la page 318), la suite $(G_k^2)_{k\in\,\mathbb{N}}$ converge presque partout sur I vers G^2 et G^2 est presque partout finie. Par suite G est presque partout finie et la série $\sum f_k$ est presque partout absolument convergente, donc convergente, notons F sa somme (en convenant de $F(x) = 0$ aux points où $G(x) = +\infty$). La fonction F est limite simple de fonctions mesurables, donc elle est mesurable. On a $|F|^2 \leqslant G^2$, donc $F \in L^2(I)$.

Pour tout $k \in \mathbb{N}$, $\Big\|F - \sum\limits_{h=0}^{k} f_h\Big\|_2 \leqslant \sum\limits_{h\geqslant k+1} \|f_h\|_2$, et $\sum\limits_{h\geqslant k+1} \|f_h\|_2$ tend vers 0 quand k tend vers $+\infty$, donc la série $\sum\limits_{k\in\mathbb{N}} f_k$ converge vers F dans $L^2(I)$.

(ii) Cette assertion se déduit de la précédente. Soit $(h_k)_{k\in\mathbb{N}}$ une suite de Cauchy de $L^2(I)$. Pour tout $m \in \mathbb{N}$, il existe $k_m \in \mathbb{N}$ tel que, pour tous $p, q \geqslant k_m$ l'on ait $\big\|h_p - h_q\big\|_2 \leqslant 2^{-m}$. Quitte à remplacer, par récurrence, k_m par $1 + \max(k_0, \ldots, k_m)$, on peut supposer que la suite $(k_m)_{m\in\mathbb{N}}$ est *strictement croissante*. Considérons la suite extraite $(h_{k_m})_{m\in\mathbb{N}}$. Pour tout $m \in \mathbb{N}$, $\big\|h_{k_{m+1}} - h_{k_m}\big\|_1 \leqslant 2^{-m}$. La série $\sum(h_{k_{m+1}} - h_{k_m})$ converge *normalement* dans $L^2(I)$ vers $S \in L^2(I)$, donc la sous-suite $(h_{k_m})_{m\in\mathbb{N}}$ converge pour la norme $\|.\|_2$ et aussi presque partout vers $S + h_{k_0}$. On vérifie que la suite initiale $(h_k)_{k\in\mathbb{N}}$ converge vers $S + h_{k_0}$ pour la norme $\|.\|_2$.

(iii) L'application $(f, g) \in L^2(I) \mapsto (\,f \mid g\,)$ est évidemment bilinéaire et symétrique et l'on a $(\,f \mid f\,) = \|f\|_2^2$, donc $(\,f \mid f\,) = 0$ si, et seulement si, $f = 0$. ■

Corollaire 175 (Inégalité de Cauchy-Schwarz).
Soit I un intervalle réel. Pour tous $f,\ g \in L^2(I)$, on a $(\,f \mid g\,) \leqslant \|f\|_2 \|g\|_2$.

La preuve du résultat suivant fait l'objet de l'exercice III.1.34.

Proposition 176. Soit $I \subset \mathbb{R}$ un intervalle.
L'espace $\mathcal{C}_c(I)$ des fonctions continues à support compact est *dense* dans $L^2(I)$.

Espaces de Hilbert complexes $L^2(I,\mathbb{C})$ *— Application aux séries de Fourier*

Nous allons compléter les résultats du module *Séries de Fourier* du livre [L2] en utilisant les espaces L^2. Nous utiliserons dans ce paragraphe les définitions et résultats des modules *Espaces préhilbertiens* (*Espaces préhilbertiens complexes, Espaces hermitiens*) et *Espaces vectoriels normés (Espaces de Hilbert)* du livre [L2].
On étend les notions de fonctions mesurables et de fonctions négligeables aux fonctions à valeurs complexes : une fonction $f : I \to \mathbb{C}$ est mesurable (respectivement négligeable) si, et seulement si, sa partie réelle et sa partie imaginaire le sont). On peut alors définir *mutatis mutandis* les espaces de fonctions *sommables* $\tilde{L}^1(I,\mathbb{C})$ (respectivement *de carré sommable* $\tilde{L}^2(I,\mathbb{C})$) sur un intervalle réel I, puis, par passage au quotient par les sous-espaces de fonctions négligeables, les espaces $L^1(I,\mathbb{C})$ (respectivement de carré sommable $L^2(I,\mathbb{C})$). L'espace $L^1(I,\mathbb{C})$ est muni de la norme $f \mapsto \displaystyle\int_I |f|$ et l'espace $L^2(I,\mathbb{C})$ est muni de la norme $f \mapsto \left(\displaystyle\int_I |f|^2\right)^{1/2}$. En utilisant les théorèmes 171 et 174, on montre facilement que les espaces $L^1(I,\mathbb{C})$ et $L^2(I,\mathbb{C})$ sont *complets*.
Si $f,\ g \in L^2(I,\mathbb{C})$, on vérifie que $fg \in L^1(I,\mathbb{C})$ et que l'application :

$$(f,g) \in L^2(I,\mathbb{C}) \times L^2(I,\mathbb{C}) \to (\, f \mid g\,) := \int_I \bar{f}g \in \mathbb{C},$$

définit un *produit scalaire* sur $L^2(I)$ tel que $(\, f \mid f\,) = \|f\|_2^2$.
Soit $T > 0$. On considère le $\mathbb{C}$-espace vectoriel des fonctions $f : \mathbb{R} \to \mathbb{C}$ périodiques de période T noté $\mathcal{F}_T$. On identifie une fonction $f \in \mathcal{F}_T$ à sa restriction $\tilde{f} := f_{|_{[0,T]}} : [0,T] \to \mathbb{C}$, vérifiant $\tilde{f}(0) = \tilde{f}(T)$. Ainsi $\mathcal{F}_T$ est identifié au sous-espace de $\mathcal{F}([0,T]\,,\mathbb{C})$ formé des fonctions prenant la même valeur en 0 et T.
On note $\mathcal{C}_T$ (respectivement $\mathcal{R}_T$) l'espace des fonctions continues (respectivement réglées) T-périodiques sur $\mathbb{R}$ à valeurs complexes. On l'identifie au sous-espace de $\mathcal{C}([0,T]\,,\mathbb{C})$ (respectivement $\mathcal{R}([0,T]\,,\mathbb{C})$) formé des fonctions prenant la même valeur en 0 et T.
On munit l'espace des fonctions de carré-intégrable (au sens de Lebesgue) sur $[0,T]$ du produit scalaire $(f,g) \mapsto (\, f \mid g\,) := \frac{1}{T}\displaystyle\int_0^T \bar{f}g$.
La norme correspondante est $f \mapsto \|f\|_2 := \dfrac{1}{\sqrt{T}}\Big(\displaystyle\int_0^T \bar{f}f\Big)^{1/2}$.
On prendra garde à la différence avec la définition antérieure associée à la période T.
L'espace des fonctions de carré-intégrable sur $[0,T]$ muni de la norme $\|.\|_2$ est noté L^2_T. On vérifie que c'est un espace de Hilbert.
On peut identifier $\mathcal{C}_T$ à un sous-espace de L^2_T (si $f \in \mathcal{C}_T$ est telle que $\|f\|_2 = 0$, alors $f = 0$).
Dans le module *Séries de Fourier* du livre [L2] nous avons introduit un espace de fonctions réglées $\underline{\mathcal{R}}_T$ (vérifiant la condition dite de Dirichlet). On a $\underline{\mathcal{R}}_T \subset \mathcal{R}_T$ et une application

linéaire $\mathcal{R}_T \to \underline{\mathcal{R}}_T$, appelée *régularisation*, induisant l'identité sur $\underline{\mathcal{R}}_T$. Les fonctions réglées périodiques sont de carré intégrable sur T. Une fonction réglée et sa régularisée diffèrent par une fonction réglée qui est non nulle au plus sur un ensemble dénombrable, donc négligeable, elles ont donc même image dans $L^2(T)$. L'application naturelle $\underline{\mathcal{R}}_T \to L^2$ est *injective*. On peut ainsi identifier $\underline{\mathcal{R}}_T$ à un sous-espace de $L^2(T)$ contenant l'espace des fonctions continues $\mathcal{C}_T$.

Dans le module *Séries de Fourier* de [L2] nous avons vu que les fonctions $(e_n : \mathbb{R} \mapsto \mathbb{C})_{n\in\mathbb{Z}})$ définies, pour tout $n \in \mathbb{Z}$, par $e_n(t) := e^{2i\pi nt/T}$ forment une *famille orthonormée totale* de $\mathcal{C}_T$. Elles forment donc aussi une famille orthonormée totale de L^2_T, c'est-à-dire une *base hilbertienne* de cet espace de Hilbert, d'après la proposition 86 de la page 537 du livre [L2].

Définition 37. Si $f \in L^2_T$ ses *coefficients de Fourier* sont :

$$\forall n \in \mathbb{Z}, \quad c_n(f) := (\, e_n \mid f\,) = \frac{1}{T}\int_0^T f(t)e^{-2i\pi nt/T}\,\mathrm{d}t.$$

En utilisant l'égalité de Parseval (théorème 86 de la page 537 du livre [L2] et son corollaire) on obtient le résultat suivant.

Théorème 177.

(i) La famille $\{e_n : \mathbb{R} \mapsto \mathbb{C}\}_{n\in\mathbb{Z}}$ est une *base hilbertienne* de l'espace de Hilbert L^2_T.

(ii) Pour tout $f \in L^2_T$, la « suite » $\big(c_n(f)\big)_{n\in\mathbb{Z}}$ appartient à $\ell^2_{\mathbb{Z}}$ et l'on a :

$$\|f\|_2^2 = \frac{1}{T}\int_0^T |f|^2 = \sum_{n\in\mathbb{Z}} \big|c_n(f)\big|^2 \qquad \textit{(égalité de Parseval)}$$

(iii) Pour tous $f,\ g \in L^2_T$: $(\, f \mid g\,) = \dfrac{1}{T}\displaystyle\int_0^T \bar{f}g = \sum_{n\in\mathbb{Z}} \overline{c_n(f)}\, c_n(g)$.

(iv) La série $\sum\limits_{n\in\mathbb{Z}} c_n(f)e_n$ converge vers f dans L^2_T.

(v) L'application $f \in L^2_T \mapsto \big(c_n(f)\big)_{n\in\mathbb{Z}}$ et un *isomorphisme d'espaces de Hilbert* de L^2_T sur $\ell^2_{\mathbb{Z}}$.

Remarque. La famille $\{e_n : \mathbb{R} \mapsto \mathbb{C}\}_{n\in\mathbb{Z}}$ est une *famille totale orthonormée* pour $\underline{\mathcal{R}}_T$. Les assertions (i), (ii), (iii), (iv) du théorème ci-dessus restent vraies si l'on remplace L^2_T par $\underline{\mathcal{R}}_T$. En revanche, l'application $f \mapsto \big(c_n(f)\big)_{n\in\mathbb{Z}}$ de $\underline{\mathcal{R}}_T$ dans $\ell^2_{\mathbb{Z}}$ est *injective* mais elle *n'est pas surjective* (*cf.* l'exercice III.1.39).

3.3.3 Intégrabilité d'une fonction et intégrabilité de sa valeur absolue

Les relations entre l'intégrabilité d'une fonction et celle de sa valeur absolue dépendent de la théorie de l'intégration considérée (au sens de Riemann, Riemann généralisé, Henstock-Kurzweil, Lebesgue). Pour aider le lecteur « à s'y retrouver » nous terminons par une liste des diverses possibilités que nous avons rencontrées dans ce module et antérieurement dans le module *Intégration* du livre [L2]. Nous laissons le lecteur chercher les références à titre d'exercice « de révision ».

1. Si f est intégrable au sens de Riemann sur un segment, $|f|$ est intégrable au sens de Riemann sur ce segment.
2. Il est possible que $|f|$ soit intégrable au sens de Riemann sur un segment sans que f le soit.
3. Soit f localement intégrable au sens de Riemann sur un intervalle $I \subset \mathbb{R}$. Si $|f|$ est intégrable au sens de Riemann généralisé sur I, f l'est aussi. Par contre f peut être intégrable au sens de Riemann généralisé sur I sans que $|f|$ le soit.
4. Il est possible que f soit intégrable au sens de Henstock-Kurzweil sur un segment sans que $|f|$ le soit.
5. Si f est intégrable au sens de Henstock-Kurzweil sur un segment I et si elle est *bornée*, alors $|f|$ est intégrable au sens de Henstock-Kurzweil sur I.
6. Il est possible que $|f|$ soit intégrable au sens de Henstock-Kurzweil sur un segment sans que f le soit.
7. Si f est intégrable au sens de Lebesgue sur un intervalle, il en est de même de $|f|$.
8. Il est possible que $|f|$ soit intégrable au sens de Lebesgue sur un intervalle sans que f le soit.
9. Si $|f|$ est intégrable au sens de Lebesgue sur un intervalle I et si f est mesurable, alors f est intégrable au sens de Lebesgue sur I.

4 Intégrales multiples au sens de Henstock-Kurzweil

Pour terminer ce module nous allons donner une introduction à la notion d'intégrale multiple *dans le cadre de la théorie de Henstock-Kurzweil*. Cela complétera la section 1. On verra que, dans ce nouveau cadre, on obtient non seulement des preuves mais aussi des versions plus satisfaisantes du théorème de Fubini et de la formule de changement de variables que nous avions admis dans la section 1.

4.1 Définition et premières propriétés

On reprend les définitions et notations de la partie 1.1.1. Par abus de langage, si $\mathcal{D}$ est une subdivision pointée, on notera $(\mathbf{J}, \xi) \in \mathcal{D}$ si $(\mathbf{J}, \xi)$ est l'un des pavés pointés de la subdivision.

On rappelle en particulier que la somme de Riemann associée à une subdivision pointée $\mathcal{D} := ((\mathbf{P}_k, \xi_k))_{k\in[\![1,N]\!]}$ de $\mathbf{P}$ est :

$$S_{\mathcal{D}}(f) := \sum_{k\in[\![1,N]\!]} \mathrm{vol}\,(\mathbf{P}_k)\, f(\xi_k).$$

On notera aussi $S_{\mathcal{D}}(f) = \sum_{(\mathbf{J},\xi)\in\mathcal{D}} \mathrm{vol}(\mathbf{J}) f(\xi)$.

Définition 38. Une *jauge* δ sur un pavé $\mathbf{P} \subset \mathbb{R}^n$ est une fonction *strictement positive* sur $\mathbf{P}$. Une subdivision pointée $\mathcal{D}$ de $\mathbf{P}$ est δ*-fine* si pour tout $(\mathbf{J}, \xi) \in \mathcal{D}$, on a $\mathbf{J} \subset \mathbf{P}\big(\xi, \delta(\xi)\big)$.

On prendra garde à la « petite » variante par rapport à la définition de subdivision δ-fine donnée en dimension 1.

Proposition 178 (Lemme de Cousin). Pour tout pavé $\mathbf{P} \subset \mathbb{R}^n$ et toute jauge δ sur $\mathbf{P}$, il existe une subdivision pointée $\mathcal{D}$ de $\mathbf{P}$ qui est δ-fine.

Démonstration. On procède comme dans la proposition 70 de la page 249. La différence est qu'au lieu de diviser chaque intervalle en deux intervalles deux fois plus petits à chaque étape, on divise chaque pavé en 2^n pavés de côtés deux fois plus petits à chaque étape. On utilise alors le théorème des compacts emboîtés au lieu du théorème des segments emboîtés. ∎

Définition 39. Une fonction $f : \mathbf{P} \to \mathbb{R}$ est $\mathscr{HK}$*-intégrable sur un pavé* $\mathbf{P} \subset \mathbb{R}^n$ s'il existe un réel A tel que pour tout $\varepsilon > 0$, on puisse trouver une fonction jauge $\delta : \mathbf{P} \to \mathbb{R}$ en sorte que si $\mathcal{D}$ est une subdivision δ-fine de $\mathbf{P}$, alors $\big|S_{\mathcal{D}}(f) - A\big| \leqslant \varepsilon$. On appelle A l'intégrale de f sur $\mathbf{P}$, on note $A = \displaystyle\int_{\mathbf{P}} f$ et l'on dit que δ est une jauge ε-adaptée à f sur $\mathbf{P}$.

Comme dans le cas de l'intégrale de Henstock-Kurzweil d'une fonction sur un segment ou dans celui de l'intégrale de Riemann sur un pavé, on vérifie facilement que la propriété de linéarité de l'intégrale et le théorème de comparaison sont satisfaits.

La $\mathscr{R}$-intégrabilité correspond au cas des jauges *constantes* (*cf.* la proposition 19 de la page 187) on a donc le résultat suivant.

Proposition 179. Si une fonction $f : \mathbf{P} \to \mathbb{R}$ est $\mathscr{R}$-intégrable sur un pavé $\mathbf{P} \subset \mathbb{R}^n$, elle est $\mathscr{HK}$-intégrable sur ce pavé et les intégrales sont les mêmes.

En utilisant la proposition 8 de la page 182, on obtient le résultat suivant.

Théorème 180.
Toute fonction $f : \mathbf{P} \to \mathbb{R}$ continue sur un pavé $\mathbf{P} \subset \mathbb{R}^n$ est $\mathscr{HK}$-intégrable.

On dispose du critère de Cauchy pour vérifier la $\mathscr{HK}$-intégrabilité des fonctions de plusieurs variables.

Théorème 181 (Critère de Cauchy).
Une fonction $f : \mathbf{P} \to \mathbb{R}$ est $\mathscr{HK}$-intégrable sur un pavé $\mathbf{P} \subset \mathbb{R}^n$ si, et seulement si, pour tout $\varepsilon > 0$, il existe une jauge δ sur $\mathbf{P}$ telle que pour toutes subdivisions $\mathcal{D}$ et $\mathcal{D}'$ δ-fines de $\mathbf{P}$ on ait : $\left|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\right| \leqslant \varepsilon$.
Une telle jauge est ε-adaptée à f sur $\mathbf{P}$.

La démonstration est en tous points identique à celle du théorème 78 de la page 265 où il suffit de remplacer $[a, b]$ par $\mathbf{P}$.
Enfin, la relation de Chasles en dimension $n = 1$ se généralise de la manière suivante.

Théorème 182 (Relation de Chasles). Soit $\mathbf{P}$ un pavé de $\mathbb{R}^n$ et $\{\mathbf{P}_j\}_{j\in[\![1,N]\!]}$ une famille de pavés d'intérieurs deux à deux disjoints tels que $\mathbf{P} = \bigcup_{j\in[\![1,N]\!]} \mathbf{P}_j$.
Une fonction $f : \mathbf{P} \to \mathbb{R}$ est $\mathscr{HK}$-intégrable sur $\mathbf{P}$ si, et seulement si, elle est $\mathscr{HK}$-intégrable sur chaque pavé $\mathbf{P}_j$. Dans ce cas $\displaystyle\int_{\mathbf{P}} f = \sum_{j\in[\![1,N]\!]} \int_{\mathbf{P}_j} f$.

Démonstration. Commençons par montrer que si f est $\mathscr{HK}$-intégrable sur $\mathbf{P}$, alors la restriction de f à n'importe quel pavé $\mathbf{P}_j$ est $\mathscr{HK}$-intégrable. Fixons $j_0 \in [\![1, N]\!]$. Étant donné $\varepsilon > 0$, considérons une jauge δ sur $\mathbf{P}$ qui est ε-adaptée à f. Soient Δ et Δ' des subdivisions de $\mathbf{P}_{j_0}$ qui sont δ-fines. En choisissant des subdivisions δ-fines des $\mathbf{P}_j$ pour $j \neq j_0$, on peut compléter Δ et Δ' en des subdivisions $\mathcal{D}$ et $\mathcal{D}'$ de $\mathbf{P}$ qui sont δ-fines et qui coïncident en dehors de $\mathbf{P}_{j_0}$. Alors :

$$\left|S_{\Delta}(f) - S_{\Delta'}(f)\right| = \left|S_{\mathcal{D}}(f) - S_{\mathcal{D}'}(f)\right| \leqslant \varepsilon.$$

Le critère de Cauchy est satisfait, f est $\mathscr{HK}$-intégrable sur $\mathbf{P}_{j_0}$ et la restriction de δ à $\mathbf{P}_{j_0}$ est une jauge ε-adaptée à f sur $\mathbf{P}_{j_0}$.

Supposons maintenant que la restriction de f à tous les pavés $\mathbf{P}_j$ soit $\mathscr{HK}$-intégrable. Il faut montrer que f est $\mathscr{HK}$-intégrable sur $\mathbf{P}$. Il y a une difficulté similaire à celle que nous avons rencontrée en dimension 1. Si, pour une jauge δ, on considère une subdivision δ-fine $\mathcal{D} := \{(\mathbf{J}, \xi)\}$, il n'est pas en général possible de la « découper » en subdivisions pointées des $\mathbf{P}_j$. Pour pouvoir le faire, nous allons utiliser la même idée qu'en dimension $n = 1$ (*cf.* la preuve du théorème 76 de la page 264). On suppose $n \geqslant 2$ (pour $n = 1$ le résultat est connu). Le point de départ est *une version multidimensionnelle du principe droite-gauche* (*cf.* le *Point méthode* de la page 248). Soit $(\mathbf{J}, \xi)$ un pavé pointé, on suppose que ξ *appartient à tous les pavés* $\mathbf{P}_i$ *qui rencontrent* $\mathbf{J}$. On peut alors considérer la famille (non vide) $A_{\mathbf{J}} \subset [\![1, N]\!]$ des indices j tels que $\mathbf{J}$ intersecte *l'intérieur* de $\mathbf{P}_j$ et noter, pour tout $j \in A_{\mathbf{J}}$, $\mathbf{J}_j := \mathbf{J} \cap \mathbf{P}_j$. Les pavés pointés $\{(\mathbf{J}_j, \xi)\}_{j\in A_{\mathbf{J}}}$ forment une *subdivision pointée* de J. On a, pour toute fonction $f : \mathbf{P} \to \mathbb{R}$:

$$f(\xi) \operatorname{vol} \mathbf{J} = f(\xi) \sum_{j\in A_{\mathbf{J}}} \operatorname{vol} \mathbf{J}_j = \sum_{j\in A_{\mathbf{J}}} f(\xi) \operatorname{vol} \mathbf{J}_j,$$

Par suite, si $(\mathbf{J},\xi)$ est un pavé pointé d'une subdivision $\mathcal{D}$, on peut, dans celle-ci, le remplacer par la famille $\{(\mathbf{J}_j,\xi)\}_{j\in A_{\mathbf{J}}}$ *sans changer la somme de Riemann* $S_{\mathcal{D}}(f)$.
Pour pouvoir mettre en œuvre la méthode décrite ci-dessus, nous allons définir une jauge convenable $\delta_{\mathcal{P}}$ « forçant » les points de subdivision à « bien se comporter » (une « variante » de la jauge $\delta_{\mathbf{c}}$ du lemme 71 de la page 251, où les points c_j sont remplacés par les frontières $\partial\mathbf{P}_j$).
Notons $\mathcal{P}:\{P_j\}_{j\in[\![1,N]\!]}$. On définit une jauge $\delta_{\mathcal{P}}$ sur P par :

$$\delta_{\mathcal{P}}:\mathbf{x}\mapsto\begin{cases}\frac{1}{2}\displaystyle\min_{\{i\in[\![1,N]\!]\,|\,\mathbf{x}\notin\partial\mathbf{P}_i\}} d(\mathbf{x},\partial\mathbf{P}_i) & \text{si}\quad x\notin\displaystyle\bigcap_{j\in[\![1,N]\!]}\partial\mathbf{P}_j\\ 1 & \text{si}\quad x\in\displaystyle\bigcap_{j\in[\![1,N]\!]}\partial\mathbf{P}_j.\end{cases}$$

Cette jauge, et (par suite) toute jauge majorée par elle, vérifient les conditions suivantes.

- Si $\mathbf{x}$ appartient à l'intérieur du pavé $\mathbf{P}_j$, alors le pavé fermé $\mathbf{P}\big(\mathbf{x},\delta(\mathbf{x})\big)$ est contenu dans $\operatorname{Int}\mathbf{P}_j$. En effet $\delta(\mathbf{x})\big)\leqslant\frac{1}{2}d(\mathbf{x},\partial\mathbf{P}_j)<d(\mathbf{x},\partial\mathbf{P}_j)$.
- Si $\mathbf{x}\notin\mathbf{P}_j$, alors le pavé fermé $\mathbf{P}\big(\mathbf{x},\delta(\mathbf{x})\big)$, de centre $\mathbf{x}$ et de rayon $\delta(\mathbf{x})$, ne rencontre pas $\mathbf{P}_j$.
  En effet $d(\mathbf{x},\partial\mathbf{P}_j)>0$ et, d'après la définition de la jauge, $\mathbf{P}\big(\mathbf{x},\delta(\mathbf{x})\big)\cap\partial\mathbf{P}_j=\varnothing$, donc $\mathbf{P}\big(\mathbf{x},\delta(\mathbf{x})\big)\cap\mathbf{P}_j=\varnothing$ (si $\mathbf{y}\in\operatorname{Int}\mathbf{P}_j$, le segment $[\mathbf{x},\mathbf{y}]$ rencontre $\partial\mathbf{P}_j$).

Soient $\mathcal{D}$ une subdivision $\delta_{\mathcal{P}}$-fine de $\mathbf{P}$ et $(\mathbf{J},\xi)\in\mathcal{D}$. On a alors *l'alternative* suivante.

- Soit le point marqué ξ appartient à *l'intérieur* de l'un des pavés $\mathbf{P}_j$, noté $\mathbf{P}_{j_0}$, alors j_0 est déterminé de manière unique, $\mathbf{J}\subset\operatorname{Int}\mathbf{P}_{j_0}$ et, si $j\neq j_0$, $\mathbf{P}_j$ ne rencontre pas $\mathbf{J}$.
- Soit le point marqué ξ appartient à $\displaystyle\bigcup_{j\in[\![1,N]\!]}\partial\mathbf{P}_j$. Si pour un indice $k\in[\![1,N]\!]$, $\xi\notin\partial\mathbf{P}_k$, alors $\xi\notin\mathbf{P}_k$ et $\mathbf{P}_k$ ne rencontre pas $\mathbf{J}$. Par suite ξ *appartient à tous les pavés* $\mathbf{P}_j$ *qui rencontrent* $\mathbf{J}$.

Ainsi, dans tous les cas, on peut appliquer la généralisation du principe droite-gauche décrite ci-dessus et remplacer *chaque* pavé pointé $(\mathbf{J},\xi)$ de la subdivision $\mathcal{D}$ par une réunion de pavés marqués $(\mathbf{J}_j=\mathbf{J}\cap\mathbf{P}_j,\xi)$ (prise sur les indices $j\in A_{\mathbf{J}}$) *sans changer la somme de Riemann* $S_{\mathcal{D}}(f)$. On obtient alors une nouvelle subdivision pointée $\delta_{\mathcal{P}}$-fine $\mathcal{D}'$ telle que $S_{\mathcal{D}'}(f)=S_{\mathcal{D}}(f)$. Par ailleurs, pour $j\in[\![1,N]\!]$ *fixé*, on peut considérer *tous* les pavés $\mathbf{J}_j=\mathbf{J}\cap\mathbf{P}_j$ ($(\mathbf{J},\xi)\in\mathcal{D}$) d'intérieur non vide. On vérifie que les $(\mathbf{J}_j,\xi)$ correspondants forment une subdivision pointée $\mathcal{D}_j$ de $\mathbf{P}_j$, puis que $\mathcal{D}'=\displaystyle\bigcup_{j=1}^{N}\mathcal{D}_j$. Par suite :

$$S_{\mathcal{D}}(f)=S_{\mathcal{D}'}(f)=\sum_{j\in[\![1,N]\!]}S_{\mathcal{D}_j}(f_{|\mathbf{P}_j}).$$

Soit maintenant $\varepsilon>0$, on choisit, pour tout $j\in[\![1,N]\!]$, une jauge δ_j ε/N-adaptée à f sur $\mathbf{P}_j$. On définit ensuite une jauge δ' sur P par :

$$\delta':\mathbf{x}\mapsto\min_{\{i\in[\![1,N]\!]|\mathbf{x}\in P_i\}}\delta_i(\mathbf{x}),$$

puis on définit une jauge $\delta:=\min(\delta',\delta_{\mathcal{P}})$.

Soit $\mathcal{D}$ une subdivision δ-fine. Puisque $\delta \leqslant \delta_{\mathcal{P}}$, on peut, en utilisant les remarques faites ci-dessus, trouver une subdivision δ-fine $\mathcal{D}'$ de $\mathbf{P}$ et, pour tout $j \in [\![1, N]\!]$ une subdivision δ-fine $\mathcal{D}_j$ de $\mathbf{P}_j$, telles que $S_{\mathcal{D}'}(f) = S_{\mathcal{D}}(f)$ et $S_{\mathcal{D}'}(f) = \sum_{j\in[\![1,N]\!]} S_{\mathcal{D}_j}(f_{|\mathbf{P_j}})$. Alors (en utilisant $\delta \leqslant \delta' \leqslant \delta_j$, pour tout $j \in [\![1, N]\!]$) :

$$\left| S_{\mathcal{D}}(f) - \sum_{j\in[\![1,N]\!]} \int_{\mathbf{P}_j} f \right| = \left| \sum_{j\in[\![1,N]\!]} S_{\mathcal{D}_j}(f) - \sum_{j\in[\![1,N]\!]} \int_{\mathbf{P}_j} f \right|$$
$$\leqslant \sum_{j\in[\![1,N]\!]} \left| S_{\mathcal{D}_j}(f) - \int_{\mathbf{P}_j} f \right| \leqslant \sum_{j\in[\![1,N]\!]} \frac{\varepsilon}{N} = \varepsilon.$$

4.2 Théorème de Fubini

Nous avons vu dans la partie 1.3 que pour calculer une intégrale multiple (au sens de Riemann) on dispose, sous-certaines hypothèses, du théorème de Fubini.
Nous allons étendre le résultat au cadre de l'intégrale de Henstock-Kurzweil. La preuve du théorème suivant est un peu délicate. Nous la ferons plus loin (*cf.* page 346).

Théorème 183 (Théorème de Fubini). Soient $\mathbf{P}_1$ un pavé de $\mathbb{R}^n$ et $\mathbf{P}_2$ un pavé de $\mathbb{R}^m$. Soit f une fonction $\mathscr{HK}$-intégrable sur $\mathbf{P}_1 \times \mathbf{P}_2$. On suppose que, pour tout $\mathbf{x} \in \mathbf{P}_1$, la fonction $\mathbf{y} \mapsto f(\mathbf{x}, \mathbf{y})$ est $\mathscr{HK}$-intégrable sur $\mathbf{P}_2$, alors la fonction $F : \mathbf{x} \mapsto \int_{\mathbf{P}_2} f(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{y}$ est $\mathscr{HK}$-intégrable sur $\mathbf{P}_1$, et de plus :

$$\int_{\mathbf{P}_1\times\mathbf{P}_2} f = \int_{\mathbf{P}_1} F = \int_{\mathbf{P}_1} \left(\int_{\mathbf{P}_2} f(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x}. \tag{36}$$

Remarque. Si f est $\mathscr{HK}$*-absolument intégrable* sur $\mathbf{P}_1 \times \mathbf{P}_2$, l'hypothèse d'intégrabilité des fonctions $\mathbf{y} \mapsto f(\mathbf{x}, \mathbf{y})$ n'est pas nécessaire, on peut montrer qu'elle est automatiquement satisfaite *presque partout* et que la formule (36) reste vraie (*cf.* par exemple [MPA] ou [Ru]).

Corollaire 184. Soient $\mathbf{P}_1$ un pavé de $\mathbb{R}^n$, $\mathbf{P}_2$ un pavé de $\mathbb{R}^m$ et f une fonction $\mathscr{HK}$-intégrable sur $\mathbf{P}_1 \times \mathbf{P}_2$. Supposons que pour tout $\mathbf{x} \in \mathbf{P}_1$, la fonction $\mathbf{y} \mapsto f(\mathbf{x}, \mathbf{y})$ soit $\mathscr{HK}$-intégrable et que pour tout $\mathbf{y} \in \mathbf{P}_2$, la fonction $\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{y})$ soit $\mathscr{HK}$-intégrable. Alors les fonctions $\mathbf{x} \mapsto \int_{\mathbf{P}_2} f(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{y}$ et $\mathbf{y} \mapsto \int_{\mathbf{P}_1} f(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{x}$ sont $\mathscr{HK}$-intégrables et :

$$\int_{\mathbf{P}_1} \left(\int_{\mathbf{P}_2} f(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x} = \int_{\mathbf{P}_2} \left(\int_{\mathbf{P}_1} f(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{x} \right) \mathrm{d}\mathbf{y}.$$

Rappelons que les fonctions continues sur des pavés y sont $\mathscr{HK}$ -intégrables. On a donc le cas particulier suivant.

Corollaire 185. Soient $\mathbf{P}_1$ un pavé de $\mathbb{R}^n$, $\mathbf{P}_2$ un pavé de $\mathbb{R}^m$ et f une fonction *continue* sur $\mathbf{P}_1 \times \mathbf{P}_2$. Alors les fonctions $\mathbf{x} \mapsto \displaystyle\int_{\mathbf{P}_2} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}$ et $\mathbf{y} \mapsto \displaystyle\int_{\mathbf{P}_1} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}$ sont $\mathscr{HK}$ -intégrables et :

$$\int_{\mathbf{P}_1}\left(\int_{\mathbf{P}_2} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}\right)\mathrm{d}\mathbf{x} = \int_{\mathbf{P}_2}\left(\int_{\mathbf{P}_1} f(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{x}\right)\mathrm{d}\mathbf{y}.$$

Démonstration.

Pour tout $\mathbf{x} \in \mathbf{P}_1$, la fonction $\mathbf{y} \mapsto f(\mathbf{x},\mathbf{y})$ est continue, donc $\mathscr{HK}$ -intégrable et, pour tout $\mathbf{y} \in \mathbf{P}_2$, la fonction $\mathbf{x} \mapsto f(\mathbf{x},\mathbf{y})$ est continue, donc $\mathscr{HK}$ -intégrable. ■

Démonstration du théorème de Fubini La fonction F a été définie dans l'énoncé du théorème 183. On peut supposer que $\mathbf{P}_1$ et $\mathbf{P}_2$ sont d'intérieurs non vide. Posons $A := \displaystyle\int_{\mathbf{P}_1\times\mathbf{P}_2} f$, fixons $\varepsilon > 0$ et considérons une jauge δ qui soit $\varepsilon/2$-adaptée à f sur $\mathbf{P}_1 \times \mathbf{P}_2$.

Pour chaque $\mathbf{x} \in \mathbf{P}_1$, posons $f_{\mathbf{x}} := f(\mathbf{x},\cdot)$ et $\delta_{\mathbf{x}} := \delta(\mathbf{x},\cdot)$, et choisissons $\mathcal{D}_{\mathbf{x}}$ une subdivision $\delta_{\mathbf{x}}$-fine de $\mathbf{P}_2$, telle que :

$$\left|S_{\mathcal{D}_{\mathbf{x}}}(f_{\mathbf{x}}) - F(\mathbf{x})\right| = \left|S_{\mathcal{D}_{\mathbf{x}}}(f_{\mathbf{x}}) - \int_{\mathbf{P}_2} f_{\mathbf{x}}\right| \leqslant \frac{\varepsilon}{2\,\mathrm{vol}(\mathbf{P}_1)}.$$

Cela est possible car, par hypothèse, $f_{\mathbf{x}}$ est intégrable pour tout $\mathbf{x} \in \mathbf{P}_1$.

Soit δ_1 la jauge sur $\mathbf{P}_1$ définie par $\delta_1(\mathbf{x}) := \min\limits_{(\mathbf{J},\eta)\in\mathcal{D}_{\mathbf{x}}} \delta(\mathbf{x},\eta)$. Nous allons montrer que si $\mathcal{D}$ est une subdivision partielle δ_1-fine de $\mathbf{P}_1$, on a $\left|S_{\mathcal{D}}(F) - A\right| \leqslant \varepsilon$. Cela complétera la démonstration : F est $\mathscr{HK}$ -intégrable sur $\mathbf{P}_1$, d'intégrale égale à A, et δ_1 est une jauge ε-adaptée à F sur $\mathbf{P}_1$.

Commençons par définir :

$$\mathbf{D} := \left\{\big(\mathbf{J}_1 \times \mathbf{J}_2, (\xi,\eta)\big) \;\middle|\; (\mathbf{J}_1,\xi) \in \mathcal{D} \text{ et } (\mathbf{J}_2,\eta) \in \mathcal{D}_\xi\right\}.$$

Alors, $\mathbf{D}$ est une subdivision partielle δ-fine de $\mathbf{P}_1 \times \mathbf{P}_2$.

Notons que :

$$S_{\mathbf{D}}(f) = \sum_{(\mathbf{J}_1,\xi)\in\mathcal{D}}\ \sum_{(\mathbf{J}_2,\eta)\in\mathcal{D}_\xi} \mathrm{vol}(\mathbf{J}_1 \times \mathbf{J}_2)\cdot f(\xi,\eta) = \sum_{(\mathbf{J}_1,\xi)\in\mathcal{D}} \mathrm{vol}(\mathbf{J_1})\cdot S_{\mathcal{D}_\xi}(f_\xi).$$

Par conséquent :

$$\begin{aligned}\left|S_{\mathcal{D}}(F) - A\right| &\leqslant \left|S_{\mathbf{D}}(f) - A\right| + \left|S_{\mathcal{D}}(F) - S_{\mathbf{D}}(f)\right| \\ &\leqslant \frac{\varepsilon}{2} + \sum_{(\mathbf{J}_1,\xi)\in\mathcal{D}} \mathrm{vol}(\mathbf{J}_1)\cdot\left|F(\xi) - S_{\mathcal{D}_\xi}(f_\xi)\right| \\ &\leqslant \frac{\varepsilon}{2} + \sum_{(\mathbf{J}_1,\xi)\in\mathcal{D}} \mathrm{vol}(\mathbf{J}_1)\cdot\frac{\varepsilon}{2\,\mathrm{vol}(\mathbf{P}_1)} \leqslant \varepsilon.\end{aligned}$$

4.3 Intégrales sur un ouvert de $\mathbb{R}^n$

Tandis qu'à une variable, il est tout à fait naturel de se contenter d'étudier l'intégrale d'une fonction sur un intervalle (fermé ou non, borné ou non), dans $\mathbb{R}^n$ ($n \geqslant 2$), il n'est pas raisonnable de restreindre l'étude aux pavés. Poursuivant la démarche entreprise dans la section 1, dans le cadre de l'intégrale de Riemann, nous nous proposons maintenant d'étudier l'intégration d'une fonction $f : U \subset \mathbb{R}^n \to \mathbb{R}$ où U est un *ouvert quelconque* de $\mathbb{R}^n$.
Pour étudier l'intégration sur un ouvert quelconque, nous limiterons notre étude aux fonctions *absolument intégrables*, celles pour lesquelles f et $|f|$ sont intégrables (on peut montrer que, comme en dimension 1, ce cadre coïncide avec celui de la version classique de l'intégrale de Lebesgue, *cf.*, par exemple, [De]).
Nous utiliserons la notion de *partie pavable* et ses propriétés (*cf.* la partie 1.2.1).

Observons d'abord que l'on peut étendre la définition de l'intégrale de Henstock-Kurzweil $\displaystyle\int_K f$ au cas où $K := \displaystyle\bigcup_{j=1}^{N} \mathbf{P}_j$ est une réunion finie de pavés fermés $\mathbf{P}_j$ d'intérieurs deux à deux disjoints (c'est-à-dire une partie pavable) :

$$\int_K f = \sum_{j=1}^{N} \left(\int_{\mathbf{P}_j} f \right).$$

Grâce à la relation de Chasles, la valeur de l'intégrale ne dépend pas du choix de la décomposition de K comme réunion de pavés d'intérieurs deux à deux disjoints.

En effet, si $\displaystyle\bigcup_{j=1}^{N} \mathbf{P}_j = \bigcup_{\ell=1}^{N'} \mathbf{J}_\ell$, on a :

$$\sum_{j=1}^{N} \left(\int_{\mathbf{P}_j} f \right) = \sum_{\left\{(j,\ell) \middle| \mathring{\mathbf{P}}_j \cap \mathring{\mathbf{J}}_\ell \neq \varnothing\right\}} \int_{\mathbf{P}_j \cap \mathbf{J}_\ell} f = \sum_{\ell=1}^{N'} \left(\int_{\mathbf{J}_\ell} f \right).$$

Définition 40. On appelle *pavage exhaustif* d'un ouvert $U \subset \mathbb{R}^n$ une suite $(K_p)_{p\in\mathbb{N}}$ de parties qui sont des réunions finies de pavés fermés d'intérieurs disjoints (c'est-à-dire des parties pavables), de sorte que :

$$K_0 \subset K_1 \subset \cdots \subset K_p \subset \cdots \subset U \quad \text{et} \quad U = \bigcup_{p\in\mathbb{N}} \operatorname{Int} K_p.$$

Notons que, si la condition de cette définition est satisfaite, alors

$$U = \bigcup_{p\in\mathbb{N}} \operatorname{Int} K_p = \bigcup_{p\in\mathbb{N}} K_p.$$

Proposition 186. Soit $U \subset \mathbb{R}^n$ un ouvert, alors il existe un *pavage exhaustif* de U.

Démonstration. On peut par exemple considérer, pour chaque entier $p \geqslant 0$, les cubes fermés de côté $1/2^p$: $C_{p,\mathbf{s}} := \displaystyle\prod_{i=1}^{N} \left[\frac{s_i}{2^p}, \frac{s_i+1}{2^p}\right]$ avec $\mathbf{s} = (s_1, \ldots, s_n) \in \mathbb{Z}^n$.

On définit alors K_p comme la réunion de tous les cubes $C_{p,\mathbf{s}}$ contenus dans $U \cap [-2^p, 2^p]^n$. Montrons que la suite $(K_p)_{p\in\mathbb{N}}$ est un pavage exhaustif.

Soit $\mathbf{x} \in U$. Il existe $\varepsilon > 0$ tel que le pavé fermé $\mathbf{P_x}$ de centre $\mathbf{x}$ et de rayon ε soit contenu dans l'ouvert U. Soit $p \in \mathbb{N}$ tel que $1/2^p \leqslant \varepsilon$ et que $\mathbf{P_x} \subset [-2^p, 2^p]^n$. Alors, si un pavé fermé $\mathbf{Q}$ est l'un des $C_{p,\mathbf{s}}$ et si $\mathbf{x} \in \mathbf{Q}$, $\mathbf{Q} \subset \mathbf{P_x} \subset U \cap [-2^p, 2^p]^n$, donc $Q \subset K_p$. Par suite la réunion des pavés de $C_{p,\mathbf{s}}$ contenant $\mathbf{x}$ est contenue dans K_p. On vérifie (en utilisant les projections sur les axes) que $\mathbf{x}$ est *intérieur* à cette réunion. Il en résulte que $\mathbf{x}$ est un *point intérieur* de K_p. Ainsi $U = \bigcup\limits_{p\in\mathbb{N}} \operatorname{Int} K_p$. ■

Un pavage exhaustif est une suite exhaustive de compacts (*cf.* la définition 13 de la page 225 et l'exercice III.1.15).

Supposons maintenant que f et $|f|$ soient intégrables sur tout pavé fermé $\mathbf{Q} \subset U$ et qu'il existe une constante M telle que, pour toute réunion finie $K \subset U$ de pavés fermés, on ait : $\displaystyle\int_K |f| \leqslant M$. Alors, si (K_p) est un pavage exhaustif de U, la suite $\left(\displaystyle\int_{K_p} f\right)$ est de Cauchy (donc convergente). En effet, la suite $\left(A_p := \displaystyle\int_{K_p} |f|\right)$ est croissante et majorée par M. Elle est donc convergente et de Cauchy : pour tout $\varepsilon > 0$, il existe un entier p_0 tel que pour tout $q \geqslant p \geqslant p_0$, on ait $0 \leqslant A_q - A_p \leqslant \varepsilon$. Comme $K_q = K_p \cup L$ avec L une réunion de pavés d'intérieurs disjoints et $\overset{\circ}{K}_p \cap \overset{\circ}{L} = \varnothing$ (*cf.* la proposition 31 de la page 197), on a :

$$\left|\int_{K_q} f - \int_{K_p} f\right| = \left|\int_L f\right| \leqslant \int_L |f| \leqslant A_q - A_p \leqslant \varepsilon.$$

Notons également que la limite de la suite $\left(\displaystyle\int_{K_p} f\right)$ ne dépend pas du choix de pavage exhaustif. En effet, si (L_q) est un autre pavage exhaustif de U, pour tout $p \geqslant 0$, on peut, en utilisant le fait que les pavages exhaustifs sont des suites exhaustives de compacts, trouver des entiers $p' \geqslant q \geqslant p$ tels que :

$$K_p \subset \operatorname{Int} L_q \subset L_q \subset \operatorname{Int} K_{p'} \subset K_{p'}.$$

On vérifie, en utilisant la proposition 31, que $L_q = K_p \cup M$ et $K_{p'} = L_q \cup N$ avec M et N des réunions finies de pavés fermés, $\overset{\circ}{K}_p \cap \overset{\circ}{M} = \varnothing$ et $\overset{\circ}{L}_q \cap \overset{\circ}{N} = \varnothing$. On a donc :

$$\left|\int_{L_q} f - \int_{K_p} f\right| = \left|\int_M f\right| \leqslant \int_M |f| \leqslant \int_{M\cup N} |f| = \int_{K_{p'}} |f| - \int_{K_p} |f| \underset{p\to+\infty}{\longrightarrow} 0.$$

D'où :

$$\lim_{p\to+\infty} \int_{K_p} f = \lim_{q\to+\infty} \int_{L_q} f.$$

La définition suivante a donc un sens.

Définition 41. Soit U un ouvert de $\mathbb{R}^n$. La fonction $f : U \to \mathbb{R}$ est *absolument intégrable sur* U si f et $|f|$ sont $\mathscr{HK}$-intégrables sur tout pavé fermé $\mathbf{Q} \subset U$ et s'il existe une constante M telle que, pour toute réunion finie $K \subset U$ de pavés fermés, on ait $\int_K |f| \leqslant M$. On définit alors :

$$\int_U f := \lim_{p \to +\infty} \int_{K_p} f$$

avec $(K_p)_{p \in \mathbb{N}}$ un pavage exhaustif de U.

On devrait préciser $\mathscr{HK}$-absolument intégrable (pour ne pas confondre avec $\mathscr{R}$-absolument intégrable). Nous ne le ferons que si c'est utile.

Remarques.

1. Dans le cas de la dimension 1, le théorème de Hake implique que la définition donnée ci-dessus de l'intégrale sur un intervalle ouvert coïncide avec la définition 21 de la page 274.
2. Dans le cas de la dimension quelconque, on vérifie facilement qu'une fonction *continue* sur l'ouvert U est $\mathscr{HK}$-absolument intégrable sur U si, et seulement si, elle est intégrable au sens de Riemann généralisé sur U (*cf.* la proposition-définition 62 de la page 229).
3. Si la fonction $f : U \to \mathbb{R}$ est absolument intégrable sur U, alors elle est absolument intégrable sur tout ouvert $V \subset U$. Si, de plus, $f \geqslant 0$, $\int_V f \leqslant \int_U f$.

On vérifie facilement que la propriété de linéarité de l'intégrale et le théorème de comparaison sont satisfaits. On voit également que si f est absolument intégrable sur U, alors :

$$f^+ := \max(f, 0) = \frac{1}{2}\big(|f| + f\big) \quad \text{et} \quad f^- := \max(-f, 0) = \frac{1}{2}\big(|f| - f\big)$$

le sont. De plus :

$$\int_U f = \int_U f^+ - \int_U f^-.$$

Remarque. Notons qu'une fonction positive $f : U \to [0, +\infty[$ est intégrable sur U si, et seulement si, elle est intégrable sur tout pavé $\mathbf{Q} \subset U$ et si :

$$\sup_K \int_K f < +\infty$$

où le supremum est pris sur toutes les réunions finies $K \subset U$ de pavés fermés (c'est-à-dire sur les parties cubables). L'intégrale de f sur U est alors égale à ce supremum.

Définition 42. Si U est un ouvert de $\mathbb{R}^n$, on dit que le *volume de* U *est fini* si la fonction constante 1 est $\mathscr{HK}$-intégrable sur U. On note alors :

$$\mathrm{vol}(U) := \int_U 1.$$

Sinon, le volume de U est infini.

La fonction 1 est $\mathscr{HK}$ *-intégrable* sur U si, et seulement si, elle est intégrable au sens de Riemann généralisé sur U. On peut donc énoncer de façon équivalente la définition ci-dessus en utilisant l'intégrale de Riemann généralisée.

Exercice 30.
Soit $\mathbf{P} \subset \mathbb{R}^n$ un pavé et soit $U := \operatorname{Int} \mathbf{P}$ son intérieur. Montrer que $\operatorname{vol}(U) = \operatorname{vol}(\mathbf{P})$.
Solution. En effet, $\mathbf{P} = [a_1, b_1] \times \cdots \times [a_n, b_n]$. Pour $k \in [\![1, n]\!]$, notons $\ell_k := b_k - a_k$. Pour $p \geqslant 2$ posons :

$$K_p := \left[a_1 + \frac{\ell_1}{2^p}, b_1 - \frac{\ell_1}{2^p}\right] \times \cdots \times \left[a_n + \frac{\ell_n}{2^p}, b_n - \frac{\ell_n}{2^p}\right].$$

La suite $(K_p)_{p \geqslant 2}$ est un pavage exhaustif de U et :

$$\operatorname{vol}(K_p) = \operatorname{vol}(\mathbf{P}) \cdot \left(1 - \frac{1}{2^{p-1}}\right)^n \underset{p \to +\infty}{\longrightarrow} \operatorname{vol}(\mathbf{P}).$$

Plus généralement, d'après le résultat suivant, si U est cubable (*cf.* la définition 12 de la page 198) l'intégrale $\displaystyle\int_U 1$ au sens $\mathscr{HK}$ et l'intégrale $\displaystyle\int 1_U$ au sens $\mathscr{R}$ sont égales. La définition ci-dessus est donc compatible avec celle utilisée plus haut. Par contre, pour un ouvert borné, la notion de volume introduite ci-dessus est *strictement plus générale* que la notion antérieure : un tel ouvert n'est pas nécessairement cubable (*cf.* l'*attention* de la page 232).

Proposition 187. Soit $U \subset \mathbb{R}^n$ un ouvert contenu dans un pavé $\mathbf{P} \subset \mathbb{R}^n$. Alors, le volume de U est fini, 1_U est $\mathscr{HK}$ -intégrable sur $\mathbf{P}$ et :

$$\operatorname{vol}(U) := \int_U 1 = \int_{\mathbf{P}} 1_U = \lambda^-(U).$$

Pour la définition de $\lambda^-(U)$, *cf.* la page 198.
En dimension $n = 1$ la notion de *volume* est remplacée par celle de *longueur*, ainsi $\operatorname{long} U$ est la mesure de Lebesgue de l'ouvert $U \subset \mathbb{R}$ (*cf.* la partie 3). On peut montrer qu'en dimension quelconque $\operatorname{vol}(U)$ est également la mesure de Lebesgue de U définie « classiquement ». Notons aussi que la fonction 1 est intégrable au sens de Riemann généralisé sur U et que l'intégrale coïncide avec l'intégrale au sens $\mathscr{HK}$ (*cf.* l'exercice 12 de la page 232).

Démonstration. Montrons d'abord que la fonction 1_U est intégrable sur U. La fonction 1_U est constante donc intégrable sur tout pavé contenu dans U. Si $K \subset U$ est une réunion finie de pavés, alors $\displaystyle\int_K 1_U \leqslant \operatorname{vol}(\mathbf{P})$. La fonction 1_U est donc intégrable sur U.
On a $\displaystyle\int_U 1_U = \lambda^-(U)$ d'après les définitions.
Montrons maintenant que la fonction 1_U est intégrable sur $\mathbf{P}$ et que son intégrale est égale à $\operatorname{vol}(U)$. Considérons pour cela un pavage exhaustif (K_p) de U et, choisissons, pour

tout $p \in \mathbb{N}$, une jauge δ_p sur $\mathbf{P}$ telle que les pavés fermés de centre $\mathbf{x}$ et de rayon $\delta_p(\mathbf{x})$ satisfassent les conditions suivantes :

- $\mathbf{P}\big(\mathbf{x}, \delta_p(\mathbf{x})\big) \subset U$ si $\mathbf{x} \in U$ et
- $\mathbf{P}\big(\mathbf{x}, \delta_p(\mathbf{x})\big) \subset \mathbb{R}^n \setminus K_p$ si $\mathbf{x} \in \mathbf{P} \setminus U$.

Cela est possible car U et $\mathbb{R}^n \setminus K_p$ sont des ouverts.

Soit $\mathcal{D}$ une subdivision δ_p-fine de $\mathbf{P}$. Pour tout $(\mathbf{J}, \xi) \in \mathcal{D}$, on peut faire les observations suivantes.

- Si $1_U(\xi) = 1$, c'est-à-dire $\xi \in U$, alors $\mathbf{J} \subset U$. Par conséquent, les pavés $\mathbf{J}$ qui contribuent à la somme de Riemann $S_{\mathcal{D}}(f)$ sont tous contenus dans U. D'où :
$$S_{\mathcal{D}}(f) \leqslant \int_U 1_U = \mathrm{vol}(U).$$
- Si $(\mathbf{J}, \xi) \in \mathcal{D}$ et si $\mathbf{J} \cap K_p \neq \varnothing$, alors $\xi \in U$ et $1_U(\xi) = 1$. Donc, les pavés de la subdivision $\mathcal{D}$ qui rencontrent K_p contribuent tous à la somme de Riemann $S_{\mathcal{D}}(f)$. ils forment un recouvrement de K_p, donc : $\displaystyle\int_{K_p} 1_U \leqslant S_{\mathcal{D}}(f) \leqslant \mathrm{vol}(U)$.

Cela termine la démonstration, puisque $\displaystyle\int_{K_p} 1_U \underset{p \to +\infty}{\longrightarrow} \mathrm{vol}(U)$. ■

Exercice 31.

Soit $(A_k)_{k \in \mathbb{N}^*}$ une suite décroissante de bornés de $\mathbb{R}^n$ tels que, pour tout $k \in \mathbb{N}^*$, 1_{A_k} soit $\mathscr{HK}$-intégrable sur $\mathbb{R}^n$. On suppose que la suite $\displaystyle\left(\int_{\mathbb{R}^n} 1_{A_k}\right)_{k \in \mathbb{N}^*}$ tend vers 0 quand n tend vers $+\infty$. Montrer que l'ensemble $\displaystyle\bigcap_{k \in \mathbb{N}^*} A_k$ est négligeable.

Solution. La suite de fonctions $\big(1_{A_k}\big)_{k \in \mathbb{N}^*}$ est décroissante et converge simplement vers $1_{\left(\bigcap_{k \in \mathbb{N}^*} A_k\right)}$. On en déduit que $\displaystyle\int_{\mathbb{R}^n} 1_{\left(\bigcap_{k \in \mathbb{N}^*} A_k\right)} = 0$.

Pour l'exercice suivant, on admet que pour tout ouvert borné $U \subset \mathbb{R}^n$ et tout $\lambda > 0$, on a $\mathrm{vol}(\lambda U) = \lambda^n \,\mathrm{vol}(U)$. Ce résultat sera prouvé plus loin (*cf.* le théorème 189).

Exercice 32.

Soit $C \subset \mathbb{R}^n$ un convexe compact.

(i) On suppose que C est d'intérieur vide. Montrer que C est négligeable.

(ii) On suppose que $0 \in \mathrm{Int}\, C$. Montrer que, pour tout $\varepsilon > 0$, $\mathrm{Int}\, C \subset C \subset (1+\varepsilon)\,\mathrm{Int}\, C$. En déduire que $\partial C \subset (1+\varepsilon)\mathrm{Int}\, C \setminus \mathrm{Int}\, C$.

(iii) Montrer que la frontière de C est négligeable (on utilisera l'exercice précédent).

Solution. Nous utiliserons des résultats de la partie 5 du module *Espaces vectoriels normés* du livre [L2].

(i) Le convexe C est d'intérieur vide si, et seulement s'il engendre un sous-espace affine E de $\mathbb{R}^n$ strictement contenu dans $\mathbb{R}^n$. Tout borné de E est négligeable, donc C est négligeable.

(ii) Soit $\varepsilon > 0$. On a $\operatorname{Int} C \cup \partial C = C$, il suffit donc de montrer que $\partial C \subset (1+\varepsilon)\operatorname{Int} C$. Soit $a \in \partial C$, on sait que $]0, a] \subset \operatorname{Int} C$ (*cf.* le lemme 63 de la page 520 du livre [L2]), donc $\frac{1}{1+\varepsilon} a \in \operatorname{Int} C$ et $a \in (1+\varepsilon)\operatorname{Int} C$.

(iii) Si C est d'intérieur vide, il est négligeable d'après (i). On peut donc supposer $\operatorname{Int} C \neq \varnothing$. Quitte à faire une translation on peut supposer que $0 \in \operatorname{Int} C$. Alors, d'après (ii), pour tout $\varepsilon > 0$, $\operatorname{Int} C \subset (1+\varepsilon)\operatorname{Int} C$ et $\partial C \subset (1+\varepsilon)\operatorname{Int} C \setminus \operatorname{Int} C$.
Pour tout $k \in \mathbb{N}^*$, on note $A_k := (1+\frac{1}{k})\operatorname{Int} C \setminus \operatorname{Int} C$.
On a $1_{A_k} = 1_{(1+\frac{1}{k})\operatorname{Int} C} - 1_{\operatorname{Int} C}$, donc 1_{A_k} est intégrable sur $\mathbb{R}^n$ et l'on a :

$$\int_{\mathbb{R}^n} 1_{A_k} = \left(\left(1+\frac{1}{k}\right)^n - 1\right) \operatorname{vol}(\operatorname{Int} C).$$

La suite $\left(\int_{\mathbb{R}^n} 1_{A_k}\right)_{k\in\mathbb{N}^*}$ tend donc vers 0 quand $n \to +\infty$. Donc, d'après l'exercice précédent $\bigcap_{k\in\mathbb{N}^*} A_k$ est négligeable.
Par ailleurs $\partial C \subset \bigcap_{k\in\mathbb{N}^*} A_k$, par suite ∂C est négligeable.

Proposition 188.
Soit $K \subset \mathbb{R}^n$ un compact contenu dans l'intérieur d'un pavé $\mathbf{P} \subset \mathbb{R}^n$. Alors, 1_K est $\mathscr{HK}$-intégrable sur $\mathbf{P}$ et $\int_{\mathbf{P}} 1_K$ est indépendante du choix de $\mathbf{P}$. On note $\mu(K) = \int 1_K := \int_{\mathbf{P}} 1_K$ et l'on dit que $\mu(K)$ est la *mesure de Lebesgue* de K.

On peut montrer que la mesure de Lebesgue d'un compact ainsi définie coïncide avec la notion « classique » (*cf.* [De] et [Ru]).

Démonstration. Notons $U := \operatorname{Int} \mathbf{P} \setminus K$. C'est un ouvert contenu dans $\mathbf{P}$, donc 1_U est $\mathscr{HK}$-intégrable sur $\mathbf{P}$. On a $\operatorname{vol} \mathbf{P} = \int_{\mathbf{P}} 1_{\mathbf{P}} = \int_{\mathbf{P}} 1_{\operatorname{Int} \mathbf{P}}$ et $1_{\operatorname{Int} \mathbf{P}} = 1_U + 1_K$, donc 1_K est $\mathscr{HK}$-intégrable sur $\mathbf{P}$ et $\int_{\mathbf{P}} 1_K = \operatorname{vol}(\mathbf{P}) - \int_{\mathbf{P}} 1_U$. On montre que le résultat est indépendant du choix de $\mathbf{P}$ en utilisant la formule de Chasles. ■

Remarque.
Soient $\mathbf{P}_1$ un pavé fermé de $\mathbb{R}^n$ et $\mathbf{P}_2$ un pavé fermé de $\mathbb{R}^m$. On note $\mathbf{P} := \mathbf{P}_1 \times \mathbf{P}_2$.
Soient $U \subset \mathbf{P}$ un *ouvert* et $K \subset \mathbf{P}$ un *compact*.

Alors on peut appliquer à 1_U et 1_K le théorème de Fubini (théorème 183). En effet, si l'on note, pour tout $\mathbf{x} \in \mathbf{P}_1$, $U_{\mathbf{x}} := \{\mathbf{y} \in \mathbb{R}^m \mid (\mathbf{x}, \mathbf{y}) \in U\}$ (respectivement $K_{\mathbf{x}} := \{\mathbf{y} \in \mathbb{R}^m \mid (\mathbf{x}, \mathbf{y}) \in K\}$), les fonctions $1_{U_{\mathbf{x}}}$ et $1_{K_{\mathbf{x}}}$ sont $\mathscr{HK}$-intégrables ($U_{\mathbf{x}}$ est un *ouvert* de $\mathbb{R}^m$ et $K_{\mathbf{x}}$ est un *compact* de $\mathbb{R}^m$).

Alors les fonctions $F : \mathbf{x} \mapsto \displaystyle\int_{\mathbf{P}_2} 1_{U_{\mathbf{x}}} \,\mathrm{d}\mathbf{y}$ et $G : \mathbf{x} \mapsto \displaystyle\int_{\mathbf{P}_2} 1_{K_{\mathbf{x}}} \,\mathrm{d}\mathbf{y}$ sont $\mathscr{HK}$-intégrables. De plus :

$$\int 1_U = \int_{\mathbf{P}_1} \left(\int_{\mathbf{P}_2} 1_{U_{\mathbf{x}}} \,\mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x}. \tag{37}$$

$$\int 1_K = \int_{\mathbf{P}_1} \left(\int_{\mathbf{P}_2} 1_{K_{\mathbf{x}}} \,\mathrm{d}\mathbf{y} \right) \mathrm{d}\mathbf{x}. \tag{38}$$

On a évidemment des résultats analogues en échangeant les rôles de $\mathbf{P}_1$ et $\mathbf{P}_2$.

Notons que, même si U est cubable, $U_{\mathbf{x}}$ ne l'est pas nécessairement, on a donc amélioré le résultat obtenu dans le cas de l'intégrale de Riemann.

4.4 Formule de changement de variables

Théorème 189 (Formule de changement de variables). Soient U_1 et U_2 deux ouverts de $\mathbb{R}^n$ et $\mathbf{h} : U_1 \to U_2$ un difféomorphisme de classe $\mathcal{C}^1$. Une fonction *continue* $f : U_2 \to \mathbb{R}^n$ est $\mathscr{HK}$-absolument intégrable sur U_2 si, et seulement si, $f \circ \mathbf{h} \cdot |\operatorname{Jac} \mathbf{h}|$ est $\mathscr{HK}$-absolument intégrable sur U_1. Dans ce cas :

$$\int_{U_2} f(\mathbf{y}) \,\mathrm{d}\mathbf{y} = \int_{U_1} f \circ \mathbf{h}(\mathbf{x}) \cdot \big|\operatorname{Jac} \mathbf{h}(\mathbf{x})\big| \,\mathrm{d}\mathbf{x}.$$

Notons qu'une fonction *continue* sur U est $\mathscr{HK}$-absolument intégrable sur U si, et seulement si, elle est absolument intégrable au sens de Riemann généralisé sur U (*cf.* la remarque 2. de la page 349). On peut donc énoncer, *de façon équivalente*, le théorème ci-dessus dans le cadre de l'intégrale de Riemann généralisée (*cf.* le corollaire 192 ci-dessous). La preuve, par contre, utilise la théorie de Henstock-Kurzweil (*cf.* la preuve de la proposition 190 ci-dessous).

Preuve de la formule de changement de variables

Avant de nous attaquer à la démonstration de la formule de changement de variables, nous allons traiter un cas particulier (qui nous servira dans la démonstration du cas général). La preuve de ce cas particulier utilise la théorie de Henstock-Kurzweil (ou si l'on préfère celle de Lebesgue).

Proposition 190.
Soit $L : \mathbb{R}^n \to \mathbb{R}^n$ une *application linéaire inversible* et $U \subset \mathbb{R}^n$ un *ouvert borné*. Alors :

$$\operatorname{vol}\big(L(U)\big) = \big|\det(L)\big| \cdot \operatorname{vol}(U).$$

Démonstration. Rappelons que toute matrice inversible $A \in M_n(\mathbb{R})$ peut être transformée en la matrice I_n par des opérations élémentaires sur les colonnes (voir le module *Calcul matriciel élémentaire* du livre [L1]). La même suite d'opérations sur les colonnes transforme I_n en A^{-1}.

Il s'en suit que toute matrice inversible peut s'écrire comme produit de matrices d'échange, de transvection et de dilatation.
En termes d'applications linéaires, cela signifie que tout endomorphisme inversible $L : \mathbb{R}^n \to \mathbb{R}^n$ peut s'écrire

$$L = L_m \circ L_{m-1} \circ \cdots \circ L_1$$

avec $L_i : \mathbb{R}^n \to \mathbb{R}^n$ des endomorphismes élémentaires du type :

– *échange :* $\sigma_{i,j} : (x_1, \ldots, x_n) \mapsto (y_1, \ldots, y_n)$ avec :

$$y_k = x_k \quad \text{pour } k \neq i \text{ et } k \neq j, \qquad\qquad y_j = x_i \quad \text{et} \quad y_i = x_j \ ;$$

– *transvection :* $\tau_{i,j}(\alpha) : (x_1, \ldots, x_n) \mapsto (y_1, \ldots, y_n)$ avec :

$$y_k = x_k \quad \text{pour } k \neq i \qquad \text{et} \qquad y_i = x_i + \alpha x_j \ ;$$

– *dilatation :* $\delta_i(\lambda) : (x_1, \ldots, x_n) \mapsto (y_1, \ldots, y_n)$ avec :

$$y_k = x_k \quad \text{pour } k \neq i \qquad \text{et} \qquad y_i = \lambda x_i.$$

Il suffit donc de montrer la proposition pour les endomorphismes élémentaires, le cas général suivant par récurrence sur m.
Le résultat est vrai si L est un échange ou une dilatation. En effet, dans ce cas, l'image d'un pavé $\mathbf{P} \subset \mathbb{R}^n$ est un pavé et l'on vérifie facilement que $\mathrm{vol}\big(L(\mathbf{P})\big) = \big|\det(L)\big| \cdot \mathrm{vol}(\mathbf{P})$. L'image par L d'un pavage exhaustif (K_p) de U est donc un pavage exhaustif (K'_p) de $L(U)$ et :

$$\mathrm{vol}\big(L(U)\big) = \lim_{p \to +\infty} \mathrm{vol}(K'_p) = \lim_{p \to +\infty} \big|\det(L)\big| \cdot \mathrm{vol}(K_p) = \big|\det(L)\big| \cdot \mathrm{vol}(U).$$

Supposons que L soit une transvection, disons :

$$L : (x_1, x_2, x_3, \ldots, x_n) \mapsto (x_1, x_2 + \alpha x_1, x_3, \ldots, x_n).$$

Le déterminant d'une transvection est égal à 1 et l'on doit montrer que le volume de $V := L(U)$ est égal au volume de U. Comme U et V sont des ouverts bornés, on peut trouver un pavé $\mathbf{P} := I_1 \times \cdots \times I_n$ qui les contient tous les deux. La proposition 187 de la page 350 dit que :

$$\mathrm{vol}(U) = \int_{\mathbf{P}} 1_U \quad \text{et} \quad \mathrm{vol}(V) = \int_{\mathbf{P}} 1_V.$$

Nous allons utiliser le théorème de Fubini pour comparer ces volumes (*cf.* la remarque de la page 352). En tranchant l'ouvert U à x_1 constant, on voit que :

$$1_U(x_1, x_2, \ldots, x_n) = 1_{U_{x_1}}(x_2, \ldots, x_n)$$

avec U_{x_1} un ouvert contenu dans le pavé $I_2 \times \cdots \times I_n$. La fonction $1_{U_{x_1}}$ est intégrable sur $I_2 \times \cdots \times I_n$ et :

$$F(x_1) := \int_{I_2 \times \cdots \times I_n} 1_{U_{x_1}} = \mathrm{vol}(U_{x_1}).$$

Le théorème de Fubini implique que $\mathrm{vol}(U) = \displaystyle\int_{I_1} F$. De même, en tranchant l'ouvert V à x_1 constant, on voit que :

$$1_V(x_1, x_2, \ldots, x_n) = 1_{V_{x_1}}(x_2, \ldots, x_n)$$

avec V_{x_1} un ouvert contenu dans le pavé $I_2 \times \cdots \times I_n$. Si l'on pose $G(x_1) := \mathrm{vol}(V_{x_1})$, le théorème de Fubini implique que $\mathrm{vol}(V) = \displaystyle\int_{I_1} G$. Comme pour tout $x_1 \in I_1$, l'ouvert V_{x_1} s'obtient en translatant l'ouvert U_{x_1} du vecteur $(\alpha x_1, 0 \ldots, 0)$, et comme une translation envoie un pavé sur un pavé de même volume (et donc un ouvert borné sur un ouvert de même volume), on a $F \equiv G$ et $\mathrm{vol}(V) = \mathrm{vol}(U)$. ∎

Revenons maintenant à la preuve du cas général.
Nous allons démontrer que la formule est vraie si $f : U_2 \to [0, +\infty[$ est une fonction continue et positive. Le cas général s'en déduit immédiatement en écrivant :

$$f = f^+ - f^- \quad \text{avec} \quad f^+ := \max(f, 0) \quad \text{et} \quad f^- := \max(-f, 0)$$

et en utilisant la propriété de linéarité de l'intégrale.
Supposons donc que $f : U_2 \to [0, +\infty[$ soit une fonction continue et positive. Soit $g : U_1 \to [0, +\infty[$ la fonction continue et positive définie par $g := f \circ \mathbf{h} \cdot |\operatorname{Jac} \mathbf{h}|$. Comme g est continue sur U_1, elle est intégrable sur tout pavé $\mathbf{P} \subset U_1$. Comme f est (absolument) intégrable sur U_2, elle est (absolument) intégrable sur l'ouvert $\mathbf{h}(\operatorname{Int} \mathbf{P}) \subset U_2$.

Lemme 191. Pour tout pavé $\mathbf{P} \subset U_1$, on a $\displaystyle\int_{\mathbf{P}} g \leqslant \int_{\mathbf{h}(\operatorname{Int} \mathbf{P})} f$.

Démonstration. C'est, une fois établi le cas d'un changement linéaire de variables, le point clé de la démonstration. L'idée de la preuve est de se ramener au voisinage d'un point $\mathbf{x}_0$ par des « zooms » successifs (de rapport $1/2$) sur des pavés bien choisis. Alors le difféomorphisme h est « voisin » de sa différentielle H en $\mathbf{x}_0$, ce qui permet de conclure en utilisant la proposition 190.

Soit $A \in \mathbb{R}$ le réel défini par $A := \dfrac{1}{\operatorname{vol}(\mathbf{P})} \left(\displaystyle\int_{\mathbf{h}(\operatorname{Int} \mathbf{P})} f - \int_{\mathbf{P}} g \right)$. Nous devons montrer que $A \geqslant 0$.
Le pavé $\mathbf{P}$ est la réunion de 2^n pavés $\mathbf{P}_j$, dont les arêtes sont deux fois plus petites et dont les intérieurs sont disjoints. On a alors, en utilisant $A \operatorname{vol} \mathbf{P} = \displaystyle\int_{\mathbf{P}} A$:

$$\sum_{j \in [\![1, 2^n]\!]} \left(\int_{\mathbf{h}(\operatorname{Int} \mathbf{P}_j)} f \right) \leqslant \int_{\mathbf{h}(\operatorname{Int} \mathbf{P})} f = \int_{\mathbf{P}} (g + A) = \sum_{j \in [\![1, 2^n]\!]} \left(\int_{\mathbf{P}_j} (g + A) \right).$$

Pour au moins un des pavés $\mathbf{P}_{j_0}$, on a donc :

$$\int_{\mathbf{h}(\operatorname{Int} \mathbf{P}_{j_0})} f \leqslant \int_{\mathbf{P}_{j_0}} (g + A).$$

En posant $\mathbf{Q}_0 := \mathbf{P}$, $\mathbf{Q}_1 := \mathbf{P}_{j_0}$ et en itérant l'argument ci-dessus, on construit une suite de pavés emboîtés $(\mathbf{Q}_k)_{k \geqslant 0}$ dont les diamètres tendent vers 0 et tels que pour tout $k \geqslant 0$:

$$\int_{V_k} f \leqslant \int_{\mathbf{Q}_k} (g + A) \quad \text{avec} \quad V_k := \mathbf{h}(\operatorname{Int} \mathbf{Q}_k).$$

Par conséquent :

$$\operatorname{vol}(V_k) \cdot \min_{V_k} f \leqslant \operatorname{vol}(\mathbf{Q}_k) \max_{\mathbf{Q}_k} (g + A) \quad \text{et} \quad A \geqslant \frac{\operatorname{vol}(V_k)}{\operatorname{vol}(\mathbf{Q}_k)} \cdot \min_{V_k} f - \max_{\mathbf{Q}_k} g.$$

L'intersection des pavés $\mathbf{Q}_k$ est réduite à un point $\mathbf{x}_0$. Comme g et $\mathbf{h}$ sont continues en $\mathbf{x}_0$ et comme f est continue en $\mathbf{h}(\mathbf{x}_0)$, on a :

$$\max_{\mathbf{Q}_k} g \underset{k \to +\infty}{\longrightarrow} g(\mathbf{x}_0) \quad \text{et} \quad \min_{V_k} f \underset{k \to +\infty}{\longrightarrow} f\big(\mathbf{h}(\mathbf{x}_0)\big).$$

Nous prétendons que :

$$\liminf_{k \to +\infty} \frac{\operatorname{vol} V_k}{\operatorname{vol} \mathbf{Q}_k} \geqslant \big|\operatorname{Jac} \mathbf{h}(\mathbf{x_0})\big|, \tag{39}$$

ce qui implique l'inégalité requise :

$$A \geqslant \left|\operatorname{Jac} \mathbf{h}(\mathbf{x_0})\right| f \circ \mathbf{h}(\mathbf{x}_0) - g(\mathbf{x}_0) = 0.$$

Pour étudier la suite $\left(\frac{\operatorname{vol} V_k}{\operatorname{vol} \mathbf{Q}_k}\right)$ et prouver (39), observons que, par construction :

$$\mathbf{Q}_k = \mathbf{a}_k + \frac{1}{2^k}\mathbf{R} \quad \text{avec} \quad \|\mathbf{a}_k - \mathbf{x}_0\| \leqslant 1/2^k,$$

où a_k est le centre de $\mathbf{Q}_k$ et où $\mathbf{R}$ désigne le pavé fermé de centre 0 dont les arêtes sont égales à celles de $\mathbf{P}$. Soit $\mathbf{R}'$ un pavé fermé de centre 0 contenu dans *l'intérieur* de $\mathbf{R}$ et $c > 0$ un réel tel que tout point de $\mathbb{R}^n \setminus \mathbf{R}$ soit au moins à distance c de $\mathbf{R}'$ (on peut choisir pour $\mathbf{R}'$ un homothétique de $\mathbf{R}$ par une homothétie de centre 0 et de rapport $\lambda \in \,]0,1[$, et l'on a $\operatorname{vol}(\mathbf{R}') = \lambda^n \operatorname{vol}(\mathbf{R})$). Posons $(\mathbf{Q}'_k) := \mathbf{a}_k + \frac{1}{2^k}\mathbf{R}'$. Pour k suffisamment grand, $V_k = \mathbf{h}(\operatorname{Int} \mathbf{Q}_k)$ contient l'image de $\mathbf{Q}'_k$ par l'application affine :

$$H : \mathbf{x} \mapsto \mathbf{h}(\mathbf{x}_0) + D_{\mathbf{x}_0}\mathbf{h}(\mathbf{x} - \mathbf{x}_0).$$

En effet, si ce n'était pas le cas, nous pourrions trouver une suite de points $\mathbf{y}_k \in \mathbf{Q}'_k$ tels que $\mathbf{z}_k := \mathbf{h}^{-1} \circ H(\mathbf{y}_k)$ ne soit pas dans $\mathbf{Q}_k$. Cela n'est pas possible puisque, la différentielle de $\mathbf{h}^{-1} \circ H$ en $\mathbf{x}_0$ étant l'identité, on aurait alors :

$$\mathbf{h}^{-1} \circ H(\mathbf{y}_k) = x_0 + D_{\mathbf{x}_0}(\mathbf{h}^{-1} \circ H)(\mathbf{y}_k - \mathbf{x}_0) + \mathrm{o}\big(\|\mathbf{y}_k - \mathbf{x}_0\|\big) = y_k + \mathrm{o}\big(\|\mathbf{y}_k - \mathbf{x}_0\|\big),$$

d'où :

$$\frac{c}{2^k} \leqslant \left\|\mathbf{h}^{-1} \circ H(\mathbf{y}_k) - \mathbf{y}_k\right\| = \mathrm{o}\big(\|\mathbf{y}_k - \mathbf{x}_0\|\big) = \mathrm{o}\left(\frac{1}{2^k}\right),$$

ce qui est impossible.

D'après l'inclusion $H(\mathbf{Q}'_k) \subset V_k$ et la proposition 190, on a donc :

$$\operatorname{vol} V_k \geqslant \operatorname{vol}\big(H(\mathbf{Q}'_k)\big) = \operatorname{vol} \mathbf{Q}'_k \left|\operatorname{Jac} \mathbf{h}(\mathbf{x_0})\right|.$$

Par conséquent, pour k suffisamment grand :

$$\frac{\operatorname{vol} V_k}{\operatorname{vol} \mathbf{Q}_k} \geqslant \frac{\operatorname{vol}(\mathbf{Q}'_k)}{\operatorname{vol} \mathbf{Q}_k} \left|\operatorname{Jac} \mathbf{h}(\mathbf{x_0})\right| = \frac{\operatorname{vol} \mathbf{R}'}{\operatorname{vol} \mathbf{R}} \left|\operatorname{Jac} \mathbf{h}(\mathbf{x_0})\right|.$$

Le résultat (c'est-à-dire (39)) suit puisque l'on peut prendre $\operatorname{vol} \mathbf{R}'$ arbitrairement proche de $\operatorname{vol} \mathbf{R}$. ∎

On déduit du lemme 191 ci-dessus que, pour toute réunion finie $K := \bigcup_{j \in [\![1,N]\!]} \mathbf{P}_j \subset U_1$ de pavés $\mathbf{P}_j$ d'intérieurs disjoints :

$$\int_K g = \sum_{j \in [\![1,N]\!]} \int_{\operatorname{Int} \mathbf{P}_j} \leqslant \sum_{j \in [\![1,N]\!]} \int_{\mathbf{h}(\operatorname{Int} \mathbf{P}_j)} f \leqslant \int_{U_2} f.$$

Cela montre que g est intégrable sur U_1 et que $\displaystyle\int_{U_1} g \leqslant \int_{U_2} f$.

En appliquant le même argument à $g : U_1 \to [0, +\infty[$ et $\mathbf{h}^{-1} : U_2 \to U_1$, et en utilisant :

$$g \circ \mathbf{h}^{-1} \cdot \left|\operatorname{Jac} \mathbf{h}^{-1}\right| = f \left|\operatorname{Jac} \mathbf{h}\right| \left|\operatorname{Jac} \mathbf{h}^{-1}\right| = f,$$

on obtient $\displaystyle\int_{U_2} f \leqslant \int_{U_1} g$, d'où l'égalité des intégrales.

On peut *reformuler* le théorème 189 en utilisant la notion d'intégrale généralisée au sens de Riemann.

Corollaire 192. Soient U_1 et U_2 deux ouverts de $\mathbb{R}^n$ et $\mathbf{h} : U_1 \to U_2$ un difféomorphisme de classe $\mathcal{C}^1$. Une fonction continue $f : U_2 \to \mathbb{R}^n$ est intégrable au sens de Riemann généralisé sur U_2 si, et seulement si, $f \circ \mathbf{h} \cdot |\operatorname{Jac} \mathbf{h}|$ est intégrable au sens de Riemann généralisé sur U_1. Dans ce cas :

$$\int_{U_2} f(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_{U_1} f \circ \mathbf{h}(\mathbf{x}) \cdot \big|\operatorname{Jac} \mathbf{h}(\mathbf{x})\big|\,\mathrm{d}\mathbf{x}.$$

On peut aussi prouver ce résultat en utilisant la théorie de Riemann généralisée au lieu de la théorie de Henstock-Kurzweil (*cf.* l'exercice III.1.43).

Exercice type corrigé

III.1.0 On dit qu'une fonction $f : I \to \mathbb{R}$ $\mathscr{HK}$-intégrable sur un intervalle $I \subset \mathbb{R}$ est *localement intégrable au sens de Lebsgue* en $c \in I$ s'il existe un segment d'intérieur non vide centré en c sur lequel f est intégrable au sens de Lebesgue. Au vu des exemples donnés dans le cours, on pourrait croire que l'ensemble des points où f n'est pas intégrable au sens de Lebesgue est « petit ». Le but de l'exercice est de montrer que ce n'est pas en général le cas en construisant une famille de fonctions $\mathscr{HK}$-intégrables sur $[0,1]$ telle que, pour chacune d'entre elles, l'ensemble des points où elle n'est pas localement intégrable au sens de Lebesgue soit d'intérieur vide mais de mesure de Lebesgue *strictement positive*. On verra même que cette mesure peut être choisie arbitrairement dans $]0,1[$. Le principe de la construction est de définir une fonction « oscillante non bornée » sur chaque composante connexe de $[0,1] \setminus K$, où K est une « modification » de l'ensemble de Cantor. L'exercice est inspiré d'un exemple de Jitan Lu et Peng Yee Lee (1999). L'idée est une variante d'une construction antérieure due à Vito Volterra.
Pour tous $a, b \in \mathbb{R}$, $a < b$, on définit une fonction $f_{a,b} : \mathbb{R} \to \mathbb{R}$ par :

$$f_{a,b} : x \mapsto \begin{cases} (x-a)^2(x-b)^2 \sin \frac{1}{(x-a)^2(x-b)^2} & \text{si} \quad x \in]a,b[\\ 0 & \text{si} \quad x \in \mathbb{R} \setminus]a,b[\end{cases}.$$

1. a) Montrer que $f_{a,b}$ est partout dérivable sur $\mathbb{R}$. Montrer que $f'_{a,b}$ est $\mathscr{HK}$-intégrable sur $\mathbb{R}$.

   b) Montrer que $f'_{a,b}$ n'est pas localement intégrable au sens de Lebesgue en a et b.

2. On reprend la méthode de construction de l'ensemble de Cantor avec une « petite modification ». On divise à chaque étape k de la récurrence les segments selon des proportions données à l'avance (dépendant de n) au lieu de les diviser en trois.

   Soit $(\varepsilon_k)_{k\in\mathbb{N}}$ une suite de nombres réels telle que $\varepsilon_0 := 0$ et, pour tout $k \in \mathbb{N}^*$, $0 < \varepsilon_k < 1$. On définit par récurrence des fermés $K_k \subset [0,1]$ formés de 2^k segments deux à deux disjoints de même longueur. On choisit $K_0 := [0,1]$, $K_1 := \left[0, \frac{1-\varepsilon_1}{2}\right] \cup \left[\frac{1+\varepsilon_1}{2}, 1\right]$. Pour $k \geqslant 1$, on définit K_{k+1} en divisant chacun des 2^k segments de K_k en trois parties *proportionnellement* à $\frac{1-\varepsilon_{k+1}}{2}, \varepsilon_{k+1}, \frac{1-\varepsilon_{k+1}}{2}$, puis en supprimant l'intervalle central *ouvert*. On définit un

« ensemble de Cantor généralisé » $K := \bigcap_{k\in\mathbb{N}} K_k$. Si, pour tout $k \in N^*$, $\varepsilon_k := 1/3$ on retrouve l'ensemble de Cantor.

a) Calculer $\operatorname{long} K_n$.

b) Montrer que K est un compact d'intérieur vide.

c) Montrer que K est intégrable et calculer sa mesure de Lebesgue $\mu(K)$. Donner une condition nécessaire et suffisante sur la suite $(\varepsilon_k)_{k\in\mathbb{N}^*}$ pour que $\mu(K) > 0$.

d) Montrer que si $\lambda \in [0,1[$, on peut choisir la suite $(\varepsilon_k)_{k\in\mathbb{N}^*}$ de telle sorte que $\mu(K) = \lambda$.

3. Le compact K étant défini comme en 2, l'ouvert $U := [0,1]\backslash K$ est une réunion dénombrable disjointe d'intervalles ouverts $]a_n, b_n[$ ($n \in \mathbb{N}$). On définit une fonction $f : [0,1] \to \mathbb{R}$ par :

$$f : x \mapsto \begin{cases} f_{a_n,b_n}(x) & \text{s'il existe } n \in \mathbb{N} \text{ tel que } x \in]a_n, b_n[\\ 0 & \text{si } \quad x \in K \end{cases}.$$

a) Montrer que f est dérivable sur $[0,1]$ avec une dérivée nulle sur K.

b) Montrer que f' est $\mathscr{HK}$-intégrable sur $[0,1]$ mais n'est intégrable au sens de Lebesgue sur aucun sous-segment contenant un point de K.

c) Quel est l'ensemble des points où f est localement intégrable au sens de Lebesgue ? Conclure.

Solution.

1. Pour cette question, il suffit (par translation et homothétie sur la variable) de traiter le cas où $a = 0$ et $b = 1$ et d'étudier les problèmes au voisinage de $x = 0$ (le cas $x = 1$ est similaire et la situation est triviale au voisinage d'un point de $\mathbb{R} \setminus \{0,\ 1\}$, puisque $f_{0,1}$ est de classe $\mathcal{C}^1$ sur $\mathbb{R}\setminus\{0,1\}$). On note, pour simplifier, $f := f_{0,1}$ (seulement dans cette question).

a) On a, pour tout $x \in]0,1[$, $\frac{1}{x}f(x) = \frac{1}{x}x^2(x-1)^2 \sin\frac{1}{x^2(x-1)^2} = x(x-1)^2 \sin\frac{1}{x^2(x-1)^2}$, donc $\lim\limits_{x\to 0^+} \frac{1}{x}f(x) = 0$. Ainsi, f est dérivable et de dérivée nulle en 0. On prouve de même que f est dérivable et de dérivée nulle en 1, par suite, f est dérivable sur $\mathbb{R}$ et, d'après le théorème fondamental (théorème 81 de la page 268), f' est $\mathscr{HK}$-intégrable sur $\mathbb{R}$.

b) On reprend la méthode de l'exemple de la page 287. La fonction $x \mapsto x^2(1-x)^2$ est strictement croissante sur $]0,1/2[$, il existe donc deux suites $(\alpha_k)_{k\geqslant 3}$ et $(\beta_k)_{k\geqslant 3}$ de $]0,1/2[$, telles que $\alpha_k^2(1-\alpha_k)^2 = \frac{1}{(2k+1)\pi}$ et $\beta_k^2(1-\beta_k)^2 = \frac{1}{2k\pi+\pi/2}$. Les segments $[\alpha_k, \beta_k]$ sont deux à deux disjoints et $|f'|$ est continue, donc intégrable sur chacun d'entre eux. Le signe de f' sur $]0,1/2[$ est celui de

$$\sin\frac{1}{x^2(x-1)^2} - \frac{1}{x^2(x-1)^2}\cos\frac{1}{x^2(x-1)^2}.$$

De plus, si $x \in [\alpha_k, \beta_k]$, on a $2k\pi + \frac{\pi}{2} \leqslant \frac{1}{x^2(1-x)^2} \leqslant 2k\pi + \pi$, donc $f' \geqslant 0$ sur chaque segment $[\alpha_k, \beta_k]$. Par suite :

$$\int_{\alpha_k}^{\beta_k} |f'| \geqslant \int_{\alpha_k}^{\beta_k} f' = f(\beta_k) - f(\alpha_k) = \frac{2}{(4k+1)\pi}.$$

Si f' était intégrable au sens de Lebesgue au voisinage de 0, $|f'|$ serait $\mathscr{HK}$-intégrable sur $[0,1/2]$ et l'on aurait :

$$\forall\, n \geqslant 3, \qquad \sum_{k\in[\![3,n]\!]} \frac{2}{(4k+1)\pi} = \sum_{k\in[\![3,n]\!]} \int_{\alpha_k}^{\beta_k} f' \leqslant \int_0^{1/2} |f'| < +\infty.$$

D'où une contradiction.

2. a) On montre par récurrence que, pour tout $n \in \mathbb{N}$: $\operatorname{long} K_n = \prod\limits_{k\in[\![1,n]\!]} (1-\varepsilon_k)$.

   b) L'ensemble K est un fermé de $[0,1]$, il est donc compact.
   On a, d'après a) : $\operatorname{long} K_n = \prod\limits_{k\in[\![1,n]\!]} (1-\varepsilon_k)$, donc la longueur de chacun des 2^n segments constituant K_n est $2^{-n} \prod\limits_{k\in[\![1,n]\!]} (1-\varepsilon_k) \leqslant 2^{-n}$. Elle tend vers 0 quand n tend vers $+\infty$, on en déduit (comme dans le cas de l'ensemble de Cantor) que K est *d'intérieur vide*.

   c) La suite $(K_k)_{k\in\mathbb{N}}$ est décroissante, donc la mesure de Lebesgue $\mu(K)$ de K est $\lim\limits_{n\to+\infty} \prod\limits_{k\in[\![1,n]\!]} (1-\varepsilon_k)$.
   On a :
   $$\ln \prod_{k\in[\![1,n]\!]} (1-\varepsilon_k) = \sum_{k\in[\![1,n]\!]} \ln(1-\varepsilon_k) \leqslant -\sum_{k\in[\![1,n]\!]} \varepsilon_k.$$
   Par suite, si la série $\sum \varepsilon_k$ est *divergente*, $\lim\limits_{n\to+\infty} \prod\limits_{k\in[\![0,n]\!]} (1-\varepsilon_k) = 0$ et $\mu(K) = 0$.
   Si, inversement la série $\sum \varepsilon_k$ est *convergente*, $\lim\limits_{k\to+\infty} \varepsilon_k = 0$ et la série $\sum\limits_{k\in[\![0,n]\!]} \ln(1-\varepsilon_k)$ est convergente.
   Si l'on note $-c$ sa somme, on a $\lim\limits_{n\to+\infty} \prod\limits_{k\in[\![0,n]\!]} (1-\varepsilon_k) = e^{-c}$ et $\mu(K) = e^{-c} \in \,]0,1[$.
   Ainsi, on a $\mu(K) > 0$ si, et seulement si, la série $\sum \varepsilon_k$ est convergente.

   d) Soit $\lambda \in [0,1[$. Pour $\lambda = 0$ on prend une série divergente, par exemple celle correspondant à l'ensemble de Cantor. Supposons donc $\lambda \in \,]0,1[$ et notons $c := -\ln\lambda \in \,]-\infty,0[$. Il existe une série $\sum\limits_{n\in\mathbb{N}^*} (-c_n)$ à termes tous strictement négatifs de somme $-c$. On définit, pour tout $k \in \mathbb{N}^*$, $\varepsilon_n := 1-e^{-c_n}$. La suite $(\varepsilon_k)_{k\in\mathbb{N}}$ répond à la question.

3. a) La dérivabilité de f sur tout intervalle $]a_n,b_n[$ ($n \in \mathbb{N}$) est évidente ($f_{a,b}$ est dérivable sur $]a,b[$). Montrons que f est dérivable, de dérivée nulle, en tout point $x_0 \in K$. Soit $x > x_0$. Si $x \in K$, $f(x) = 0$, donc $\frac{f(x)-f(x_0)}{x-x_0} = 0$. Sinon, il existe $n \in \mathbb{N}$ (unique) tel que $x \in \,]a_n,b_n[$ et l'on a $x_0 \leqslant a_n$ (puisque $x_0 \notin \,]a_n,b_n[$). Alors :
   $$f(x) \leqslant (b_n-a_n)^2(x-a_n)^2 \leqslant (x-a_n)^2 \leqslant (x_0-a_n)^2$$
   et $\frac{f(x)}{x-x_0} = \frac{f(x)-f(x_0)}{x-x_0} \leqslant x-x_0$. On a un résultat similaire pour $x < x_0$.

b) D'après le théorème fondamental (théorème 81), f' est $\mathscr{HK}$-intégrable sur $\mathbb{R}$. Soit J un segment de $[0,1]$ contenant un point de K. Si k est assez grand, J contient l'un des segments de K_k, par suite, il contient un segment $[a_n, b_n]$ (dans son intérieur), pour un $n \in \mathbb{N}$ convenable. D'après la question 1.b), la fonction f' n'est pas localement intégrable au sens de Lebesgue en a_n et b_n. Par suite, f' n'est pas intégrable au sens de Lebesgue sur J.

c) Notons V l'ensemble des points $x \in [0,1]$ tels que f' soit localement intégrable au sens de Lebesgue en x. Il est évidemment ouvert et il contient U (puisque f' est continue sur U). D'après b), il ne peut pas contenir de point de K, donc $V = U$ et l'ensemble des points où f' n'est pas localement intégrable au sens de Lebesgue est K, il est donc d'intérieur vide (d'après 2 a)) et de mesure de Lebesgue *strictement positive* si la série $\sum \varepsilon_k$ est convergente. Cette mesure peut être choisie arbitrairement d'après 2.d).

Pour aller plus loin

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

Exercices sur la section 1

III.1.1 Le but de cet exercice est de terminer la preuve de la proposition 31 de la page 197. Soient $X,\ Y \subset \mathbb{R}^n$ deux parties pavables. On choisit un pavé fermé $\mathbf{P}$ tel que $X,\ Y \subset \operatorname{Int} \mathbf{P}$.

1. Montrer que l'on peut écrire $X = X' \cup X''$ comme réunion de deux parties pavables, où $X' = \overline{\operatorname{Int} X'}$ et X'' est d'intérieur vide. Montrer que $\overline{\operatorname{Int} X}$ est une partie pavable.
2. a) Montrer que $\mathbf{P} \setminus \operatorname{Int} Y$ est pavable.

   b) Montrer que $X \setminus \operatorname{Int} Y$ est pavable.
3. Montrer $\operatorname{vol}(X \cup Y) + \operatorname{vol}(X \cap Y) = \operatorname{vol} X + \operatorname{vol} Y$.

III.1.2 Soient $n, p \in \mathbb{N}^*$ avec $p < n$, $\mathbf{P} : [0,1]^p \subset \mathbb{R}^p$ et $\gamma : \mathbf{P} \to \mathbb{R}^n$ une application M-lipschitzienne. Montrer que l'image de γ est $\mathscr{R}$-négligeable.

III.1.3 Montrer qu'une partie $A \subset \mathbb{R}^n$ est $\mathscr{R}$-négligeable si, et seulement si, pour tout $\varepsilon > 0$, on peut la recouvrir par un nombre fini de boules *de même rayon* dont la somme des volumes est majorée par ε.

III.1.4 Soient $n, m \in \mathbb{N}^*$, avec $n \leqslant m$ et $A \subset \mathbb{R}^n$ une partie $\mathscr{R}$-négligeable.

1. Soit $\varphi : A \to \mathbb{R}^m$ une application χ-lipshitzienne. Montrer que $\varphi(A) \subset \mathbb{R}^m$ est $\mathscr{R}$-négligeable. (On utilisera le résultat de l'exercice précédent.)
2. Soient $U \subset \mathbb{R}^n$ un ouvert et $A \subset \mathbb{R}^n$ une partie *bornée* telle que $\bar{A} \subset U$. Soit $\varphi : U \to \mathbb{R}^m$ une application de classe $\mathcal{C}^1$, montrer que $\varphi(A) \subset \mathbb{R}^m$ est $\mathscr{R}$-négligeable.

III.1.5 *Exemple du problème* 3. *pour les intégrales itérées, signalé page 210.*
On définit une fonction $f : [0,1]^2 \to \mathbb{R}$ par :

$$f : (x,y) \mapsto \begin{cases} 1 & \text{s'il existe } n \in \mathbb{N},\ k,\ l \in \mathbb{N}^*, \quad \text{tels que } (x,y) = \left(\frac{2k-1}{2^n}, \frac{2l-1}{2^n}\right) \\ 0 & \text{sinon.} \end{cases}$$

1. Montrer que, pour tout $x \in [0,1]$, la fonction $y \mapsto f(x,y)$ est $\mathscr{R}$-intégrable sur $[0,1]$ et que $\displaystyle\int_0^1 f(x,y)\,\mathrm{d}y = 0$. En déduire que $\displaystyle\int_0^1 \mathrm{d}x \int_0^1 f(x,y)\,\mathrm{d}y$ existe et est égale à 0.
   Montrer que $\displaystyle\int_0^1 \mathrm{d}y \int_0^1 f(x,y)\,\mathrm{d}x$ existe et est égale à 0.
2. Montrer que les ensembles :
   $$\{(x,y) \in [0,1]^2 \mid f(x,y) = 1\} \qquad \text{et} \qquad \{(x,y) \in [0,1]^2 \mid f(x,y) = 0\}$$
   sont *denses* dans $[0,1]^2$. En déduire que f n'est pas $\mathscr{R}$-intégrable sur $[0,1]^2$.

III.1.6 *Exemple du problème* 1. *pour les intégrales itérées, signalé page 210.*
On définit une fonction $f : [0,1]^2 \to \mathbb{R}$ par :

$$f : (x,y) \mapsto \begin{cases} 1 & \text{si } (x = 0 \text{ ou } x = 1) \text{ et } y \in \mathbb{Q} \\ 1 & \text{si } (y = 0 \text{ ou } y = 1) \text{ et } x \in \mathbb{Q} \\ 0 & \text{sinon.} \end{cases}$$

1. Montrer que, pour $x = 0$ et $x = 1$, la fonction $y \mapsto f(x,y)$ n'est pas $\mathscr{R}$-intégrable sur $[0,1]$.
2. Montrer que $\{(x,y) \in [0,1]^2 \mid f(x,y) \neq 0\}$ est $\mathscr{R}$-négligeable.
   Montrer que f est $\mathscr{R}$-intégrable sur $[0,1]^2$ et que $\displaystyle\int_{[0,1]^2} f = 0$.

III.1.7 *Exemple du problème* 1. *pour les intégrales itérées, signalé page 210.*
Pour tout $x \in \mathbb{Q}$, on note $x = p_x/q_x$, où $p_x \in \mathbb{Z}$ et $q_x \in \mathbb{N}^*$, p_x et q_x étant premiers entre eux.
On définit une fonction $f : [0,1]^2 \to \mathbb{R}$ par :

$$f : (x,y) \mapsto \begin{cases} 1/q_x + 1/q_y & \text{si} \quad x,\ y \in \mathbb{Q} \\ 0 & \text{si} \quad x \notin \mathbb{Q} \quad \text{ou} \quad y \notin \mathbb{Q}. \end{cases}$$

1. Montrer que la fonction f est continue sur $A := \big([0,1] \setminus \mathbb{Q}\big)^2$.
2. Montrer que $[0,1]^2 \setminus A$ est négligeable et que f est $\mathscr{R}$-intégrable sur $[0,1]^2$.
3. Soit $x \in \mathbb{Q}$ fixé.
   Montrer que la fonction $y \in [0,1] \to f(x,y) \in \mathbb{R}$ n'est pas $\mathscr{R}$-intégrable sur $[0,1]$.

III.1.8 On note $A := \big\{(\lambda + \mu\sqrt{2}, \lambda - \mu\sqrt{2}) \in [0,1]^2 \ \big|\ \lambda,\ \mu \in \mathbb{Q}\big\}$ et l'on définit une fonction $f : [0,1]^2 \to \mathbb{R}$ par $f := 1_A$.

1. Montrer que A est dense dans $[0,1]^2$. En déduire que f n'est pas $\mathscr{R}$-intégrable sur $[0,1]^2$.

2. Soit $x \in [0,1]$ fixé. Montrer que, pour tout $y \in [0,1]$, $f(x,y) = 0$ sauf peut-être pour une valeur de y. Montrer que la fonction $y \in [0,1] \to f(x,y) \in \mathbb{R}$ est intégrable sur $[0,1]$ et que $\displaystyle\int_0^1 f(x,y)\,\mathrm{d}y = 0$.
3. Montrer que $\displaystyle\int_0^1 \mathrm{d}x \int_0^1 f(x,y)\,\mathrm{d}y = \int_0^1 \mathrm{d}y \int_0^1 f(x,y)\,\mathrm{d}x = 0$.

III.1.9 *Exemple du problème* 2. *pour les intégrales itérées, signalé page 210.*
On définit une fonction $f : [0,1]^2 \to \mathbb{R}$ par :

$$f : (x,y) \to f(x,y) := 1_{\mathbb{Q}}(x) + 2y\,1_{\mathbb{R}\setminus\mathbb{Q}}(x).$$

1. Soit $x \in [0,1]$ fixé. Montrer que la fonction $y \in [0,1] \longmapsto f(x,y) \in \mathbb{R}$ est $\mathscr{R}$-intégrable sur $[0,1]$ et que $\displaystyle\int_0^1 f(x,y)\,\mathrm{d}y = 1$. En déduire que $\displaystyle\int_0^1 \mathrm{d}x \int_0^1 f(x,y)\,\mathrm{d}y = 1$.
2. Soit $y \in [0,1]$, $y \neq 1/2$, fixé. Montrer que la fonction $x \in [0,1] \to f(x,y) \in \mathbb{R}$ n'est pas $\mathscr{R}$-intégrable sur $[0,1]$.

III.1.10 Dans cet exercice on montre une *généralisation* du théorème d'Arzela (*cf.* le corollaire 122 de la page 706 du module *Intégration* du livre [L2]) en utilisant *seulement* le résultat de l'exercice 7 de la page 638 de ce même module :
$(\star)$ Soit $(g_n)_{n\in\mathbb{N}}$ une suite décroissante de fonctions en escalier sur $[a,b]$, convergeant simplement vers 0 sur $[a,b]$. Alors $\displaystyle\lim_{n\to+\infty} \int_a^b g_n = 0$.
On obtient donc, en particulier, une nouvelle preuve très simple du théorème d'Arzela.

1. Soit $(f_k)_{k\in[\![1,n]\!]}$ une suite finie décroissante de fonctions bornées définies sur $[a,b]$. Soit $(g_k)_{k\in[\![1,n]\!]}$ une suite finie de fonctions $\mathscr{R}$-intégrables sur $[a,b]$ telles que, pour tout $k \in [\![1,n]\!]$, l'on ait $g_k \leqslant f_k$. On pose, pour tout $k \in [\![1,n]\!]$, $h_k := \min\limits_{j\in[\![1,k]\!]} g_j$. Montrer que :
$$\int_a^b h_n \geqslant \sum_{k\in[\![1,n]\!]} \int_a^b g_k - \sum_{k\in[\![1,n-1]\!]} {}_*\!\!\int_a^b f_k = {}_*\!\!\int_a^b f_n - \sum_{k\in[\![1,n]\!]} \left({}_*\!\!\int_a^b f_k - \int_a^b g_k \right). \quad (40)$$
2. Soit $(f_n)_{n\in\mathbb{N}}$ une suite uniformément bornée de fonctions positives définies sur $[a,b]$. On suppose que cette suite tend simplement vers 0 sur $[a,b]$. Montrer que $\displaystyle\lim_{n\to+\infty} {}_*\!\!\int_a^b f_n = 0$.
3. En déduire une nouvelle preuve du théorème d'Arzela.

III.1.11 Le but de cet exercice est de prouver la proposition 55 de la page 212 (théorème de Warusfel).
Soient $\mathbf{P} := [a,b] \times [c,d]$ et $f : \mathbf{P} \to [-M,M]$ une fonction bornée.
On *suppose* que, pour tout $x \in [a,b]$, la fonction $f_x : y \mapsto f(x,y)$ est $\mathscr{R}$-intégrables sur $[c,d]$, et, pour tout $y \in [c,d]$, la fonction $g_y : x \mapsto f(x,y)$ est $\mathscr{R}$-intégrables sur $[a,b]$. On note :

$$F(x) := \int_c^d f(x,y)\,\mathrm{d}y = \int_c^d f_x \qquad \text{et} \qquad G(y) := \int_a^b f(x,y)\,\mathrm{d}x = \int_a^b g_y.$$

1. Soit $\mathbf{u} := \{x_0, \ldots, x_N\}$ une subdivision de $[a, b]$. On note, pour tout $y \in [c, d]$:
$$\forall\, j \in [\![0, N-1]\!], \quad \varepsilon_j(y) := \sup_{x \in [x_j, x_{j+1}]} f(x, y) - \inf_{x \in [x_j, x_{j+1}]} f(x, y)$$
et :
$$\Sigma_{\mathbf{u}}(y) := \sum_{j \in [\![0, N-1]\!]} \varepsilon_j(y)(x_{j+1} - x_j) = D_{\mathbf{u}} f_y - d_{\mathbf{u}} f_y.$$
Montrer que, pour tous $\xi,\ \xi' \in [x_j, x_{j+1}]$, $\left|F(\xi) - F(\xi')\right| \leqslant \displaystyle\int_c^d \varepsilon_j(y)\,\mathrm{d}y$.
En déduire $D_{\mathbf{u}} F - d_{\mathbf{u}} F \leqslant \displaystyle\int_c^d \Sigma_{\mathbf{u}}(y)\,\mathrm{d}y$.
2. Étudier, pour tout $y \in [c, d]$, $\Sigma_{\mathbf{u}}(y)$ « quand le pas de la subdivision $\mathbf{u}$ tend vers 0 ». En déduire que, pour tout $\varepsilon > 0$, il existe une subdivision $\mathbf{u}$ telle que $\displaystyle\int_c^d \Sigma_{\mathbf{u}}(y)\,\mathrm{d}y \leqslant \varepsilon$ (on utilisera l'exercice III.1.10 ci-dessus).
Montrer que F est $\mathscr{R}$-intégrable sur $[a, b]$ et que G est $\mathscr{R}$-intégrable sur $[c, d]$.
3. Soit $y \in [c, d]$.
Pour tout $n \in \mathbb{N}^*$ on considère la *somme de Cauchy* de la fonction $f_y : x \mapsto f(x, y)$ associée à la subdivision de $[a, b]$ en n parties égales $\{x_j := a + j\frac{b-a}{n}\}_{j \in [\![0, n]\!]}$, c'est-à-dire $G_n(y) := \frac{b-a}{n} \sum\limits_{j \in [\![0, n-1]\!]} f(x_j, y)$. Montrer que $(G_n)_{n \in \mathbb{N}^*}$ est une suite bornée de fonctions convergeant simplement vers G sur $[a, b]$.
Exprimer $\displaystyle\int_c^d G_n(y)\,\mathrm{d}y$ comme une somme de Riemann associée à F. Conclure.

III.1.12 Calculer $\displaystyle\int_{[0,1]^2} \frac{1}{(1 + x + y)^2}\,\mathrm{d}x\,\mathrm{d}y$.

III.1.13 Calculer l'aire de l'ellipse $E := \left\{(x, y) \in \mathbb{R}^2 \;\middle|\; \dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} < 1.\right\}$.

III.1.14 Calculer les intégrales suivantes :

1. $\displaystyle\int_U \frac{1}{(x + y)^3}\,\mathrm{d}x\,\mathrm{d}y$ avec $U := \left\{(x, y) \in \mathbb{R}^2 \;\middle|\; x > 1,\ y > 1 \text{ et } x + y < 3\right\}$.
2. $\displaystyle\int_V (x + y)e^{-x}e^{-y}\,\mathrm{d}x\,\mathrm{d}y$ avec $V := \left\{(x, y) \in \mathbb{R}^2 \;\middle|\; x > 0,\ y > 0 \text{ et } x + y < 1\right\}$.

III.1.15 Soient $U \subset \mathbb{R}^n$ un ouvert et $(K_k)_{k \in \mathbb{N}}$ une suite croissante de compacts contenus dans U telle que $\bigcup\limits_{k \in \mathbb{N}} K_k = U$.
Montrer que c'est une suite exhaustive de compacts de U si, et seulement si, $\bigcup\limits_{k \in \mathbb{N}} \operatorname{Int} K_k = U$.

III.1.16
On se propose de montrer que la suite $(\cos n^2)_{n \in \mathbb{N}}$ n'a pas de limite quand n tend vers $+\infty$.

1. Montrer que la suite $\big(\cos(2n + 1)\big)_{n \in \mathbb{N}}$ ne peut pas tendre vers 1 quand n tend vers $+\infty$.

2. Montrer, en utilisant $\cos(n+1)^2$, que la suite $(\cos n^2)_{n\in\mathbb{N}}$ ne peut pas tendre vers 1 quand n tend vers $+\infty$.
3. Montrer qu'il existe un nombre infini d'entiers $n \in \mathbb{N}$ tels que $2n+1$ soit un carré. En déduire, en utilisant $\cos(n+1)^2$, que la seule limite possible pour la suite $(\cos n^2)_{n\in\mathbb{N}}$, quand n tend vers $+\infty$, est $-1/2$.
4. En utilisant :

$$\cos 9x = T_9(\cos x) = 256\cos^9 x - 576\cos^7 x + 432\cos^5 x - 120\cos^3 x + 9\cos x$$

(*cf.* polynômes de Tchebychev), montrer que $-1/2$ ne peut pas être limite de $(\cos n^2)_{n\in\mathbb{N}}$. Conclure.

III.1.17 On définit une fonction $f : \mathbb{R}^2 \to \mathbb{R}$ par :

$$f : (x,y) \mapsto f(x,y) = \begin{cases} 0 & \text{si} \quad (x,y) = (0,0) \\ \frac{x^2-y^2}{(x^2+y^2)^2} & \text{si} \quad (x,y) \neq (0,0). \end{cases}$$

Pour tout $n \in \mathbb{N}^*$, on note :

$$K_n := \{(x,y) \in \mathbb{R}^2 \mid 1/n \leqslant x \leqslant 1,\ 1/n \leqslant y \leqslant 1\}$$
$$K'_n := \{(x,y) \in \mathbb{R}^2 \mid 1/n^2 \leqslant x \leqslant 1,\ 1/n \leqslant y \leqslant 1\}$$
$$K''_n := \{(x,y) \in \mathbb{R}^2 \mid 1/n \leqslant x \leqslant 1,\ 1/n^2 \leqslant y \leqslant 1\}.$$

1. Montrer que $(K_n)_{n\in\mathbb{N}^*}$, $(K'_n)_{n\in\mathbb{N}^*}$ et $(K''_n)_{n\in\mathbb{N}^*}$ sont des suites exhaustives de compacts de $]0,1]^2$.
2. Montrer que, pour tout $n \in \mathbb{N}^*$, $\displaystyle\int_{K_n} f = 0$.
3. Calculer $\displaystyle\int_{K'_n} f$ et montrer que $\displaystyle\lim_{n\to+\infty}\int_{K'_n} f = -\pi/4$.
4. Montrer que $\displaystyle\lim_{n\to+\infty}\int_{K_n} f \neq \lim_{n\to+\infty}\int_{K'_n} f$.
5. Calculer $\displaystyle\int_{K''_n} f$, puis $\displaystyle\lim_{n\to+\infty}\int_{K''_n} f$. Conclure.

III.1.18 Soit $\Sigma := \sum a_{n,m}$ une *série double* à termes réels.
Pour les définitions et propriétés relatives aux séries doubles, on se reportera au module *Suites et séries de fonctions* du livre [L2].
On note $\Delta := \{(x,y) \in \mathbb{R}^2 \mid x \geqslant 0,\ y \geqslant 0\}$.
Pour tous $n, m \in \mathbb{N}$, on note $L_{n,m} := [n, n+1[\times [m, m+1[$.

1. On suppose que la série double Σ est *absolument convergente*. Soit $\varepsilon > 0$. Montrer qu'il existe un sous-ensemble *fini* $I \subset \mathbb{N}^2$ tel que, pour tout sous-ensemble fini $J \subset \mathbb{N}^2$ tel que $I \subset J$, l'on ait $\displaystyle\left|\sum_{n\in\mathbb{N}}\sum_{m\in\mathbb{N}} a_{n,m} - \sum_{(n,m)\in J} a_{n,m}\right| \leqslant \varepsilon$.

2. Montrer que les $\{L_{n,m}\}_{(n,m)\in\mathbb{N}^2}$ forment une partition de Δ.

   Soit $\varphi : \mathbb{N} \to \mathbb{N}^2$ une *bijection*. Pour tout $k \in \mathbb{N}$, on désigne par K_k l'adhérence de $\bigcup\limits_{h\in[\![0,k]\!]} L_{\varphi(h)}$. Montrer que $\{K_k\}_{k:in\mathbb{N}}$ est une suite exhaustive de compacts de Δ.

3. À la série double Σ, on associe la fonction $f : \Delta \to \mathbb{R}$ définie par $f(x,y) := a_{n,m}$ si $(x,y) \in L_{n,m}$.
   a) Montrer que f est localement intégrable au sens de Riemann sur Δ.
   b) Montrer que les conditions suivantes sont équivalentes.
      (i) La fonction f est intégrable au sens de Riemann généralisé sur Δ.
      (ii) La série double Σ est absolument convergente.
      (iii) La série $\sum a_{\varphi(h)}$ est absolument convergente.

      Montrer que si ces conditions sont satisfaites :

$$\int_\Delta f = \sum_{n\in\mathbb{N}}\sum_{m\in\mathbb{N}} a_{n,m} = \sum_{k\in\mathbb{N}} a_{\varphi(k)}.$$

III.1.19 On propose dans cet exercice une preuve récente (et simple...) de l'égalité $\zeta(2) = \frac{\pi^2}{6}$ par *changement de variables trigonométrique* dans une intégrale double *généralisée* due à F. Beukers, J.A.C. Kolk et E. Calabi (1993).

1. Montrer que $\displaystyle\sum_{n\in\mathbb{N}} \frac{1}{(2n+1)^2} = \frac{3}{4}\zeta(2)$.

2. Montrer que l'intégrale $\displaystyle\iint_{[0,1[^2} \frac{\mathrm{d}x\,\mathrm{d}y}{1-x^2y^2}$ est convergente et que :

$$I := \iint_{[0,1[^2} \frac{\mathrm{d}x\,\mathrm{d}y}{1-x^2y^2} = \iint_{]0,1[^2} \frac{\mathrm{d}x\,\mathrm{d}y}{1-x^2y^2} = \frac{3}{4}\,\zeta(2).$$

3. On note $T := \left\{(\varphi,\psi) \in \mathbb{R}^2 \;\middle|\; \varphi \geqslant 0,\ \psi \geqslant 0,\ \varphi+\psi \leqslant \pi/2\right\}$ et l'on définit une application $h : \operatorname{Int} T \to \mathbb{R}^2$ par :

$$(\varphi,\psi) \in \operatorname{Int} T \mapsto h(\varphi,\psi) = \left(x = \frac{\sin\varphi}{\cos\psi},\ y = \frac{\sin\psi}{\cos\varphi}\right).$$

   Montrer que h induit un difféomorphime de classe $\mathcal{C}^1$ de $\operatorname{Int} T$ sur $]0,1[^2$ et calculer l'inverse de ce difféomorphisme. Calculer le jacobien de h.

4. Montrer que $I = \displaystyle\iint_T \mathrm{d}\varphi\,\mathrm{d}\psi$. Conclure.

III.1.20 On propose dans cet exercice une preuve récente de l'égalité $\zeta(2) = \frac{\pi^2}{6}$ par *changement de variables algébrique* dans une intégrale double *généralisée*, c'est-à-dire par égalité de *périodes* (*cf. Un peu d'histoire* page 177). Cette preuve est due à Zagier et Kontsevitch.

1. Calculer $\displaystyle\sum_{k\in\mathbb{N}^*} \frac{1}{(k+1/2)^2}$.

2. Montrer que l'intégrale $\displaystyle\iint_{]0,1[^2} \frac{1}{1-xy}\frac{\mathrm{d}x\,\mathrm{d}y}{\sqrt{xy}}$ est convergente et que :
$$I := \iint_{]0,1[^2} \frac{1}{1-xy}\frac{\mathrm{d}x\,\mathrm{d}y}{\sqrt{xy}} = 3\,\zeta(2).$$
3. On définit une application $g : \mathbb{R}^2 \to \mathbb{R}^2$ par :
$$g : (\xi,\eta) \mapsto \left(x = \xi^2\frac{1+\eta^2}{1+\xi^2},\ y = \eta^2\frac{1+\xi^2}{1+\eta^2}\right)$$
Calculer le jacobien de g. Montrer que g induit un difféomorphisme de classe $\mathcal{C}^1$ de $\{(\xi,\eta)\in\mathbb{R}^2 \mid 0<\xi,\ 0<\eta,\ \xi\eta<1\}$ sur $]0,1[^2$.
4. Montrer que $I = 4\displaystyle\int_{\{0<\xi,\ 0<\eta,\ \xi\eta<1\}} \frac{\mathrm{d}\xi}{1+\xi^2}\frac{\mathrm{d}\eta}{1+\eta^2}\cdot$.
En déduire (en utilisant le changement de variables $\xi\mapsto\xi^{-1}$, $\eta\mapsto\eta^{-1}$) :
$$I = 2\int_{[0,+\infty[^2} \frac{\mathrm{d}\xi}{1+\xi^2}\frac{\mathrm{d}\eta}{1+\eta^2} \qquad \text{puis} \qquad \zeta(2) = \frac{\pi^2}{6}\cdot$$

III.1.21 On propose dans cet exercice une preuve de l'égalité $\zeta(2) = \frac{\pi^2}{6}$ basée sur la *formule de Fubini* pour les intégrales doubles.

1. On note $\Delta := [0,+\infty[^2$.
Montrer que l'intégrale généralisée $I = \displaystyle\int_\Delta \frac{\mathrm{d}x\,\mathrm{d}y}{(1+y)(1+x^2y)}$ existe et la calculer de deux façons. En déduire $\displaystyle\int_0^{+\infty} \frac{\ln x}{x^2-1}\,\mathrm{d}x = \pi^2/4$.
2. Montrer que $\displaystyle\int_0^1 \frac{\ln x}{x^2-1}\,\mathrm{d}x = \sum_{n\in\mathbb{N}} \frac{1}{(2n+1)^2}$ et conclure.

III.1.22 Cet exercice est une variante du précédent où l'on utilise deux fois la formule de Fubini et un changement de coordonnées polaires. La solution est *beaucoup plus élémentaire*, il n'y a pas d'intégrales généralisées à deux variables (seulement une intégrale généralisée à une variable très simple). C'est, parmi celles proposées, la preuve la plus élémentaire de $\zeta(2) = \pi^2/6$ et celle utilisant le moins d'astuces.

Soient $T := \{(x,y)\in\mathbb{R}^2 \mid 0\leqslant y\leqslant x\leqslant 1$ et $A := \displaystyle\int_T \frac{\mathrm{d}x\,\mathrm{d}y}{(1+x^2)(1+y^2)}$.

1. Calculer A.
2. Montrer que les intégrales généralisées :
$$B := \int_0^{\pi/4} \frac{\ln(2\cos^2\theta)}{\cos 2\theta}\,\mathrm{d}\theta, \quad C := \int_0^{\pi/4} \frac{\ln(2\sin^2\theta)}{\cos 2\theta}\,\mathrm{d}\theta \quad \text{et} \quad D := \int_0^1 \frac{\ln t}{1-t^2}\,\mathrm{d}t.$$
sont convergentes respectivement sur $[0,\pi/4[$, $]0,\pi/4[$ et $[0,1[$.
3. Calculer $B+C$ et $B-C$ en fonction de D. Calculer D, puis B et C.
4. Comparer A et B (on utilisera un passage en coordonnées polaires).
En déduire $\zeta(2) = \pi^2/6$.

Exercices sur la section 2

III.1.23 Soient $a, b \in \mathbb{R}$, $a < b$ et $(J_j)_{j \in [\![1,m]\!]}$ une famille finie d'intervalles ouverts (non vides) de $\mathbb{R}$ recouvrant $[a, b]$. Montrer, en utilisant une jauge convenable, qu'il existe une subdivision $\mathbf{u} := \{x_0, x_1, \ldots, x_n\}$ de $[a, b]$ telle que tout segment $[x_i, x_{i+1}]$ ($i = [\![1, n-1]\!]$) soit contenu dans l'un des J_j.

III.1.24 On définit une fonction $f : [0, 1] \to \mathbb{R}$ par : $f(x) := q$ si $x \in \mathbb{Q}$ et s'écrit $x = p/q$, avec p et q entiers positifs premiers entre eux, et $f(x) = 0$ si x est irrationnel. Montrer que f n'est intégrable au sens de Riemann sur aucun sous-segment (non réduit à un point) de $[0, 1]$ mais qu'elle est $\mathscr{HK}$ -intégrable sur $[0, 1]$.

III.1.25 $**$
Soit $\sum\limits_{k \in \mathbb{N}^*} a_k$ une série *convergente* à termes réels, de somme S.
Soit, pour tout $k \in \mathbb{N}$, $c_k := 1 - 1/2^k$. On définit une fonction $f : [0, 1] \to \mathbb{R}$ par :

$$f : x \mapsto \begin{cases} 2^k a_k & \text{si} \quad x \in [c_{k-1}, c_k[\,,\ k \in \mathbb{N}^* \\ 0 & \text{si} \quad x = 1 \end{cases}.$$

1. Montrer que la fonction f est $\mathscr{HK}$ -intégrable sur le *segment* $[0, 1]$, en utilisant le théorème de Hake, et que $\displaystyle\int_0^1 f = S$
2. Montrer que la fonction f est $\mathscr{HK}$ -intégrable sur le *segment* $[0, 1]$ et que $\displaystyle\int_0^1 f = S$ en construisant des jauges appropriées.
3. On suppose que la série $\sum\limits_{k \in \mathbb{N}^*} |a_k|$ est *divergente*. Montrer que $|f|$ n'est pas $\mathscr{HK}$ -intégrable sur $[0, 1]$. Donner un exemple de fonction $\mathscr{HK}$ -intégrable sur $[0, 1]$ dont la valeur absolue n'est pas $\mathscr{HK}$ -intégrable sur $[0, 1]$.
4. On note $g : [0, 1] \to \mathbb{R}$ la fonction définie par $g := f\, 1_{\mathbb{R} \setminus \mathbb{Q}}$. Montrer que g est $\mathscr{HK}$ -intégrable sur $[0, 1]$. Calculer $\displaystyle\int_0^1 g$. Montrer que, si $a_k \neq 0$, g n'est pas $\mathscr{R}$-intégrable sur $[k-1, k]$. Montrer que, sauf si f est identique nulle, g n'est pas localement intégrable au sens de Riemann sur $[0, 1]$.

III.1.26 Soit $\sum\limits_{k \in \mathbb{N}^*} a_k$ une série *convergente* à termes réels, de somme S. On définit une fonction $f : [0, +\infty[\to \mathbb{R}$ par $f(x) = a_k$ si $x \in [k-1, k[$, avec $k \in \mathbb{N}^*$.
Montrer que f est $\mathscr{HK}$ -intégrable sur $[0, +\infty[$ et calculer $\displaystyle\int_0^{+\infty} f$.

III.1.27 Soient $f : [a, b] \to \mathbb{R}$, une fonction $\mathscr{HK}$ -intégrable sur $[a, b]$ et *minorée*, et $g : [a, b] \to \mathbb{R}$ une fonction *réglée*.
Montrer que fg est $\mathscr{HK}$ -intégrable sur $[a, b]$.

III.1.28 Soit $\sum a_n$ une série *divergente* à termes réels positifs. On suppose que $\lim\limits_{n\to+\infty} a_n = 0$. On définit, pour tout $n \in \mathbb{N}$, $f_n := a_n 1_{[n,n+1]}$.

1. Montrer que la suite $(f_n)_{n\in\mathbb{N}}$ converge simplement vers 0, que les f_n sont $\mathscr{HK}$-intégrables sur $\mathbb{R}$ et que $\lim\limits_{n\to+\infty} \int_{\mathbb{R}} f_n = 0$.
2. Montrer que la suite $(f_n)_{n\in\mathbb{N}}$ n'est pas dominée, c'est-à-dire qu'il n'existe pas de fonction g, intégrable au sens de Henstock-Kurzweil sur $\mathbb{R}$, telle que, pour tout $n \in \mathbb{N}$, $f_n \leqslant g$.
3. Montrer qu'il existe des sous-suites de $(f_n)_{n\in\mathbb{N}}$ vérifiant la condition de domination.

III.1.29 Pour tout $n \in \mathbb{N}^*$, on définit une fonction f_n sur $\mathbb{R}$ par :

$$f_n : x \mapsto \begin{cases} -n & \text{si} \quad -1/n \leqslant x < 0 \\ n & \text{si} \quad 0 < x \leqslant 1/n \\ 0 & \text{sinon.} \end{cases}$$

1. Montrer que la suite $(f_n)_{n\in\mathbb{N}}$ converge simplement vers 0, que les f_n sont intégrables sur $\mathbb{R}$ et que $\lim\limits_{n\to+\infty} \int_{\mathbb{R}} f_n = 0$.
2. Montrer que ni la suite $(f_n)_{n\in\mathbb{N}}$ ni aucune de ses suites extraites ne sont dominées.

Exercices sur la section 3

III.1.30 $**$ Le but de cet exercice est de prouver la proposition 151 de la page 326.
On se propose de montrer, en raisonnant par l'absurde, qu'il existe *une* famille dénombrable $(A_n)_{n\in\mathbb{N}}$ de sous-ensembles *disjoints* de $[0,1]$ telle que :

$$\mu^+\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \neq \sum_{n\in\mathbb{N}} \mu^+(A_n). \tag{41}$$

On suppose donc que, *pour toute* famille dénombrable $(A_n)_{n\in\mathbb{N}}$ de sous-ensembles *disjoints* de $[0,1]$, on a :

$$\mu^+\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu^+(A_n). \tag{42}$$

On introduit sur $[0,1]$ la relation d'équivalence $x\mathcal{R}y$ si, et seulement si, $x - y \in \mathbb{Q}$. Pour tout réel $x \in [0,1]$, on note $Q_x := \{x \in [0,1] \mid x - y \in \mathbb{Q}\}$ sa classe d'équivalence.
Montrer qu'il existe $A \subset [0,1]$ contenant, pour toute classe d'équivalence définie par $\mathcal{R}$, un et un seul élément de la classe. Soit $q : n \in \mathbb{N} \mapsto q_n \in \mathbb{Q} \cap [0,1[$ une *bijection*. On note alors, pour tout $n \in \mathbb{N}$, A_n l'ensemble des $x + q_n$ modulo 1, x appartenant à A.

1. Montrer que, pour tout $n \in \mathbb{N}$, $\mu^+(A_n) = \mu^+(A)$.
2. Montrer que, pour tous $m,\ n \in \mathbb{N}$, $m \neq n$, $A_m \cap A_n = \varnothing$.
3. Montrer que $\bigcup\limits_{n\in\mathbb{N}} A_n = [0,1]$.
4. Conclure.

III.1.31 Montrer que la famille de parties de $\mathbb{R}$ formée des réunions dénombrables d'intervalles réels n'est pas égale à la tribu de Borel.

III.1.32 Pour tout $m \in \mathbb{N}^*$ on divise $I := [0,1]$ en m segments de longueurs égales $I_{m,h} := \left[\frac{h}{m}, \frac{h+1}{m}\right]$, pour $h \in [\![0, m-1]\!]$ et l'on note $\psi_{m,h} := 1_{I_{m,h}}$. On ordonne les $\psi_{m,h}$ par *l'ordre lexicographique* sur les couples ordonnés (m,h), on obtient une suite $(f_k)_{k\in\mathbb{N}}$:

$$f_0 := 1_I,\ f_1 = 1_{[0,1/2]},\ \ f_2 = 1_{[1/2,1]},\ f_3 = 1_{[0,1/3]},\ f_4 = 1_{[1/3,2/3]},\ f_5 = 1_{[2/3,1]}, \ldots$$

1. Montrer que la suite $(f_k)_{k\in\mathbb{N}}$ converge dans $L^1(I)$, calculer sa limite.
2. Montrer que, quel que soit $x \in I$, la suite $\big(f_k(x)\big)_{k\in\mathbb{N}}$ diverge.

III.1.33 Soit $f : \mathbb{R} \to \mathbb{R}$. On suppose que, pour tout $a \in \mathbb{Q}$, $f^{-1}\big(]a, +\infty[\big)$ est mesurable. Montrer que f est mesurable.

III.1.34 Soit $I \subset \mathbb{R}$ un intervalle, on se propose de montrer que l'espace $\mathcal{C}_c(I)$ des fonctions continues à support compact est *dense* dans les espaces de Banach $\big(L^1(I), \|.\|_1\big)$ et $\big(L^2(I), \|.\|_2\big)$. On a $\mathcal{C}_c(\operatorname{Int} I) \subset \mathcal{C}_c(I)$ et $L^1(I) = L^1(\operatorname{Int} I)$, $L^2(I) = L^2(\operatorname{Int} I)$. Il suffit donc de prouver le résultat quand I est *ouvert*. Soit $I \subset \mathbb{R}$ un intervalle ouvert. On note $E_1 \subset L^1(I)$ (respectivement $E_2 \subset L^2(I)$) l'adhérence de $\mathcal{C}_c(I)$ dans $L^1(I)$ (respectivement $L^2(I)$).

1. Montrer que l'espace vectoriel des *fonctions étagées* sur I est dense dans $\big(L^1(I), \|.\|_1\big)$ et $\big(L^2(I), \|.\|_2\big)$.
2. Soient $A \subset \mathbb{R}$ une partie bornée intégrable, $\varepsilon > 0$, $U \subset \mathbb{R}$ un *ouvert borné* et $F \subset \mathbb{R}$ un *fermé* tels que $F \subset A \subset U$ et $\mu(U \setminus F) \leqslant \varepsilon$.
   a) Montrer qu'il existe une fonction *continue à support compact* $\varphi : \mathbb{R} \to [0,1]$ telle que $1_F \leqslant \varphi \leqslant 1_U$.
   b) Montrer que $\displaystyle\int_{\mathbb{R}} \big|1_A - \varphi\big| \leqslant \varepsilon$ et $\displaystyle\int_{\mathbb{R}} \big|1_A - \varphi\big|^2 \leqslant \varepsilon$.
3. Montrer que si $A \subset I$ est une partie intégrable bornée, $1_A \in E_1$ et $1_A \in E_2$. En déduire que si A' est une partie intégrable, $1_{A'} \in E_1$ et $1_{A'} \in E_2$.
4. Montrer que $E_1 = L^1(I)$ et $E_2 = L^2(I)$.

III.1.35 Soient $I \subset \mathbb{R}$ un segment et $f : I \to \mathbb{R}$ une fonction $\mathscr{HK}$-intégrable sur I. Montrer que les deux conditions suivantes sont équivalentes :

(i) f est intégrable au sens de Lebesgue sur I ;

(ii) pour toute fonction $g : I \to \mathbb{R}$ mesurable et bornée sur I, fg est $\mathscr{HK}$-intégrable sur I.

III.1.36 Soit $A \subset \mathbb{R}$ un sous-ensemble *mesurable* de mesure de Lebesgue *strictement positive* (éventuellement infinie). Montrer que la fonction $f : x \mapsto \frac{1_{A(x)}}{1+x^2}$ est intégrable au sens de Lebesgue sur $\mathbb{R}$ et que son intégrale est strictement positive.

III.1.37 Soit $A \subset \mathbb{R}$ un sous-ensemble *mesurable* de mesure de Lebesgue *strictement positive* (éventuellement infinie). Montrer que la fonction $f : x \mapsto \frac{1_A}{1+x^2}$ est intégrable au sens de Lebesgue sur $\mathbb{R}$ et que son intégrale est strictement positive.

III.1.38 $\star\star$

1. Quel est le cardinal de l'ensemble $\mathbb{R}^{\mathbb{N}}$? Quel est le cardinal de l'ensemble $\mathbb{R}^{\mathbb{R}}$?
2. Montrer que la tribu borélienne $\mathcal{B}$ de $\mathbb{R}$ est engendrée par la famille des intervalles ouverts de $\mathbb{R}$ dont l'origine et l'extrémité appartiennent à $\mathbb{Q}$.
3. Soit $\mathcal{A}$ une famille de parties d'un ensemble E. On note $\mathcal{A}_\alpha$ l'ensemble des *unions dénombrables* d'éléments de $\mathcal{A}$, $\mathcal{A}_\beta$ l'ensemble des *intersections dénombrables* d'éléments de $\mathcal{A}$ et $\mathcal{A}_\gamma := \mathcal{A}_\alpha \cup \mathcal{A}_\beta$.

   Soient $\mathcal{U}$ la famille des intervalles ouverts de $\mathbb{R}$ dont l'origine et l'extrémité appartiennent à $\mathbb{Q}$ et $\mathcal{C}$ la famille formée des éléments de $\mathcal{U}$ et de leurs complémentaires. On note $\mathcal{F}_0 = \mathcal{C}_\gamma$ et l'on définit par récurrence des familles (croissantes) $\mathcal{F}_n$ de parties de $\mathbb{R}$ par $\mathcal{F}_{n+1} := \big(\mathcal{F}_n\big)_\gamma$ $(n \geqslant 1)$.

   a) Montrer que le cardinal de $\bigcup\limits_{n\in\mathbb{N}} \mathcal{F}_n$ est égal à celui de $\mathbb{R}$.

   b) On *admet* que la suite $(\mathcal{F}_n)_{n\in\mathbb{N}}$ est *strictement croissante*. Montrer que le sous-ensemble $\bigcup\limits_{n\in\mathbb{N}} \mathcal{F}_n$ de $\mathcal{B}$ est *distinct* de $\mathcal{B}$.
4. On admet[6] que le cardinal d'une tribu engendrée par un ensemble dénombrable est égal à celui de $\mathbb{R}$. Quel est le cardinal de $\mathcal{B}$?
5. Montrer que le cardinal de l'ensemble des parties mesurables au sens de Lebesgue de $\mathbb{R}$ est égal à celui de l'ensemble des parties de $\mathbb{R}$.
6. Montrer qu'il existe des parties mesurables au sens de Lebesgue non boréliennes. Quel est le cardinal de l'ensemble de ces parties ?

III.1.39

On utilise les résultats de l'exercice II.3.6 de la page 725 du livre [L2] dont on reprend les notations. Soit $K \subset [0,1]$ le compact de Smith-Volterra-Cantor.

1. Soit $N \subset [0,1]$ un ensemble négligeable.

   Montrer que $K \setminus N \neq \varnothing$ et que, pour tout $x \in K \setminus N$ et tout $\varepsilon > 0$, l'ensemble
   $$]x-\varepsilon, x+\varepsilon[\cap \big([0,1] \setminus (K \cup N)\big)$$
   est non vide.
2. Soit $f : [0,1] \to \mathbb{R}$. On suppose que f est presque partout égale à 1_K. Montrer que f est discontinue sur $K \setminus N$. En déduire que f n'est pas intégrable au sens de Riemann sur $[0,1]$.
3. Montrer qu'il existe une fonction g de carré sommable sur $[0,1]$ telle qu'il n'existe pas de fonction réglée h sur $[0,1]$ telle que $g-h$ soit négligeable.

[6] Cf. par exemple : D. Revuz, *Intégration*, Hermann, Paris

III.1.40 *Mesure de comptage.*
Soient Ω un ensemble. On le munit de la tribu formée de l'ensemble de ses parties $\mathcal{P}(\Omega)$. Pour tout $A \subset \Omega$, on définit $\mu(A)$ par $\mu(A) := \operatorname{card} A$ si A est *fini* et $\mu(A) := \infty$ si A est *infini*.

1. Montrer que μ est une mesure sur $\big(\Omega, \mathcal{P}(\Omega)\big)$.
2. Définir, par analogie avec le cas de la mesure de Lebesgue, une notion d'intégrale pour μ. Caractériser les fonctions intégrables dans le cas où $\Omega = \mathbb{N}$.

III.1.41 Soit $\sum f_{n,m}$ une série double de fonctions *absolument intégrables* au sens de Henstock-Kurzweil sur un intervalle I. On suppose que la série numérique double $\sum \left(\int_I |f_{n,m}| \right)$ est convergente. Montrer qu'alors la série double $\sum f_{n,m}$ converge *absolument* presque partout sur I vers une fonction F qui est *absolument intégrable* au sens de Henstock-Kurzweil sur I et l'on a :

$$\sum_{n\in\mathbb{N}} \sum_{m\in\mathbb{N}} \left(\int_I f_{n,m} \right) = \int_I F = \int_I \left(\sum_{n\in\mathbb{N}} \sum_{m\in\mathbb{N}} f_{n,m} \right).$$

III.1.42 $**$
Soit $A \subset \mathbb{R}^2$. Pour tout $x \in \mathbb{R}$, on note $A_x := \{y \in \mathbb{R} \mid (x,y) \in A\}$.
Montrer que, si A est négligeable dans $\mathbb{R}^2$, alors, pour presque tout $x \in \mathbb{R}$, A_x est négligeable dans $\mathbb{R}$. Exprimer le résultat sous forme d'une formule du type de Fubini.

Exercices sur la section 4

III.1.43 On se propose de démontrer la formule de changement de variables (théorème 189 de la page 353) dans le cadre de l'intégrale de Riemann généralisée. Dans cet exercice on se limite donc à utiliser cette théorie.

1. Soit $L : \mathbb{R}^n \to \mathbb{R}^n$ une transvection.
   a) Soit $\mathbf{P}$ un pavé fermé de $\mathbb{R}^n$. Montrer que $L(\mathbf{P})$ est cubable et que
   $$\operatorname{vol}\big(L(\mathbf{P})\big) = \operatorname{vol}(\mathbf{P}).$$
   b) Soit A un sous-ensemble pavable de $\mathbb{R}^n$.
   Montrer que $L(A)$ est cubable et que $\operatorname{vol}(L(A)) = \operatorname{vol}(A)$.
   c) Soient $U \subset \mathbb{R}^n$ un ouvert borné et $(K_p)_{p\in\mathbb{N}}$ un pavage exhaustif de U.
   Montrer que $\big(L(K_p)\big)_{p\in\mathbb{N}}$ est une suite exhaustive de compacts de $L(U)$.
   En déduire que $\operatorname{vol}\big(L(U)\big) = \operatorname{vol}(U)$.
2. Soit $L : \mathbb{R}^n \to \mathbb{R}^n$ une *application linéaire inversible* et $U \subset \mathbb{R}^n$ un *ouvert borné*. Montrer que $\operatorname{vol}\big(L(U)\big) = \big|\det(L)\big| \operatorname{vol}(U)$.
3. Terminer la preuve du théorème de changement de variables dans le cadre de l'intégrale de Riemann généralisée.

III.1.44 Soient $n,\ m \in \mathbb{N}^*$ On note $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$ des ouverts bornés et $K \subset \mathbb{R}^n$, $L \subset \mathbb{R}^m$ des compacts. Calculer $\operatorname{vol}(U \times V)$ et $\mu(K \times L)$.

III.1.45 Soit $K \subset \mathbb{R}$ un compact non cubable d'intérieur vide (*cf.* l'exercice type corrigé III.1.0). Montrer que K^n est un compact non cubable de $\mathbb{R}^n$.

Introduction aux équations aux dérivées partielles

III.2

1 Généralités

1.1 Définitions et premiers exemples

On a vu dans le livre [L2] des équations différentielles ordinaires. Ce sont des équations qui expriment une relation entre une fonction inconnue $u(t)$ d'une variable réelle $t \in \mathbb{R}$ et un certain nombre de ses dérivées $\frac{\mathrm{d}u}{\mathrm{d}t}$, $\frac{\mathrm{d}^2u}{\mathrm{d}t^2}$,...

On va examiner dans ce module des équations aux dérivées partielles (EDP). Ce sont des équations qui expriment une relation entre une fonction inconnue $u(\mathbf{x})$ de deux ou plusieurs variables réelles $\mathbf{x} := (x_1, \ldots, x_n) \in \mathbb{R}^n$ et un certain nombre de ses dérivées partielles $\frac{\partial u}{\partial x_i}$, $\frac{\partial^2 u}{\partial x_i \partial x_j}$,...

Soit n un entier strictement positif. On commence par introduire une notation standard dans la théorie des EDP. On note ∇ le vecteur des dérivées partielles :

$$\nabla := \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right) .$$

Cela veut dire que, pour toute fonction différentiable à valeurs réelles $\varphi : \mathbb{R}^n \to \mathbb{R}$, la fonction $\nabla\varphi : \mathbb{R}^n \to \mathbb{R}^n$ est à valeurs dans $\mathbb{R}^n$ et donnée par :

$$\nabla\varphi(\mathbf{x}) = \left(\frac{\partial \varphi}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial \varphi}{\partial x_n}(\mathbf{x}) \right) , \quad \mathbf{x} \in \mathbb{R}^n .$$

Dans le module II.5 du livre [L2], on avait noté $\nabla\varphi(\mathbf{x}) = D_{\mathbf{x}}\varphi$, mais la notation ∇ est la plus couramment utilisée dans le domaine des EDP. Pour alléger les notations, on note les dérivées partielles par des indices, de sorte que :

$$\nabla\varphi(\mathbf{x}) = (\varphi_{x_1}(\mathbf{x}), \ldots, \varphi_{x_n}(\mathbf{x})) .$$

Définition 1.
Soient n un entier strictement positif, Ω un ouvert de $\mathbb{R}^n$ et $F : \mathbb{R}^n \times \mathbb{R} \times \Omega \to \mathbb{R}$. Une expression de la forme :

$$F(\nabla u, u, \mathbf{x}) = 0, \quad \mathbf{x} \in \Omega\,, \tag{1}$$

s'appelle une *EDP du premier ordre*.
On dit que la fonction $u \in \mathcal{C}^1(\Omega, \mathbb{R})$ est *solution de l'EDP* si, pour tout $\mathbf{x} \in \Omega$, $F(\nabla u(\mathbf{x}), u(\mathbf{x}), \mathbf{x}) = 0$.

Les EDP modélisent très souvent des problèmes issus de la physique, décrivant des systèmes dynamiques, des problèmes d'évolution, de transport, de diffusion, de propagation d'ondes,... Dans ce cas, la variable $\mathbf{x} = (x_1, \ldots, x_n)$ contient une coordonnée décrivant la variable temporelle, et entre une et trois coordonnées décrivant les variables spatiales. C'est le cas dans l'exemple qui suit, qui modélise un problème d'évolution temporelle dans un milieu uni-dimensionnel.

Exemple. Si c est un réel fixé et $f : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ est une fonction donnée, le problème de transport consiste à chercher $u : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ qui satisfait pour tous $t > 0$, $x \in \mathbb{R}$:

$$u_t(t, x) + cu_x(t, x) = f(t, x)\,. \tag{2}$$

Cette équation est bien de la forme (1). Ici $n = 2$, $(x_1, x_2) = (t, x)$, $\Omega =]0, +\infty[\times \mathbb{R}$ et :

$$F(\mathbf{v}, u, (x_1, x_2)) = v_1 + cv_2 - f(x_1, x_2)\,.$$

L'équation (2) peut être utilisée pour modéliser l'évolution de la concentration $u(t, x)$ d'un polluant dans un fleuve à la position x à l'instant t. Dans ce cadre c est la vitesse du fleuve et $f(t, x)$ est un terme source, qui décrit le rejet de polluant à la position x et à l'instant t. En général, on spécifie en plus de l'équation d'évolution une condition initiale, c'est-à-dire le profil de la solution à l'instant $t = 0$, la concentration initiale de polluant. Si le fleuve est propre à l'instant $t = 0$, on rajoute la condition $u(t = 0, x) = 0$ pour tout $x \in \mathbb{R}$.

On peut généraliser la notion d'EDP à des équations faisant intervenir des dérivées d'ordre supérieur.

Définition 2. Soient Ω un ouvert de $\mathbb{R}^n$ et $F : M_n(\mathbb{R}) \times \mathbb{R}^n \times \mathbb{R} \times \Omega \to \mathbb{R}$. Une expression de la forme :

$$F(\nabla^2 u, \nabla u, u, \mathbf{x}) = 0, \quad \mathbf{x} \in \Omega\,, \tag{3}$$

s'appelle une *EDP du second ordre*.

La notation ∇^2 désigne la matrice des dérivées partielles d'ordre 2 :

$$\nabla^2 := \left(\frac{\partial^2}{\partial x_i \partial x_j} \right)_{i,j \in [\![1,n]\!]}\,.$$

Cela veut dire que, pour toute fonction de classe $\mathcal{C}^2$ à valeurs réelles $\varphi : \mathbb{R}^n \to \mathbb{R}$, la fonction $\nabla^2\varphi : \mathbb{R}^n \to M_n(\mathbb{R})$ est à valeurs dans l'espace des matrices réelles de taille $n \times n$:

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad \nabla^2\varphi(\mathbf{x}) = \left(\frac{\partial^2 \varphi}{\partial x_i \partial x_j}(\mathbf{x}) \right)_{i,j \in [\![1,n]\!]} = \left(\varphi_{x_i x_j}(\mathbf{x}) \right)_{i,j \in [\![1,n]\!]}\,.$$

Comme pour les dérivées partielles d'ordre 1, pour alléger les notations, on note les dérivées partielles par des doubles indices. Enfin, on rappelle que, par le théorème de Schwarz (voir le module II.5 du livre [L2]), on a pour tous $i, j \in [\![1, n]\!]$:

$$\varphi_{x_i x_j}(\mathbf{x}) = \varphi_{x_j x_i}(\mathbf{x}) \,.$$

Exemple. L'équation des ondes uni-dimensionnelle est :

$$u_{tt} - c^2 u_{xx} = 0 \,, \quad (t, x) \in]0, +\infty[\times]0, L[\,.$$

Cette équation peut être utilisée pour décrire la propagation d'une onde (vibration) sur une corde. Il peut s'agir par exemple d'une corde de piano (de longueur L) frappée à un instant donné (disons $t = 0$). Dans ce cas $u(t, x)$ modélise le déplacement de la corde autour de sa position d'équilibre. La constante strictement positive c, dite vitesse du son, est liée à la tension T de la corde et à sa masse linéaire m par la relation $c = \sqrt{T/m}$. Cette équation est bien de la forme (3). Ici $n = 2$, $(x_1, x_2) = (t, x)$, $\Omega =]0, +\infty[\times]0, L[$ et :

$$F(\mathbf{W}, \mathbf{v}, u, (x_1, x_2)) = W_{11} - c^2 W_{22} \,.$$

La modélisation du problème, en plus de l'équation d'évolution, peut fournir d'autres indications. Par exemple, la corde est fixée à ses deux extrémités 0 et L, ce qui impose que :

$$\forall t \geqslant 0 \,, \quad u(t, 0) = u(t, L) = 0 \,.$$

On peut aussi connaître le déplacement initial de la corde (résultat du coup de marteau) :

$$\forall x \in [0, L] \,, \quad u(0, x) = f(x) \,,$$

où $f : [0, L] \to \mathbb{R}$ est une fonction donnée. On peut aussi savoir que la vitesse initiale de chaque point de la corde est nulle :

$$\forall x \in [0, L] \,, \quad u_t(0, x) = 0 \,.$$

Bien sûr, on peut encore généraliser la notion d'EDP et introduire des EDP d'ordre k. L'ordre d'une EDP est définie de manière similaire à celui d'une équation différentielle ordinaire, à savoir l'ordre de la plus haute dérivée apparaissant dans l'équation.
Résoudre l'EDP (1) ou (3) revient à déterminer toutes les fonctions u qui sont solutions, ou bien seulement les fonctions u qui satisfont (1) ou (3) ainsi que des conditions initiales et/ou des conditions spécifiées sur le bord du domaine Ω ou une partie du bord. Ces conditions sont en général fournies par la modélisation physique (ou économique,...) du problème. La détermination des solutions consiste, idéalement, à trouver des expressions explicites de toutes les solutions. À défaut, on cherchera des propriétés plus qualitatives des solutions, leur existence, unicité, régularité,...
Dans ce module, on considérera essentiellement des EDP linéaires. Dans ce cas, la fonction F qui apparaît dans (1) ou (3) est linéaire en u, ∇u et $\nabla^2 u$.
Un système d'EDP est, grossièrement, une famille d'EDP couplées. La définition précise est la suivante :

Définition 3. Soient n, m deux entiers strictement positifs, Ω un ouvert de $\mathbb{R}^n$ et $\mathbf{F} : M_{n,m}(\mathbb{R}) \times \mathbb{R}^m \times \Omega \to \mathbb{R}^m$. Une expression de la forme :

$$\mathbf{F}(\nabla \mathbf{u}, \mathbf{u}, \mathbf{x}) = 0, \quad \mathbf{x} \in \Omega\,, \tag{4}$$

s'appelle un *système d'EDP du premier ordre*.
On dit que la fonction $\mathbf{u} \in \mathcal{C}^1(\Omega, \mathbb{R}^m)$ est *solution de ce système d'EDP* si, pour tout $\mathbf{x} \in \Omega$, $\mathbf{F}(\nabla \mathbf{u}(\mathbf{x}), \mathbf{u}(\mathbf{x}), \mathbf{x}) = 0$. On note ici $\nabla \mathbf{u}$ la matrice $(\frac{\partial u_j}{\partial x_i})_{i \in [\![1,n]\!], j \in [\![1,m]\!]}$.
On appelle $x_1, \ldots, x_n$ les *variables indépendantes du problème*.
On appelle $u_1, \ldots, u_m$ les *variables dépendantes du problème*.

Ici, on a supposé que le nombre d'équations (la dimension de l'espace d'arrivée de $\mathbf{F}$) était égal au nombre de fonctions inconnues (la dimension de la solution $\mathbf{u}$). C'est le cas le plus courant, mais on peut bien sûr considérer des situations plus générales.

1.2 Problèmes bien posés

Problèmes d'évolution. Les EDP modélisent très souvent des problèmes d'évolution, comme dans les deux exemples présentés précédemment. Dans ces situations, les variables de l'EDP ont une interprétation physique. Une des coordonnées est le temps t et les autres coordonnées sont des coordonnées spatiales $\mathbf{x} = (x_1, \ldots, x_d)$. Le domaine est typiquement de la forme $(t, \mathbf{x}) \in]0, T[\times \Omega$ ou $]0, +\infty[\times \Omega$, avec Ω un ouvert de $\mathbb{R}^d$ et d est la dimension des coordonnées spatiales ($d = 3$ dans l'espace cartésien usuel, ou $d = 1$ ou 2 pour des problèmes se déroulant sur une ligne ou dans un plan). La modélisation physique peut fournir des conditions supplémentaires que doivent vérifier les solutions de l'EDP, et l'on peut distinguer deux types de conditions.
Conditions initiales. Dans le cas d'équations d'évolution, on connaît souvent l'état du système à l'instant initial, et l'on souhaite calculer l'évolution du système au cours du temps. Cela veut dire que l'on connaît $u(t_0, \mathbf{x})$ et / ou $u_t(t_0, \mathbf{x})$ pour tout $\mathbf{x} \in \Omega$ à un certain instant t_0 (souvent $t_0 = 0$), et l'on souhaite calculer $u(t, \mathbf{x})$ pour $t > t_0$ et $\mathbf{x} \in \Omega$. De telles conditions initiales sont aussi appelées conditions de Cauchy.
Conditions aux bords. On peut avoir affaire à des équations d'évolution se déroulant dans l'espace $\mathbb{R}^d$ tout entier, mais il peut aussi arriver que le domaine spatial ouvert Ω soit borné. Le bord du domaine est défini par $\partial\Omega = \overline{\Omega} \backslash \Omega$. Il arrive alors fréquemment que le problème physique fournisse des indications sur le comportement de la solution cherchée sur le bord $\partial\Omega$, ou sur une partie Γ du bord $\partial\Omega$. On cherche alors la ou les solutions de l'EDP satisfaisant ces conditions aux bords. On peut distinguer deux grands types de conditions aux bords (mais on peut en imaginer d'autres) :

- Conditions de Dirichlet : les valeurs de la solution u sur le bord $\partial\Omega$ ou sur une partie Γ du bord $\partial\Omega$ sont spécifiées. C'est ce type de conditions aux bords que l'on manipulera dans ce module. On écrira une telle condition initiale sous la forme :

  $$u(\mathbf{x}) = f(\mathbf{x})\,, \quad \mathbf{x} \in \Gamma\,,$$

  pour une fonction f donnée.
- Conditions de Neumann : les valeurs de la dérivée normale de la solution $\frac{\partial u}{\partial n}$ sur le bord $\partial\Omega$ ou sur une partie Γ du bord $\partial\Omega$ sont spécifiées. On n'abordera pas ce type de

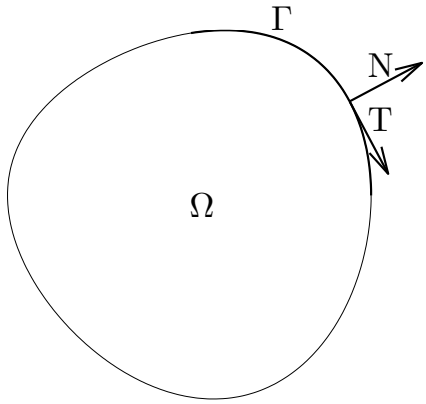


FIGURE III.2.1 – Exemple de configuration pour le domaine $\Omega \subset \mathbb{R}^2$, la partie Γ du bord $\partial\Omega$, un vecteur tangent $\mathbf{T}$ et le vecteur normal orienté vers l'extérieur $\mathbf{N}$ en un point du bord $\partial\Omega$.

conditions aux bords dans ce module, qui nécessite de manipuler des notions abordées dans le module de géométrie pour des surfaces lisses en dimension 3 par exemple, *cf.* module II.1, section 2.3 de la page 146. La droite normale à une surface en un point est la droite orthogonale au plan tangent en ce point. Les vecteurs directeurs de cette droite sont appelés vecteurs normaux à la surface. Pour les surfaces fermées comme $\partial\Omega$ on peut particulariser un vecteur normal unitaire, vecteur de norme un et orienté vers l'extérieur (voir Figure III.2.1). Notons $\mathbf{N}$ le vecteur normal unitaire orienté vers l'extérieur du domaine Ω en un point de son bord $\partial\Omega$. La dérivée normale en ce point est définie par $\frac{\partial u}{\partial n} := \mathbf{N} \cdot \nabla u$.

Problèmes bien posés

On dira qu'un problème est bien posé s'il satisfait les trois propriétés suivantes :

Existence : il existe au moins une solution.

Unicité : il y a au plus une solution.

Stabilité : la solution dépend continûment des conditions initiales / aux bords. En d'autres mots, un petit changement dans les conditions initiales / aux bords produit un petit changement de la solution.

Ce dernier critère est énoncé ici de manière vague, mais c'est une propriété critique si l'on songe aux applications, physiques ou autres, où les conditions initiales / aux bords sont très souvent connues avec une certaine imprécision. On verra un exemple à la fin de cette section.

Les propriétés d'existence et d'unicité dépendent de la classe de fonctions dans laquelle on cherche les solutions. Si l'on cherche la solution dans une classe trop petite, en demandant des conditions de régularité fortes, on peut très bien ne pas trouver de solution. Il n'existe pas de théorème général d'existence et d'unicité pour les EDP. Personne n'envisage d'ailleurs qu'il existe un tel théorème, eu égard à la richesse et la variété des phénomènes que peuvent modéliser des EDP. Le théorème de Cauchy-Kowalevski est le résultat principal d'existence et d'unicité locales pour des solutions analytiques à des EDP analytiques

(i.e. à coefficients analytiques). Un cas particulier de ce théorème fut d'abord montré par Augustin Cauchy, et le résultat complet est dû à Sofia Kovalevskaïa. Depuis, de nombreuses généralisations et extensions ont été proposées. Le théorème d'existence de Riquier est une des principales généralisations.
On commence par étendre la notion de « fonction analytique » introduites dans le module du même nom du livre [L2].

Définition 4.
1. Soit n un entier strictement positif. Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est dite *analytique* en $\mathbf{x}_0 \in \mathbb{R}^n$ s'il existe $r > 0$ et des constantes $f_{j_1,\dots,j_n} \in \mathbb{R}$, $j_1, \dots, j_n \in \mathbb{N}$, telles que :

$$\forall \mathbf{x} \in \mathbb{R}^n \mid |\mathbf{x}| < r\,, \quad f(\mathbf{x}_0 + \mathbf{x}) = \sum_{j_1,\dots,j_n=0}^{+\infty} f_{j_1,\dots,j_n} x_1^{j_1} \cdots x_n^{j_n}\,.$$

La fonction est dite *analytique* si elle est analytique en tout point.
2. Soient n et m deux entiers strictement positifs. Une fonction $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ est dite *analytique* si les fonctions coordonnées f_l, $l = 1, \dots, m$, sont analytiques.

Si $f : \mathbb{R}^n \to \mathbb{R}$ est analytique en $\mathbf{x}_0$, alors elle est indéfiniment différentiable en $\mathbf{x}_0$ et l'on peut identifier les coefficients $f_{j_1,\dots,j_n}$ du développement en série entière :

$$f_{j_1,\dots,j_n} = \frac{1}{j_1! \cdots j_n!} \frac{\partial^{j_1+\dots+j_n} f}{\partial x_1^{j_1} \cdots \partial x_n^{j_n}}(\mathbf{x}_0)\,.$$

Considérons un système d'EDP du premier ordre comprenant m variables dépendantes $\mathbf{u} := (u_1, \dots, u_m)$ et n variables indépendantes $(t, \mathbf{x})$, $\mathbf{x} := (x_1, \dots, x_{n-1})$:

$$\mathbf{u}_t = \sum_{j=1}^{n-1} \mathbf{C}_j(\mathbf{u}, \mathbf{x}) \mathbf{u}_{x_j} + \mathbf{c}(\mathbf{u}, \mathbf{x})\,, \qquad (t, \mathbf{x}) \in]0, +\infty[\times \mathbb{R}^{n-1}\,, \tag{5}$$

où $\mathbf{C}_j : \mathbb{R}^m \times \mathbb{R}^{n-1} \to M_m(\mathbb{R})$ et $\mathbf{c} : \mathbb{R}^m \times \mathbb{R}^{n-1} \to \mathbb{R}^m$ sont des fonctions analytiques. Noter que les fonctions $\mathbf{C}_j$ et $\mathbf{c}$ dans l'EDP (5) ne dépendent pas du temps t. On peut toujours se ramener à cette situation, quitte à introduire la nouvelle coordonnée $u_{m+1} = t$ au vecteur $\mathbf{u}$. Le problème de Cauchy constitué de l'EDP précédente et de la condition initiale :

$$\mathbf{u}(t = 0, \mathbf{x}) = \mathbf{u}_0(\mathbf{x})\,, \quad \mathbf{x} \in \mathbb{R}^{n-1}\,, \tag{6}$$

où $\mathbf{u}_0 : \mathbb{R}^{n-1} \to \mathbb{R}^m$ est une fonction analytique, peut se ramener à un problème de Cauchy avec condition initiale nulle. Il suffit de considérer les nouvelles variables dépendantes :

$$\widetilde{\mathbf{u}}(t, \mathbf{x}) := \mathbf{u}(t, \mathbf{x}) - \mathbf{u}_0(\mathbf{x})\,,$$

et les nouvelles fonctions :

$$\widetilde{\mathbf{C}}_j(\widetilde{\mathbf{u}}, \mathbf{x}) := \mathbf{C}_j(\widetilde{\mathbf{u}} + \mathbf{u}_0(\mathbf{x}), \mathbf{x})\,, \qquad \widetilde{\mathbf{c}}(\widetilde{\mathbf{u}}, \mathbf{x}) := \mathbf{c}(\widetilde{\mathbf{u}} + \mathbf{u}_0(\mathbf{x}), \mathbf{x})\,,$$

qui sont toujours analytiques et indépendantes de t. On vérifie que $\widetilde{\mathbf{u}}$ est solution de l'EDP :

$$\widetilde{\mathbf{u}}_t = \sum_{j=1}^{n-1} \widetilde{\mathbf{C}}_j(\widetilde{\mathbf{u}}, \mathbf{x}) \widetilde{\mathbf{u}}_{x_j} + \widetilde{\mathbf{c}}(\widetilde{\mathbf{u}}, \mathbf{x})\,, \qquad (t, \mathbf{x}) \in]0, +\infty[\times \mathbb{R}^{n-1}\,,$$

avec la condition initiale nulle $\widetilde{\mathbf{u}}(t=0,\mathbf{x})=0\ \forall \mathbf{x}\in\mathbb{R}^{n-1}$ si, et seulement si, $\mathbf{u}$ est solution de l'EDP (5) avec la condition initiale (6). La formulation du théorème de Cauchy-Kowalewski qui suit se concentre sur les cas de conditions initiales nulles.

Théorème 1. On note B_r^n la boule ouverte de $\mathbb{R}^n$ de centre 0 et de rayon r. Considérons l'EDP :

$$\mathbf{u}_t=\sum_{j=1}^{n-1}\mathbf{C}_j(\mathbf{u},\mathbf{x})\mathbf{u}_{x_j}+\mathbf{c}(\mathbf{u},\mathbf{x})\,,\quad (t,\mathbf{x})\in B_r^n\,, \tag{7}$$

avec la condition de Cauchy en $t=0$:

$$\mathbf{u}(t=0,\mathbf{x})=0\,,\quad \mathbf{x}\in B_r^{n-1}\,. \tag{8}$$

Supposons que les fonctions $\mathbf{C}_j:\mathbb{R}^m\times\mathbb{R}^{n-1}\to M_m(\mathbb{R})$ et $\mathbf{c}:\mathbb{R}^m\times\mathbb{R}^{n-1}\to\mathbb{R}^m$ sont analytiques. Alors il existe $r>0$ tel que le problème (7-8) admet une unique fonction analytique $\mathbf{u}:B_r^n\mapsto\mathbb{R}^m$.

Démonstration. Nous donnons l'idée de la preuve. Elle consiste en deux étapes : 1) identifier les termes du développement en série entière d'une solution de l'EDP et 2) montrer que cette série converge sur une boule. Pour le premier point, si l'on suppose que la fonction $\mathbf{u}$ se développe comme :

$$\mathbf{u}(t,\mathbf{x})=\sum_{i,j_1,\dots,j_{n-1}=0}^{+\infty}\mathbf{a}_{i,j_1,\dots,j_{n-1}}t^i x_1^{j_1}\cdots x_{n-1}^{j_{n-1}}\,,$$

où les coefficients vérifient :

$$\mathbf{a}_{i,j_1,\dots,j_{n-1}}=\frac{1}{i!j_1!\cdots j_{n-1}!}\frac{\partial^{i+j_1+\cdots+j_{n-1}}\mathbf{u}}{\partial t^i\partial x_1^{j_1}\cdots\partial x_{n-1}^{j_{n-1}}}(0,0)\in\mathbb{R}^m\,,$$

alors ces coefficients peuvent être identifiés par récurrence en développant de manière formelle la condition initiale (8) et les deux membres de l'EDP (7). En effet, il est facile de voir que, si une solution analytique du problème existe, alors on doit avoir :

$$\mathbf{a}_{0,0,\dots,0}=0\,,\qquad \mathbf{a}_{0,j_1,\dots,j_n}=0\,,$$

car $\mathbf{u}(t=0,\mathbf{x})=0$ d'après la condition initiale (8). En considérant l'EDP (7), on voit aussi que :

$$\mathbf{a}_{1,0,\dots,0}=\mathbf{c}(0,0)\,,$$

car $\mathbf{u}(t=0,0)=0$ et $\mathbf{u}_{x_j}(t=0,0)=0$. En dérivant plusieurs fois en les coordonnées de $\mathbf{x}$ l'EDP (7), on obtient que les coefficients $\mathbf{a}_{i,j_1,\dots,j_{n-1}}$ avec $i=1$ sont nécessairement de la forme :

$$\mathbf{a}_{1,j_1,\dots,j_n}=\frac{\partial^{j_1+\cdots+j_{n-1}}\mathbf{c}}{\partial x_1^{j_1}\cdots\partial x_{n-1}^{j_{n-1}}}(0,0)\,.$$

En dérivant ensuite en t l'EDP (7), on obtient que les coefficients $\mathbf{a}_{i,j_1,\dots,j_{n-1}}$ sont des polynômes en les dérivées de $\mathbf{C}_j$ et $\mathbf{c}$ en $(0,0)$. Ainsi, si la solution se développe sous la forme

de série entière, alors tous les termes du développement peuvent être identifiés. D'une part, cela prouve l'unicité d'une solution analytique. D'autre part, une fois ce travail d'identification effectué, le fin de la démonstration de l'existence d'une solution analytique consiste à montrer que la série ainsi obtenue converge effectivement sur une boule (en utilisant des techniques de majoration fine que l'on ne développera pas dans ce module). ■

Le théorème de Cauchy-Kowalewski est un résultat d'existence et d'unicité locale, il ne dit rien sur l'existence d'une solution globale dans tout le domaine, et il ne s'applique pas pour de nombreuses EDP qui n'admettent pas de solution analytique, mais qui admettent cependant des solutions dans des classes de fonctions de type $\mathcal{C}^k$ ou $\mathcal{C}^\infty$. Par exemple, il peut s'appliquer à l'équation de transport que l'on étudiera dans la section 2.1, à l'équation des ondes que l'on étudiera dans la section 4.1 ou à l'équation de Laplace que l'on va voir ci-dessous, mais il ne s'applique pas à l'équation de la chaleur que l'on étudiera dans la section 4.4.

Exemple. Considérons l'équation des ondes :

$$u_{tt} - u_{xx} = 0\,, \qquad (t,x) \in \,]0,+\infty[\, \times \mathbb{R}\,,$$

avec des conditions de Cauchy du type $u(t=0,x) = f(x)$, $u_t(t=0,x) = g(x)$, $x \in \mathbb{R}$, avec f et g analytiques.

On peut metttre cette équation sous la forme (5-6). En effet, en posant

$$\mathbf{v} = \begin{pmatrix} u \\ u_t \\ u_x \end{pmatrix}, \qquad \mathbf{v}_0(x) = \begin{pmatrix} f(x) \\ g(x) \\ f_x(x) \end{pmatrix},$$

on trouve que l'équation des ondes se met sous la forme :

$$\mathbf{v}_t = \begin{pmatrix} u_t \\ u_{tt} \\ u_{tx} \end{pmatrix} = \begin{pmatrix} u_t \\ u_{xx} \\ u_{xt} \end{pmatrix} = \frac{\partial}{\partial x} \begin{pmatrix} 0 \\ u_x \\ u_t \end{pmatrix} + \begin{pmatrix} u_t \\ 0 \\ 0 \end{pmatrix},$$

c'est-à-dire :

$$\mathbf{v}_t = \mathbf{C}\mathbf{v}_x + \mathbf{c}(\mathbf{v})\,,$$

avec la condition initiale $\mathbf{v}(t=0,x) = \mathbf{v}_0(x)$, $x \in \mathbb{R}$ et

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{c}(\mathbf{v}) = \begin{pmatrix} v_2 \\ 0 \\ 0 \end{pmatrix}.$$

On peut aussi metttre l'équation des ondes sous la forme (7) avec condition de Cauchy nulle en $t = 0$. En effet, en posant

$$\widetilde{\mathbf{v}} = \begin{pmatrix} u - f \\ u_t - g \\ u_x - f_x \end{pmatrix},$$

on trouve que l'équation des ondes se met sous la forme :

$$\widetilde{\mathbf{v}}_t = \begin{pmatrix} u_t \\ u_{tt} \\ u_{tx} \end{pmatrix} = \begin{pmatrix} u_t \\ u_{xx} \\ u_{xt} \end{pmatrix} = \frac{\partial}{\partial x} \begin{pmatrix} 0 \\ u_x - f_x \\ u_t - g \end{pmatrix} + \begin{pmatrix} u_t - g \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} g \\ f_{xx} \\ g_x \end{pmatrix},$$

c'est-à-dire :

$$\widetilde{\mathbf{v}}_t = \widetilde{\mathbf{C}}\widetilde{\mathbf{v}}_x + \widetilde{\mathbf{c}}(\widetilde{\mathbf{v}}, x)\,,$$

avec $\widetilde{\mathbf{v}}(t=0,x) = 0$, $x \in \mathbb{R}$ et

$$\widetilde{\mathbf{C}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad \widetilde{\mathbf{c}}(\widetilde{\mathbf{v}}, x) = \begin{pmatrix} \widetilde{v}_2 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} g(x) \\ f_{xx}(x) \\ g_x(x) \end{pmatrix}.$$

Le théorème de Cauchy-Kowalewski indique qu'il existe des solutions analytiques de l'équation des ondes au voisinage de $(0,0)$.
L'équation de Laplace :

$$u_{tt} + u_{xx} = 0\,, \qquad (t,x) \in \,]0,+\infty[\, \times \mathbb{R}\,,$$

avec des conditions de Cauchy en $t = 0$ se traite exactement de la même façon, il suffit juste de changer quelques signes dans les arguments développés ci-dessus pour l'équation des ondes.

Exemple. En ce qui concerne l'équation de la chaleur,

$$u_t = u_{xx}\,, \qquad (t,x) \in \,]0,+\infty[\, \times \mathbb{R}\,,$$

Kowalewski a proposé des conditions initiales (par exemple, $u(t=0,x) = 1/(1+x^2)$) pour lesquelles la solution u admet un développement en série formelle unique autour de $(t,x) = (0,0)$, mais pour lesquelles ce développement diverge pour tout $t \neq 0$, ce qui veut dire qu'il n'y a pas de solution analytique dans un voisinage de $(t,x) = (0,0)$. On verra dans la section 4.4 que l'équation de la chaleur admet cependant des solutions de classe $\mathcal{C}^\infty$.

On peut bien sûr généraliser l'énoncé du théorème à des situations impliquant des dérivées temporelles d'ordre plus élevé. Il suffit pour cela d'introduire de nouvelles variables dépendantes et d'augmenter la taille du système, de manière similaire à ce qui est fait avec les équations différentielles ordinaires.
On termine cette section par un exemple de problème qui admet une unique solution analytique, mais qui est cependant mal posé dans le sens où un changement infinitésimal des conditions initiales peut complètement modifier la solution. Cet exemple est dû à Hadamard.

Exemple. Considérons l'équation de Laplace en deux dimensions :

$$u_{tt} + u_{xx} = 0\,, \qquad (t,x) \in \,]0,+\infty[\, \times \mathbb{R}\,, \tag{9}$$

avec la condition de Cauchy en $t = 0$:

$$u(0,x) = 0\,, \qquad u_t(0,x) = \frac{1}{n}\sin(nx)\,, \qquad x \in \mathbb{R}\,, \tag{10}$$

où $n > 0$ est fixé. La fonction :

$$u(t,x) = \frac{1}{n^2}\sin(nx)\sinh(nt) \tag{11}$$

est solution du problème. En effet, cette fonction admet des dérivées de tout ordre, ses dérivées secondes sont données par :

$$u_{tt}(t,x) = \sin(nx)\sinh(nt)\,, \qquad u_{xx}(t,x) = -\sin(nx)\sinh(nt)\,,$$

ce qui montre que u satisfait bien l'EDP (9). On vérifie aussi directement que la fonction u vérifie les conditions initiales (10) en $t = 0$ et qu'elle est analytique. C'est donc l'unique solution analytique d'après le théorème de Cauchy-Kowalewski.
Ce problème est un exemple de problème mal posé. En effet, d'une part, lorsque $n \to +\infty$, les conditions initiales (10) convergent vers 0 dans $\mathcal{C}^0$. D'autre part, l'unique solution du problème de Cauchy avec conditions initiales nulles est la fonction identiquement nulle (on vérifie trivialement que la fonction identiquement nulle est solution et qu'il n'y a pas d'autre solution). Et pourtant, la solution du problème de Cauchy avec conditions initiales (10) ne converge pas du tout vers la fonction nulle dans $\mathcal{C}^0$ lorsque $n \to +\infty$. La solution (11) présente des oscillations dont l'amplitude explose lorsque $n \to +\infty$, car $\sinh(nt)/n^2 \to +\infty$ pour tout $t > 0$. En d'autres mots, lorsque l'on change les conditions initiales à $t = 0$ par une quantité arbitrairement petite, la solution à un temps $t > 0$ subit une modification arbitrairement grande. Le problème est donc mal posé, la solution ne dépend pas de manière continue des conditions initiales.

1.3 Équations linéaires

On a vu dans le livre [L2] que la solution générale d'une équation différentielle ordinaire linéaire homogène d'ordre n pouvait s'écrire comme une combinaison linéaire de n solutions linéairement indépendantes avec n constantes arbitraires. En d'autres mots, si $(u_1, \ldots, u_n)$ est une famille de n solutions linéairement indépendantes d'une équation différentielle ordinaire homogène d'ordre n, alors la solution générale est de la forme :

$$\sum_{k=1}^{n} \alpha_k u_k\,,$$

pour des constantes $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ arbitraires. De manière plus générale, toute combinaison linéaire de solutions est encore solution de l'équation différentielle ordinaire. Ce principe s'étend aux EDP.
Considérons une EDP linéaire du premier ordre (dans le cas où $\forall \mathbf{x} \in \Omega$, $\mathbf{a}(\mathbf{x}) = 0$) ou du second ordre (dans le cas où $\exists \mathbf{x} \in \Omega \mid \mathbf{a}(\mathbf{x}) \neq 0$) dans un domaine $\Omega \subset \mathbb{R}^n$:

$$\sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) u_{x_i x_j} + \sum_{i=1}^{n} b_i(\mathbf{x}) u_{x_i} + c(\mathbf{x}) u = 0\,, \quad \mathbf{x} \in \Omega\,.$$

On dit que cette EDP est homogène car son second membre est nul, et donc $u = 0$ est solution. Il y a en général une infinité de solutions distinctes $(u_1, \ldots, u_k, \ldots)$. Une combinaison linéaire de ces solutions est clairement solution de l'EDP. On peut en fait obtenir une très grande famille de solutions en considérant les sommes infinies :

$$u(\mathbf{x}) = \sum_{k=1}^{+\infty} \alpha_k u_k(\mathbf{x})\,,$$

où les α_k sont des constantes arbitraires. Attention, une telle somme n'a aucune raison d'être solution, car il se peut que la série ne soit pas convergente, ou bien que les dérivées et le signe somme ne puissent pas être intervertis. Cependant, ce principe de superposition est un moyen pratique de construire des solutions, comme on va le voir dans les exemples qui vont suivre.

2 Équations d'ordre 1

2.1 Équations de transport

On va commencer par regarder l'EDP la plus simple possible :

$$u_t + cu_x = 0\,, \qquad (t,x) \in \,]0,+\infty[\, \times \mathbb{R}\,, \tag{12}$$

où c est une constante réelle fixée. Une telle EDP modélise un phénomène de transport au cours du temps t le long d'une ligne décrite par l'axe des x. En général (mais pas toujours), l'état initial du système est connu à l'instant $t = 0$, c'est-à-dire que la fonction $x \mapsto u(t=0,x)$ est connue.

Dans le cas d'une telle EDP, il n'est pas difficile de trouver la forme générale de la solution. En effet, si $u \in \mathcal{C}^1(]0,+\infty[\times\mathbb{R},\mathbb{R})$ est une solution, alors la fonction :

$$v(t,x) := u(t, x+ct)$$

est de classe $\mathcal{C}^1$ sur $]0,+\infty[\times\mathbb{R}$ et vérifie pour tous $t > 0$, $x \in \mathbb{R}$:

$$\begin{aligned} v_t(t,x) &= \frac{\mathrm{d}}{\mathrm{d}t}\big[u(t,x+ct)\big] \\ &= u_t(t,x+ct) + u_x(t,x+ct)\frac{\mathrm{d}}{\mathrm{d}t}\big[x+ct\big] \\ &= u_t(t,x+ct) + cu_x(t,x+ct) = 0\,, \end{aligned}$$

ce qui montre que $v(t,x) = v(1,x)$, soit $u(t,x+ct) = f(x)$ pour tout $(t,x) \in \,]0,+\infty[\, \times \mathbb{R}$, où l'on a noté $f(x) = u(1,x+c)$. Cela s'écrit aussi :

$$\forall (t,x) \in \,]0,+\infty[\, \times \mathbb{R}\,, \qquad u(t,x) = f(x-ct)\,,$$

ce qui montre que u doit être une fonction de $x - ct$ seulement. Réciproquement, si on se donne une fonction f de classe $\mathcal{C}^1$ sur $\mathbb{R}$, alors $u(t,x) := f(x-ct)$ vérifie l'EDP (12), car :

$$u_t(t,x) = -cf'(x-ct)\,, \quad u_x(t,x) = f'(x-ct)\,.$$

On a donc montré le résultat suivant.

Proposition 2.
Les solutions de classe $\mathcal{C}^1$ de l'EDP (12) sont les fonctions de la forme :

$$u(t,x) = f(x-ct)\,,$$

où $f \in \mathcal{C}^1(\mathbb{R},\mathbb{R})$.

Comme corollaire immédiat, il n'y a qu'une seule solution de classe $\mathcal{C}^1$ sur $\mathbb{R}_+ \times \mathbb{R}$ qui vérifie de plus la condition initiale $u(t=0,x) = f(x)$, $x \in \mathbb{R}$, où f est une fonction donnée.

Cette solution est précisément $f(x-ct)$. La forme de la solution explique la terminologie « équation de transport » : la condition initiale f est transportée à la vitesse c.
Le raisonnement précédent se généralise en dimension plus grande. On va maintenant regarder l'EDP :

$$u_t + \mathbf{c}\cdot\nabla u = 0\,, \qquad (t,\mathbf{x}) \in \,]0,+\infty[\,\times\mathbb{R}^n\,, \tag{13}$$

où $\mathbf{c}\in\mathbb{R}^n$ et n est un entier strictement positif. On supposera $\mathbf{c}\neq 0$ dans la suite. Le terme $\mathbf{c}\cdot\nabla$ signifie explicitement :

$$\mathbf{c}\cdot\nabla = \sum_{j=1}^{n} c_j \frac{\partial}{\partial x_j}\cdot$$

On commence par chercher la solution générale de cette EDP. Supposons que u soit une solution de (13) de classe $\mathcal{C}^1$ sur $]0,+\infty[\times\mathbb{R}^n$. Pour tout $\mathbf{x}_0\in\mathbb{R}^n$, la fonction w à valeurs réelles définie par :

$$w(t) := u(t,\mathbf{x}_0+\mathbf{c}t)$$

est de classe $\mathcal{C}^1$ sur $]0,+\infty[$ et elle vérifie pour tout $t>0$:

$$\frac{\mathrm{d}w}{\mathrm{d}t}(t) = \mathbf{c}\cdot\nabla u(t,\mathbf{x}_0+\mathbf{c}t) + u_t(t,\mathbf{x}_0+\mathbf{c}t) = 0\,.$$

Ici, on s'est servi du fait que, pour toute fonction $\varphi\in\mathcal{C}^1(\mathbb{R}^n,\mathbb{R})$:

$$\begin{aligned}
\varphi(\mathbf{c}t) &= \varphi(c_1t,\ldots,c_nt)\,,\\
\frac{\mathrm{d}}{\mathrm{d}t}\big(\varphi(\mathbf{c}t)\big) &= \sum_{j=1}^{n}\frac{\partial\varphi}{\partial x_j}(c_1t,\ldots,c_nt)\frac{\mathrm{d}}{\mathrm{d}t}\big(c_jt\big)\\
&= \sum_{j=1}^{n}\frac{\partial\varphi}{\partial x_j}(c_1t,\ldots,c_nt)c_j = \mathbf{c}\cdot\nabla\varphi(\mathbf{c}t)\,,
\end{aligned}$$

et l'on a appliqué ce résultat avec $\varphi(\mathbf{x}) = u(t,\mathbf{x}_0+\mathbf{x})$. La fonction w est donc constante, et u vérifie $\forall t>0$:

$$u(t,\mathbf{x}_0+\mathbf{c}t) = u(1,\mathbf{x}_0+\mathbf{c})\,, \tag{14}$$

ce qui montre que u est constante sur les droites semi-infinies de l'espace $]0,+\infty[\times\mathbb{R}^n$ de direction $(1,c_1,\ldots,c_n)$. Par conséquent, si u est continue sur $\mathbb{R}_+\times\mathbb{R}^n$ et si l'on connaît la valeur de u sur l'hyperplan $\{0\}\times\mathbb{R}^n$, alors on connaît u sur tout le domaine $\mathbb{R}_+\times\mathbb{R}^n$ (il faut ici que $\mathbf{c}\neq 0$ pour que cette affirmation soit vraie).
On considère maintenant le problème avec conditions initiales :

$$u_t + \mathbf{c}\cdot\nabla u = 0\,, \qquad (t,\mathbf{x}) \in \,]0,+\infty[\,\times\mathbb{R}^n\,, \tag{15}$$

$$u(0,\mathbf{x}) = g(\mathbf{x})\,, \qquad \mathbf{x}\in\mathbb{R}^n\,, \tag{16}$$

où $g:\mathbb{R}^n\to\mathbb{R}$ est une fonction de classe $\mathcal{C}^1$.

Proposition 3.
Le problème (15-16) admet une unique solution dans $\mathcal{C}^1([0,+\infty[\times\mathbb{R}^n,\mathbb{R})$ donnée par :

$$u(t,\mathbf{x}) = g(\mathbf{x}-\mathbf{c}t)\,. \tag{17}$$

Démonstration. On vérifie facilement que (17) est solution.
En effet, $\nabla u(t,\mathbf{x}) = \nabla g(\mathbf{x}-\mathbf{c}t)$ et $u_t(t,\mathbf{x}) = -\mathbf{c}\cdot\nabla g(\mathbf{x}-\mathbf{c}t)$, donc $u_t + \mathbf{c}\cdot\nabla u = 0$.
Réciproquement, si u est solution, alors pour tout $\mathbf{x}_0 \in \mathbb{R}^n$ la fonction $t \mapsto u(t,\mathbf{x}_0+\mathbf{c}t)$ est constante sur $]0,+\infty[$ d'après (14). En utilisant la continuité de u sur $[0,+\infty[\times\mathbb{R}^n$ et la condition initiale (15) on trouve que $u(t,\mathbf{x}_0+\mathbf{c}t) = u(0,\mathbf{x}_0) = g(\mathbf{x}_0)$ pour tout $t \geqslant 0$ et $\mathbf{x}_0 \in \mathbb{R}^n$. En appliquant ce résultat avec $\mathbf{x}_0 = \mathbf{x}-\mathbf{c}t$, on trouve bien que $u(t,\mathbf{x}) = g(\mathbf{x}-\mathbf{c}t)$.

À la vue de cette preuve, on remarque que l'énoncé de la proposition reste encore vrai si l'on remplace $\mathcal{C}^1([0,+\infty[\times\mathbb{R}^n,\mathbb{R})$ par $\mathcal{C}^1(]0,+\infty[\times\mathbb{R}^n,\mathbb{R})\cap\mathcal{C}^0([0,+\infty[\times\mathbb{R}^n,\mathbb{R})$. La solution est alors nécessairement dans $\mathcal{C}^1([0,+\infty[\times\mathbb{R}^n,\mathbb{R})$. ■

On généralise le résultat précédent au problème inhomogène (i.e. avec second membre non nul) avec conditions initiales :

$$u_t + \mathbf{c}\cdot\nabla u = f\,, \quad (t,\mathbf{x}) \in]0,+\infty[\times\mathbb{R}^n\,, \tag{18}$$

$$u(0,\mathbf{x}) = g(\mathbf{x})\,, \quad \mathbf{x}\in\mathbb{R}^n\,, \tag{19}$$

où $f : \mathbb{R}_+\times\mathbb{R}^n \to \mathbb{R}$ est une fonction de classe $\mathcal{C}^1$.

Proposition 4. Le problème inhomogène (18-19) admet une unique solution dans $\mathcal{C}^1([0,+\infty[\times\mathbb{R}^n,\mathbb{R})$ donnée par :

$$u(t,\mathbf{x}) = g(\mathbf{x}-\mathbf{c}t) + \int_0^t f(s,\mathbf{x}-\mathbf{c}(t-s))\,\mathrm{d}s\,. \tag{20}$$

Démonstration. Supposons que u soit solution de classe $\mathcal{C}^1$. Fixons $(t_0,\mathbf{x}_0)\in\mathbb{R}_+\times\mathbb{R}^n$ et considérons la fonction à valeurs réelles :

$$w(t) := u(t_0+t,\mathbf{x}_0+\mathbf{c}t)\,.$$

Elle est de classe $\mathcal{C}^1$ et vérifie, pour tout $t > -t_0$:

$$\frac{\mathrm{d}w}{\mathrm{d}t}(t) = \mathbf{c}\cdot\nabla u(t_0+t,\mathbf{x}_0+\mathbf{c}t) + u_t(t_0+t,\mathbf{x}_0+\mathbf{c}t) = f(t_0+t,\mathbf{x}_0+\mathbf{c}t)\,.$$

En intégrant entre $t=-t_0$ et $t=0$, on trouve :

$$w(0) - w(-t_0) = \int_{-t_0}^0 f(t_0+t,\mathbf{x}_0+\mathbf{c}t)\,\mathrm{d}t\,.$$

On se sert d'une part de la condition initiale (19) pour obtenir que :

$$w(-t_0) = u(0,\mathbf{x}_0-\mathbf{c}t_0) = g(\mathbf{x}_0-\mathbf{c}t_0)\,.$$

D'autre part, on effectue le changement de variable $s = t + t_0$ dans l'intégrale :

$$u(t_0, \mathbf{x}_0) = g(\mathbf{x}_0 - \mathbf{c}t_0) + \int_0^{t_0} f(s, \mathbf{x}_0 - \mathbf{c}(t_0 - s))\,\mathrm{d}s\,.$$

Cela montre que, si u existe, alors elle est nécessairement de la forme (20). On a ainsi montré l'unicité de la solution.

Vérifions maintenant que u définie par (20) est effectivement solution en calculant les dérivées partielles de $u(t, \mathbf{x})$ en t et en x_j.

Considérons tout d'abord la dérivée partielle de $u(t, \mathbf{x})$ en x_j pour un certain $j \in [\![1, n]\!]$. Ici, les variables t et $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$ sont gelées et ne jouent aucun rôle. Nous devons vérifier les hypothèses du théorème de dérivation sous le signe somme (voir le théorème 124 de la page 306 de ce livre ou le corollaire 131 de la page 720 du livre [L2]).

La fonction $\widetilde{f}(x_j, s) := f(s, \mathbf{x} - \mathbf{c}(t - s))$ est continue sur $\mathbb{R} \times [0, t]$ et admet en tout point une dérivée partielle en x_j qui est elle-même continue sur $\mathbb{R} \times [0, t]$. Cela suffit pour assurer que $\widetilde{F}(x_j) := \displaystyle\int_0^t \widetilde{f}(x_j, s)\,\mathrm{d}s$ est de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et :

$$\frac{\mathrm{d}\widetilde{F}}{\mathrm{d}x_j}(x_j) = \int_0^t \frac{\partial \widetilde{f}}{\partial x_j}(x_j, s)\,\mathrm{d}s\,,$$

ce qui s'écrit aussi :

$$\frac{\partial}{\partial x_j}\left(\int_0^t f(s, \mathbf{x} - \mathbf{c}(t - s))\,\mathrm{d}s\right) = \int_0^t \frac{\partial f}{\partial x_j}(s, \mathbf{x} - \mathbf{c}(t - s))\,\mathrm{d}s\,.$$

On obtient ainsi :

$$\nabla u(t, \mathbf{x}) = \nabla g(\mathbf{x} - \mathbf{c}t) + \int_0^t \nabla f(s, \mathbf{x} - \mathbf{c}(t - s))\,\mathrm{d}s\,. \tag{21}$$

Considérons ensuite la dérivée partielle de $u(t, \mathbf{x})$ en t. Ici, les variables $(x_1, \ldots, x_n)$ sont gelées et ne jouent aucun rôle. Nous devons vérifier les hypothèses de la version du théorème de dérivation sous le signe somme dans laquelle les bornes de l'intégrale dépendent elles-mêmes du paramètre par rapport auquel on dérive (voir la proposition 132 de la page 720 du livre [L2]). Fixons $T > 0$. La fonction $\widetilde{f}(t, s) := f(s, \mathbf{x} - \mathbf{c}(t - s))$ est continue sur $[0, T] \times [0, T]$ et admet en tout point une dérivée partielle en t qui est elle-même continue sur $[0, T] \times [0, T]$. Cela suffit pour assurer que $\widetilde{F}(t) := \displaystyle\int_0^t \widetilde{f}(t, s)\,\mathrm{d}s$ est de classe $\mathcal{C}^1$ sur $[0, T]$ et :

$$\frac{\mathrm{d}\widetilde{F}}{\mathrm{d}t}(t) = \int_0^t \frac{\partial \widetilde{f}}{\partial t}(t, s)\,\mathrm{d}s + \widetilde{f}(t, t)\,,$$

ce qui s'écrit aussi :

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_0^t f(t, \mathbf{x} - \mathbf{c}(t - s))\,\mathrm{d}s\right) &= \int_0^t \frac{\partial}{\partial t}\Big(f(s, \mathbf{x} - \mathbf{c}(t - s))\Big)\,\mathrm{d}s + f(t, \mathbf{x}) \\ &= -\int_0^t \mathbf{c} \cdot \nabla f(s, \mathbf{x} - \mathbf{c}(t - s))\,\mathrm{d}s + f(t, \mathbf{x})\,.\end{aligned}$$

On obtient ainsi :

$$u_t(t,\mathbf{x}) = -\mathbf{c}\cdot\nabla g(\mathbf{x}-\mathbf{c}t) - \int_0^t \mathbf{c}\cdot\nabla f(s,\mathbf{x}-\mathbf{c}(t-s))\,\mathrm{d}s + f(t,\mathbf{x})\,. \tag{22}$$

En combinant (21) et (22) on trouve le résultat désiré :

$$u_t(t,\mathbf{x}) + \mathbf{c}\cdot\nabla u(t,\mathbf{x}) = f(t,\mathbf{x})\,.$$

■

Remarque. On a pu résoudre l'EDP (13) car on a réussi à ramener le problème à un système d'équations différentielles ordinaires. En fait, la méthode utilisée ici est un cas particulier de la méthode des caractéristiques que l'on va introduire dans la prochaine section dans le cadre général des EDP du premier ordre.

2.2 Méthode des caractéristiques

Le principe est le suivant. Considérons une EDP de la forme :

$$F(\nabla u, u, \mathbf{x}) = 0\,, \qquad \mathbf{x}\in\Omega\,, \tag{23}$$

avec des conditions aux bords $u(\mathbf{x}) = g(\mathbf{x})$, $\mathbf{x}\in\Gamma\subset\partial\Omega$. Le principe de la méthode des caractéristiques pour calculer $u(\mathbf{x})$ en un certain point $\mathbf{x}\in\Omega$ consiste à se ramener à une équation différentielle ordinaire. Pour cela, on cherche une courbe qui part d'un point $\mathbf{x}_0$ du bord Γ, qui voyage dans le domaine Ω, qui passe par $\mathbf{x}$, et sur laquelle on sait calculer u. Plus précisément, on cherche cette courbe sous la forme paramétrée $(\mathbf{X}(t))_{t\geqslant 0}$, telle que $\mathbf{X}(0) = \mathbf{x}_0\in\Gamma$, $\mathbf{X}(T) = \mathbf{x}$ (pour un certain $T>0$), et $w(t) := u(\mathbf{X}(t))$ vérifie une équation différentielle ordinaire partant de $w(0) = u(\mathbf{x}_0) = g(\mathbf{x}_0)$ connu. Si l'on sait résoudre l'équation différentielle ordinaire, alors on trouve la valeur $w(T)$ qui n'est autre que la valeur de u au point $\mathbf{x}$. Cette équation différentielle ordinaire est appelée équation caractéristique, et la courbe paramétrée $(\mathbf{X}(t))_{t\in[0,T]}$ est appelée courbe caractéristique.

Exemple. C'est exactement la stratégie que l'on a suivie avec l'équation de transport (13) :

$$\begin{aligned} &u_{\widetilde{t}} + \sum_{j=1}^n c_j u_{\widetilde{x}_j} = 0\,, \quad (\widetilde{t},\widetilde{\mathbf{x}})\in\,]0,+\infty[\,\times\mathbb{R}^n\,,\\ &u(\widetilde{t}=0,\widetilde{\mathbf{x}}) = g(\widetilde{\mathbf{x}})\,, \quad \widetilde{\mathbf{x}}\in\mathbb{R}^n\,. \end{aligned}$$

Il s'agit bien d'un cas particulier de la forme (23). Il suffit de poser :

$$\mathbf{x} = (\widetilde{t},\widetilde{\mathbf{x}}) \qquad \Omega = \,]0,+\infty[\,\times\mathbb{R}^n \qquad \Gamma = \partial\Omega = \{0\}\times\mathbb{R}^n$$

$$\nabla = \left(\frac{\partial}{\partial\widetilde{t}}, \frac{\partial}{\partial\widetilde{x}_1}, \ldots, \frac{\partial}{\partial\widetilde{x}_n}\right) \quad \text{et} \quad F(\nabla u, u, \mathbf{x}) = \begin{pmatrix}1\\ \mathbf{c}\end{pmatrix}\cdot\nabla u = u_{\widetilde{t}} + \sum_{j=1}^n c_j u_{\widetilde{x}_j}\,.$$

En notant $u\left(\begin{pmatrix}\widetilde{t}\\ \widetilde{\mathbf{x}}\end{pmatrix}\right) = u(\widetilde{t},\widetilde{\mathbf{x}})$, on a montré qu'une solution u était constante le long des droites de direction $\begin{pmatrix}1\\ \mathbf{c}\end{pmatrix}$:

$$u\left(\begin{pmatrix}\widetilde{t}\\ \widetilde{\mathbf{x}}\end{pmatrix}\right) = u\left(\begin{pmatrix}\widetilde{t}\\ \widetilde{\mathbf{x}}\end{pmatrix} + \begin{pmatrix}1\\ \mathbf{c}\end{pmatrix} t\right)\,, \quad \forall\begin{pmatrix}\widetilde{t}\\ \widetilde{\mathbf{x}}\end{pmatrix}\in\Omega,\ t\in\mathbb{R} \quad \text{tels que} \quad \begin{pmatrix}\widetilde{t}\\ \widetilde{\mathbf{x}}\end{pmatrix} + \begin{pmatrix}1\\ \mathbf{c}\end{pmatrix} t\in\Omega\,,$$

c'est-à-dire pour tout $t > -\widetilde{t}$, et que toutes ces droites venaient intersecter l'hyperplan Γ en $t = -\widetilde{t}$. C'est ainsi que l'on a pu déterminer l'expression de u :

$$u\left(\begin{pmatrix}\widetilde{t}\\ \widetilde{\mathbf{x}}\end{pmatrix}\right) = u\left(\begin{pmatrix}\widetilde{t}\\ \widetilde{\mathbf{x}}\end{pmatrix} - \begin{pmatrix}1\\ \mathbf{c}\end{pmatrix}\widetilde{t}\right) = u\left(\begin{pmatrix}0\\ \widetilde{\mathbf{x}} - \mathbf{c}\widetilde{t}\end{pmatrix}\right) = g(\widetilde{\mathbf{x}} - \mathbf{c}\widetilde{t})\,.$$

Dans cet exemple, les courbes caractéristiques sont donc des droites, les équations caractéristiques ont pour solutions des constantes, mais ce n'est bien sûr pas le cas général.

On va maintenant montrer comment on peut obtenir une équation caractéristique dans le cas général (23). On va supposer dans la suite que F, g sont suffisamment régulières, et que la solution u existe, est unique et est de classe $\mathcal{C}^2$ (on s'occupera du problème d'existence et d'unicité de la solution dans la section 2.3). On introduit alors, pour tout chemin différentiable $(\mathbf{X}(t))_{t\geqslant 0}$ dans Ω, la fonction $w(t)$ à valeurs réelles et la fonction $\mathbf{v}(t)$ à valeurs dans $\mathbb{R}^n$ données par :

$$w(t) := u(\mathbf{X}(t))\,, \qquad \mathbf{v}(t) := \nabla u(\mathbf{X}(t))\,.$$

La fonction w donne les valeurs de u le long de la courbe paramétrée, tandis que $\mathbf{v}$ donne les valeurs du gradient ∇u le long de la courbe. On va voir que l'on peut choisir le chemin $(\mathbf{X}(t))_{t\geqslant 0}$ de telle manière que l'on puisse calculer $(w(t))_{t\geqslant 0}$.
On commence par différentier les fonctions coordonnées de $\mathbf{v}(t)$:

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = \sum_{j=1}^{n} u_{x_i x_j}(\mathbf{X}(t))\frac{\mathrm{d}X_j}{\mathrm{d}t}\,, \quad i \in [\![1,n]\!]\,. \tag{24}$$

Il ne faut pas se décourager devant l'apparition de dérivées secondes de u, car on va pouvoir les éliminer. En effet, si l'on différentie (23) par rapport à x_i, alors on trouve :

$$\sum_{j=1}^{n} \frac{\partial F}{\partial v_j}(\nabla u, u, \mathbf{x})u_{x_i x_j}(\mathbf{x}) + \frac{\partial F}{\partial u}(\nabla u, u, \mathbf{x})u_{x_i}(\mathbf{x}) + \frac{\partial F}{\partial x_i}(\nabla u, u, \mathbf{x}) = 0\,. \tag{25}$$

En choisissant la courbe paramétrée $(\mathbf{X}(t))_{t\geqslant 0}$ de telle manière qu'elle satisfasse :

$$\frac{\mathrm{d}X_j}{\mathrm{d}t} = \frac{\partial F}{\partial v_j}(\mathbf{v}(t), w(t), \mathbf{X}(t))\,, \qquad j \in [\![1,n]\!]\,,$$

et en utilisant (25) au point $\mathbf{X}(t)$, l'équation (24) devient :

$$\begin{aligned}\frac{\mathrm{d}v_i}{\mathrm{d}t} &= \sum_{j=1}^{n} u_{x_i x_j}(\mathbf{X}(t))\frac{\partial F}{\partial v_j}(\mathbf{v}(t), w(t), \mathbf{X}(t))\\ &= -\frac{\partial F}{\partial u}(\mathbf{v}(t), w(t), \mathbf{X}(t))v_i(t) - \frac{\partial F}{\partial x_i}(\mathbf{v}(t), w(t), \mathbf{X}(t))\,.\end{aligned}$$

Enfin, en différentiant $w(t)$ par rapport à t :

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \sum_{j=1}^{n} \frac{\partial u}{\partial x_j}(\mathbf{X}(t))\frac{\mathrm{d}X_j}{\mathrm{d}t}(t) = \sum_{j=1}^{n} v_j(t)\frac{\partial F}{\partial v_j}(\mathbf{v}(t), w(t), \mathbf{X}(t))\,.$$

On obtient donc que le triplet $(\mathbf{v}(t), w(t), \mathbf{X}(t))_{t\geqslant 0}$ vérifie le système d'équations différentielles ordinaires suivant :

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -\nabla_{\mathbf{X}} F(\mathbf{v}(t), w(t), \mathbf{X}(t)) - \frac{\partial F}{\partial u}(\mathbf{v}(t), w(t), \mathbf{X}(t))\mathbf{v}(t)\,, \tag{26}$$

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \nabla_{\mathbf{v}} F(\mathbf{v}(t), w(t), \mathbf{X}(t)) \cdot \mathbf{v}(t)\,, \tag{27}$$

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \nabla_{\mathbf{v}} F(\mathbf{v}(t), w(t), \mathbf{X}(t))\,. \tag{28}$$

Ici, on a noté de manière évidente $\nabla_{\mathbf{x}} = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right)$ et $\nabla_{\mathbf{v}} = \left(\frac{\partial}{\partial v_1}, \ldots, \frac{\partial}{\partial v_n}\right)$.
Dans le cas où la fonction F est linéaire en ses deux premières variables :

$$F(\nabla u, u, \mathbf{x}) = \mathbf{c}(\mathbf{x}) \cdot \nabla u(\mathbf{x}) + d(\mathbf{x})u(\mathbf{x}) + e(\mathbf{x})\,, \tag{29}$$

avec $\mathbf{c} \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^n)$ et $d, e \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$, les équations caractéristiques se simplifient quelque peu, car :

$$\nabla_{\mathbf{v}} F(\mathbf{v}, u, \mathbf{x}) = \mathbf{c}(\mathbf{x})$$

est indépendant de $\mathbf{v}$ et u. L'équation (28) prend alors la forme :

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t}(t) = \mathbf{c}(\mathbf{X}(t))\,.$$

Il s'agit d'une équation différentielle ordinaire du premier ordre qui montre que la courbe caractéristique $(\mathbf{X}(t))_{t\geqslant 0}$ suit le champ de vitesses $\mathbf{c}(\mathbf{x})$. L'équation (27) se simplifie en :

$$\begin{aligned}\frac{\mathrm{d}w}{\mathrm{d}t}(t) &= \mathbf{c}(\mathbf{X}(t)) \cdot \mathbf{v}(t) = \mathbf{c}(\mathbf{X}(t)) \cdot \nabla u(\mathbf{X}(t)) \\ &= -d(\mathbf{X}(t))u(\mathbf{X}(t)) - e(\mathbf{X}(t)) \\ &= -d(\mathbf{X}(t))w(t) - e(\mathbf{X}(t))\,,\end{aligned}$$

que l'on peut intégrer explicitement dès que l'on connaît $\mathbf{X}(t)$.

Exercice 1.

Montrer que :

$$w(t) = w(0)\exp\left(-\int_0^t d(\mathbf{X}(s))\,\mathrm{d}s\right) - \int_0^t e(\mathbf{X}(s))\exp\left(-\int_s^t d(\mathbf{X}(u))\,\mathrm{d}u\right)\mathrm{d}s\,.$$

Solution. La fonction w est solution d'une équation différentielle linéaire ordinaire du premier ordre. La forme de la solution d'une telle équation résulte de la théorie sur ces équations (*cf.* le livre [L2]).

Dans le cas linéaire, on n'a pas besoin de l'équation sur $\mathbf{v}(t)$, et on obtient donc le résultat suivant.

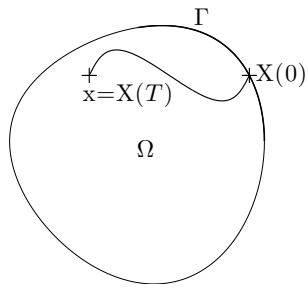


FIGURE III.2.2 – Exemple de caractéristique $(\mathbf{X}(t))_{t\in[0,T]}$ joignant un point de Γ à un point $\mathbf{x}$ à l'intérieur du domaine Ω.

Proposition 5. Soit $u \in \mathcal{C}^2(\Omega, \mathbb{R})$ solution de l'EDP :

$$\mathbf{c}(\mathbf{x}) \cdot \nabla u(\mathbf{x}) + d(\mathbf{x})u(\mathbf{x}) + e(\mathbf{x}) = 0\,, \quad \mathbf{x} \in \Omega\,, \tag{30}$$

$$u(\mathbf{x}) = g(\mathbf{x})\,, \quad \mathbf{x} \in \Gamma\,. \tag{31}$$

Les équations caractéristiques de cette EDP sont :

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t}(t) = \mathbf{c}(\mathbf{X}(t))\,, \tag{32}$$

$$\frac{\mathrm{d}w}{\mathrm{d}t}(t) = -d(\mathbf{X}(t))w(t) - e(\mathbf{X}(t))\,. \tag{33}$$

Si $\mathbf{x}_0 \in \Gamma$ et $(\mathbf{X}(t), w(t))_{t\geqslant 0}$ est solution du système (32-33) partant de $\mathbf{X}(0) = \mathbf{x}_0$ et $w(0) = g(\mathbf{x}_0)$, alors $w(t) = u(\mathbf{X}(t))$ pour tout $t \geqslant 0$, tant que $\mathbf{X}(t) \in \Omega$.

Cette proposition permet de calculer la solution de l'EDP en tous les points du domaine Ω par lesquels passe une courbe de la forme (32) partant d'un certain point du bord Γ (voir Figure III.2.2).

Il n'est pas évident que l'on puisse atteindre tous les points du domaine Ω par une solution de l'équation différentielle ordinaire (32) partant d'un point $\mathbf{x}_0 \in \Gamma$. Par exemple, si $\mathbf{c}(\mathbf{x}) = 0$ pour tout $\mathbf{x} \in \Gamma$, alors il est clair que l'on n'arrive même pas à décoller du bord Γ : $\mathbf{X}(t) = \mathbf{X}(0)\ \forall t$. En fait, le cas où $\mathbf{c}(\mathbf{x}) = 0$ pour tout $\mathbf{x} \in \Gamma$ correspond au cas où le bord Γ coïncide avec une courbe caractéristique. La prochaine section va discuter plus précisément ce problème dans le cas d'une EDP du premier ordre en dimension 2, et l'on va retrouver un problème similaire dans l'étude des équations linéaires du second ordre. Grosso modo, il n'est pas possible de résoudre le problème de Cauchy constitué d'une EDP dans un domaine Ω et d'une condition aux bords sur Γ lorsque Γ coïncide avec des caractéristiques.

Un autre cas de figure qui pourrait se présenter serait qu'un même point du domaine Ω puisse être atteint par plusieurs courbes de la forme (32), partant de différents points $\mathbf{x}_0$. Cela peut se produire pour des EDP non linéaires, mais pas pour l'EDP linéaire (30) en supposant que $\mathbf{c}$ soit lipschitzien. En effet le théorème de Cauchy-Lipschitz (*cf.* module « Équations différentielles » du livre [L2]) nous affirme que l'équation différentielle ordinaire (32)

passant par un point $\mathbf{x}$ n'a qu'une seule trajectoire solution. Les cas de figure où les courbes caractéristiques se croisent nécessitent des outils d'analyse plus élaborés que ceux dont on dispose dans le cadre de ce livre. Ils se rencontrent en fait assez couramment en physique non linéaire, et correspondent par exemple à des phénomènes de choc en mécanique des fluides.

Dans l'exercice qui suit, on applique la proposition 5 sur un exemple où tout le domaine Ω est recouvert par les courbes caractéristiques issues du bord Γ.

Exercice 2.

Résoudre par la méthode des caractéristiques l'EDP avec conditions de Dirichlet :

$$xu_y - yu_x = u\,, \quad (x,y) \in]0,+\infty[^2\,, \tag{34}$$

$$u(x=0,y) = f(y)\,, \quad y \in \mathbb{R}_+\,, \tag{35}$$

où $f \in \mathcal{C}^1(\mathbb{R}_+,\mathbb{R})$ est donnée.

Solution. Cette EDP est bien de la forme (30), avec $\Omega =]0,+\infty[^2$, $\Gamma = \{0\}\times\mathbb{R}_+ \subset \partial\Omega$, $\mathbf{c}(x,y) = \begin{pmatrix} -y \\ x \end{pmatrix}$, $d(x,y) = -1$ et $e(x,y) = 0$.

Les équations caractéristiques pour $\mathbf{X}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ et $w(t)$ sont :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -y\,, \quad \frac{\mathrm{d}y}{\mathrm{d}t} = x\,, \quad \frac{\mathrm{d}w}{\mathrm{d}t} = w\,,$$

partant d'un point de Γ : $x(0) = 0$, $y(0) = y_0$, $w(0) = f(y_0)$, $y_0 \in \mathbb{R}_+$. L'intégration de ces équations différentielles ordinaires donne :

$$x(t) = y_0 \sin t\,, \quad y(t) = y_0 \cos t\,, \quad w(t) = f(y_0)\exp t\,.$$

Soit $(x,y) \in \Omega$.

Il faut trouver $(y_0,t_0) \in \mathbb{R}_+ \times \mathbb{R}_+$ tel que $(x(t_0),y(t_0)) = (x,y)$. On vérifie que $y_0 = \sqrt{x^2+y^2}$ et $t_0 = \arctan(y/x)$ conviennent. Donc, si u est solution, on a forcément :

$$u(x,y) = w(t_0) = f(y_0)\exp t_0 = f(\sqrt{x^2+y^2})\exp\left[\arctan\frac{y}{x}\right]\,.$$

On peut alors vérifier directement que cette fonction est effectivement solution du problème de Dirichlet (34-35).

2.3 Le problème de Cauchy en dimension 2

La méthode des caractéristiques permet d'étudier l'existence et l'unicité de la solution d'une EDP lorsque l'on prescrit des conditions sur un bord, une courbe, une surface qui n'est pas caractéristique. On donne une formulation précise de ce résultat dans cette section dans le cas d'une EDP linéaire d'ordre 1 en dimension 2. On considère un domaine $\Omega \subset \mathbb{R}^2$ et une courbe $\gamma \subset \Omega$ avec sa représentation paramétrique $(x(t),y(t))_{t\in]0,1[}$:

$$\gamma := \left\{(x,y) \in \mathbb{R}^2 \mid x = x(t), y = y(t) \text{ pour } t \in]0,1[\right\}\,. \tag{36}$$

On cherche à résoudre l'EDP :

$$a(x,y)u_x + b(x,y)u_y + d(x,y)u + e(x,y) = 0\,, \quad (x,y) \in \Omega\,, \tag{37}$$

où a, b, d, e sont des fonctions données de Ω dans $\mathbb{R}$, avec des conditions prescrites sur la courbe γ :

$$\text{pour tout } t \in]0,1[\,, \quad u(x(t),y(t)) = f(t)\,, \tag{38}$$

où f est une fonction donnée de $]0,1[$ dans $\mathbb{R}$. La recherche d'une telle solution s'appelle le problème de Cauchy pour l'EDP (37) relatif à γ.

Proposition 6. Soient $\Omega \subset \mathbb{R}^2$ et $a, b, d, e \in \mathcal{C}^2(\Omega, \mathbb{R})$ définissant l'EDP (37). Soient $x, y, f \in \mathcal{C}^2(]0,1[, \mathbb{R})$ vérifiant :

$$\text{pour tout } t \in]0,1[\,, \quad y'(t)a(x(t),y(t)) - x'(t)b(x(t),y(t)) \neq 0\,. \tag{39}$$

Alors il existe une unique fonction $u \in \mathcal{C}^2$ vérifiant (37) dans un voisinage de γ et la condition (38) sur γ.

On donne ci-dessous la preuve complète de ce résultat. Elle est assez longue, mais elle peut être décomposée en des étapes élémentaires qui sont toutes simples car elles ne font intervenir que des calculs sur des équations différentielles ordinaires. Avant de se lancer dans la preuve, on peut en donner rapidement l'idée. On commence par noter que la condition (39) signifie que γ n'est pas tangente à une courbe caractéristique. On peut alors, de chaque point de γ, faire partir une courbe caractéristique qui n'est pas tangente à γ et sur laquelle on sait calculer u par la méthode des caractéristiques. Ces courbes caractéristiques permettent de recouvrir un voisinage de γ. C'est ce qui permet de prouver la proposition 6 qui est un résultat d'existence et d'unicité locale de la solution. Pour avoir l'existence et l'unicité globale de la solution, il faudrait par exemple arriver à montrer que toutes ces courbes caractéristiques recouvrent Ω.

Démonstration de la proposition 6. Pour tout $t \in]0,1[$ fixé, on note $X(s,t)$, $Y(s,t)$, et $w(s,t)$ les solutions du système d'équations différentielles ordinaires :

$$\frac{\partial X}{\partial s} = a(X,Y)\,, \quad X(s=0,t) = x(t)\,, \tag{40}$$

$$\frac{\partial Y}{\partial s} = b(X,Y)\,, \quad X(s=0,t) = x(t)\,, \tag{41}$$

$$\frac{\partial w}{\partial s} = -d(X,Y)w - e(X,Y)\,, \quad w(s=0,t) = f(t)\,. \tag{42}$$

Ce sont les solutions des équations caractéristiques (32-33) partant du point $(x(t),y(t))$ de γ. Comme a, b, d, e appartiennent à $\mathcal{C}^1(\Omega, \mathbb{R})$, ces solutions sont bien définies tant que $(X(s,t), Y(s,t))$ reste dans Ω. De plus ces solutions sont de classe $\mathcal{C}^2$.

Étape 1. Il existe un ouvert U de $\mathbb{R}^2$ contenant $\{0\}\times]0,1[$ et un ouvert $W \subset \Omega$ de $\mathbb{R}^2$ contenant γ tels que, pour tout $(x,y) \in W$, il existe un unique $(s,t) \in U$ vérifiant :

$$(x,y) = (X(s,t), Y(s,t))\,. \tag{43}$$

La fonction $(s,t) \in U \mapsto (x,y) \in W$ est de classe $\mathcal{C}^2$.

Cette étape montre que l'on peut trouver un voisinage de γ dans lequel chaque point est lié à un unique point de γ par une unique courbe caractéristique.

Démonstration. On fixe un $t_0 \in]0,1[$ et l'on note $\mathbf{x}_0 = (x(t_0), y(t_0)) \in \gamma$.
On considère $\mathbf{X} : (s,t) \to (X(s,t), Y(s,t))$ l'application qui donne la solution des équations différentielles ordinaires (40-41).
On a en particulier $\mathbf{X}(0,t_0) = (X(0,t_0), Y(0,t_0)) = (x(t_0), y(t_0)) = \mathbf{x}_0$. La matrice jacobienne de cette fonction en $(0,t_0)$ est :

$$J_{(0,t_0)}\mathbf{X} = \begin{pmatrix} \dfrac{\partial X}{\partial s}(0,t_0) & \dfrac{\partial X}{\partial t}(0,t_0) \\ \dfrac{\partial Y}{\partial s}(0,t_0) & \dfrac{\partial Y}{\partial t}(0,t_0) \end{pmatrix} = \begin{pmatrix} a(x(t_0), y(t_0)) & x'(t_0) \\ b(x(t_0), y(t_0)) & y'(t_0) \end{pmatrix},$$

dont le déterminant vaut :

$$\mathrm{Det} J_{(0,t_0)}\mathbf{X} = a(x(t_0), y(t_0))y'(t_0) - b(x(t_0), y(t_0))x'(t_0)\,.$$

Par l'hypothèse (39) ce déterminant est non nul, donc la matrice jacobienne est inversible. Le théorème d'inversion locale (*cf.* le livre [L2]) donne l'existence d'un ouvert U_{t_0} de $\mathbb{R}^2$ contenant $(0,t_0)$ et d'un ouvert W_{t_0} de $\mathbb{R}^2$ contenant $\mathbf{x}_0$ tels que pour tout $(x,y) \in W_{t_0}$, il existe un unique $(s,t) \in U_{t_0}$ vérifiant (43).
Il affirme aussi que la fonction $(s,t) \in U_{t_0} \mapsto (x,y) \in W_{t_0}$ est de classe $\mathcal{C}^1$.
Une extension du théorème précise que si la fonction $\mathbf{X}$ est de classe $\mathcal{C}^2$, alors la fonction $(s,t) \in U_{t_0} \to (x,y) \in W_{t_0}$ est aussi de classe $\mathcal{C}^2$. De plus, $U := \bigcup_{t_0 \in]0,1[} U_{t_0}$ est un ouvert qui contient $\{0\} \times]0,1[$ (on rappelle que toute union d'ouverts de $\mathbb{R}^2$ est ouvert), de même $W := \bigcup_{t_0 \in]0,1[} W_{t_0}$ est un ouvert qui contient γ. Cela achève la démonstration. ■

L'étape 2 énonce le résultat d'unicité locale.

Étape 2 : Il existe au plus une solution à l'EDP (37) dans $\mathcal{C}^2(W,\mathbb{R})$ vérifiant la condition (38).

Démonstration. Soit u une solution dans $\mathcal{C}^2(W,\mathbb{R})$. Alors, pour tout $\mathbf{x} \in W$, il existe un unique $(s,t) \in U$ tel que $\mathbf{x} = (X(s,t), Y(s,t))$. D'après la proposition 5, on a nécessairement $u(\mathbf{x}) = w(s,t)$ où w est la solution de (42). Cela montre le résultat cherché. ■

Il reste à montrer l'existence locale d'une solution. Pour tout $\mathbf{x} \in W$, on définit :

$$u(\mathbf{x}) := w(s,t)\,, \tag{44}$$

où w est la solution de (42) et $(s,t) \in U$ est l'unique point vérifiant $\mathbf{x} = (X(s,t), Y(s,t))$. Il est clair que u vérifie la condition (38). On va montrer que u vérifie l'EDP (37). Cela va se faire en cinq étapes (étapes 3 à 7 ci-dessous).

Étape 3 : Soit u définie par (44). Alors, pour tout $t \in]0,1[$, $u_x(x(t), y(t)) = \alpha(t)$ et $u_y(x(t), y(t)) = \beta(t)$ où $(\alpha(t), \beta(t))$ est donné par :

$$\begin{pmatrix} \alpha(t) \\ \beta(t) \end{pmatrix} = \begin{pmatrix} x'(t) & y'(t) \\ a(x(t), y(t)) & b(x(t), y(t)) \end{pmatrix}^{-1} \begin{pmatrix} f'(t) \\ -d(x(t), y(t))f(t) - e(x(t), y(t)) \end{pmatrix}.$$

On peut noter que, par l'hypothèse (39), la matrice qui intervient dans la définition de $(\alpha(t), \beta(t))$ est bien inversible.

Démonstration. Comme u est définie par (44), cette fonction vérifie pour tout $t \in]0,1[$: $u(x(t), y(t)) = f(t)$. Comme toutes les fonctions de cette équation sont de classe $\mathcal{C}^1$, on peut dériver en t et l'on obtient :

$$u_x(x(t), y(t))x'(t) + u_y(x(t), y(t))y'(t) = f'(t)\,.$$

La fonction u vérifie aussi $u(X(s,t), Y(s,t)) = w(s,t)$ pour s proche de 0, donc en dérivant par rapport à s en $s = 0$:

$$u_x(x(t), y(t))\frac{\partial X}{\partial s}(s=0,t) + u_y(x(t), y(t))\frac{\partial Y}{\partial s}(s=0,t) = \frac{\partial w}{\partial s}(s=0,t)\,.$$

En se servant de (40-42) on trouve :

$$u_x(x(t), y(t))a(x(t), y(t)) + u_y(x(t), y(t))b(x(t), y(t)) = -d(x(t), y(t))f(t) - e(x(t), y(t))\,.$$

Cela montre que $(\alpha(t), \beta(t)) := (u_x(x(t), y(t)), u_y(x(t), y(t)))$ vérifie le système linéaire :

$$\begin{pmatrix} x'(t) & y'(t) \\ a(x(t), y(t)) & b(x(t), y(t)) \end{pmatrix} \begin{pmatrix} \alpha(t) \\ \beta(t) \end{pmatrix} = \begin{pmatrix} f'(t) \\ -d(x(t), y(t))f(t) - e(x(t), y(t)) \end{pmatrix},$$

ce qui donne le résultat annoncé. ■

Les fonctions v_1 et v_2 introduites ci-dessous sont celles qui interviennent dans les équations caractéristiques (26). Comme le montre l'étape 4, elles sont en fait fortement liées à w.

Étape 4 : Soit $t \in]0,1[$ fixé. On définit $(v_1(s,t), v_2(s,t))$ les solutions des équations différentielles ordinaires suivantes :

$$\frac{\partial v_1}{\partial s} = -a_x(X,Y)v_1 - b_x(X,Y)v_2 - d_x(X,Y)w - e_x(X,Y) - d(X,Y)v_1\,, \tag{45}$$

$$\frac{\partial v_2}{\partial s} = -a_y(X,Y)v_1 - b_y(X,Y)v_2 - d_y(X,Y)w - e_y(X,Y) - d(X,Y)v_2\,, \tag{46}$$

partant de :

$$v_1(s=0,t) = \alpha(t)\,, \tag{47}$$

$$v_2(s=0,t) = \beta(t)\,, \tag{48}$$

où (X,Y) est la solution de (40-41). Alors, pour tout $(s,t) \in U$, on a :

$$\begin{aligned} a(X(s,t),Y(s,t))v_1(s,t) &+ b(X(s,t),Y(s,t))v_2(s,t) \\ &= -d(X(s,t),Y(s,t))w(s,t) - e(X(s,t),Y(s,t))\,. \end{aligned} \tag{49}$$

Démonstration. On note :

$$\begin{aligned} \varphi(s,t) := a(X(s,t),Y(s,t))v_1(s,t) &+ b(X(s,t),Y(s,t))v_2(s,t) \\ &+ d(X(s,t),Y(s,t))w(s,t) + e(X(s,t),Y(s,t))\,. \end{aligned}$$

Notre objectif est de montrer que $\varphi \equiv 0$. On remarque tout d'abord que :
$\varphi(s=0,t) = a(x(t),y(t))\alpha(t)+b(x(t),y(t))\beta(t)+d(x(t),y(t))f(t)+e(x(t),y(t)) = 0$,
d'après l'étape 3.
On dérive ensuite en s la fonction φ (en se servant de (40-42) et de (45-46)) :

$$\begin{aligned}
\frac{\partial\varphi}{\partial s} &= a_x\frac{\partial X}{\partial s}v_1 + a\frac{\partial v_1}{\partial s} + a_y\frac{\partial Y}{\partial s}v_1 + b_x\frac{\partial X}{\partial s}v_2 + b\frac{\partial v_2}{\partial s} + b_y\frac{\partial Y}{\partial s}v_2 \\
&\quad + d_x\frac{\partial X}{\partial s}w + d\frac{\partial w}{\partial s} + d_y\frac{\partial Y}{\partial s}w + e_x\frac{\partial X}{\partial s} + e_y\frac{\partial Y}{\partial s} \\
&= a_x a v_1 + a\big(-a_x v_1 - b_x v_2 - d_x w - e_x - d v_1\big) + a_y b v_1 \\
&\quad + b_x a v_2 + b\big(-a_y v_1 - b_y v_2 - d_y w - e_y - d v_2\big) + b_y b v_2 \\
&\quad + d_x a w + d\big(-dw - e\big) + d_y b w + e_x a + e_y b \\
&= -d\big(a v_1 + b v_2 + d w + e\big)\,,
\end{aligned}$$

et l'on trouve que φ vérifie l'équation différentielle ordinaire (en s) :

$$\frac{\partial\varphi}{\partial s}(s,t) = -d(X(s,t),Y(s,t))\varphi(s,t)\,,$$

avec la condition initiale $\varphi(s=0,t)=0$. L'unique solution de cette équation différentielle ordinaire est la fonction nulle. On a donc montré le résultat désiré. ∎

Étape 5 : On a pour tout $(s,t)\in U$ *:*

$$\frac{\partial w}{\partial s}(s,t) = v_1(s,t)\frac{\partial X}{\partial s}(s,t) + v_2(s,t)\frac{\partial Y}{\partial s}(s,t)\,, \tag{50}$$

$$\frac{\partial w}{\partial t}(s,t) = v_1(s,t)\frac{\partial X}{\partial t}(s,t) + v_2(s,t)\frac{\partial Y}{\partial t}(s,t)\,. \tag{51}$$

Démonstration. En substituant l'équation (42) dans (49), on trouve :

$$v_1(s,t)a(X(s,t),Y(s,t)) + v_2(s,t)b(X(s,t),Y(s,t)) = \frac{\partial w}{\partial s}(s,t)\,,$$

ce qui donne (50) avec (40-41). Pour montrer (51), on introduit :

$$\psi(s,t) := \frac{\partial w}{\partial t}(s,t) - v_1(s,t)\frac{\partial X}{\partial t}(s,t) - v_2(s,t)\frac{\partial Y}{\partial t}(s,t)\,,$$

et l'on va montrer que $\psi\equiv 0$. On remarque tout d'abord que

$$\psi(s=0,t) = f'(t) - \alpha(t)x'(t) - \beta(t)y'(t) = 0\,,$$

d'après l'étape 3. On dérive ensuite en s la fonction ψ (en se servant de (40-41) et de (51)) :

$$\begin{aligned}
\frac{\partial\psi}{\partial s} &= \frac{\partial^2 w}{\partial s\partial t} - \frac{\partial v_1}{\partial s}\frac{\partial X}{\partial t} - v_1\frac{\partial^2 X}{\partial s\partial t} - \frac{\partial v_2}{\partial s}\frac{\partial Y}{\partial t} - v_2\frac{\partial^2 Y}{\partial s\partial t} \\
&= \frac{\partial}{\partial t}\big(v_1 a + v_2 b\big) - \frac{\partial v_1}{\partial s}\frac{\partial X}{\partial t} - v_1\frac{\partial a}{\partial t} - \frac{\partial v_2}{\partial s}\frac{\partial Y}{\partial t} - v_2\frac{\partial b}{\partial t} \\
&= \frac{\partial v_1}{\partial t}a + \frac{\partial v_2}{\partial t}b - \frac{\partial v_1}{\partial s}\frac{\partial X}{\partial t} - \frac{\partial v_2}{\partial s}\frac{\partial Y}{\partial t}\,.
\end{aligned}$$

Puis en se servant de (45-46) et de (49) :

$$\begin{aligned}
\frac{\partial \psi}{\partial s} &= \frac{\partial v_1}{\partial t} a + \frac{\partial v_2}{\partial t} b - \big(- a_x v_1 - b_x v_2 - d_x w - e_x - d v_1 \big) \frac{\partial X}{\partial t} \\
&\qquad - \big(- a_y v_1 - b_y v_2 - d_y w - e_y - d v_2 \big) \frac{\partial Y}{\partial t} \\
&= \frac{\partial v_1}{\partial t} a + \frac{\partial v_2}{\partial t} b - \big(- \frac{\partial a}{\partial t} v_1 - \frac{\partial b}{\partial t} v_2 - \frac{\partial d}{\partial t} w - \frac{\partial e}{\partial t} \big) + d v_1 \frac{\partial X}{\partial t} + d v_2 \frac{\partial Y}{\partial t} \\
&= \frac{\partial}{\partial t} (v_1 a + v_2 b + d w + e) - d \frac{\partial w}{\partial t} + d v_1 \frac{\partial X}{\partial t} + d v_2 \frac{\partial Y}{\partial t} \\
&= -d \big(\frac{\partial w}{\partial t} - v_1 \frac{\partial X}{\partial t} - v_2 \frac{\partial Y}{\partial t} \big) ,
\end{aligned}$$

on trouve que ψ vérifie l'équation différentielle ordinaire (en s) :

$$\frac{\partial \psi}{\partial s}(s,t) = -d(X(s,t), Y(s,t)) \psi(s,t) ,$$

avec la condition initiale $\psi(s=0,t) = 0$. L'unique solution de cette équation différentielle ordinaire est la fonction nulle. On a donc montré le résultat désiré. ∎

Étape 6 : Soit u définie par (44). On a pour tout $(s,t) \in U$:

$$u_x(X(s,t), Y(s,t)) = v_1(s,t) , \tag{52}$$

$$u_y(X(s,t), Y(s,t)) = v_2(s,t) . \tag{53}$$

Démonstration. La définition (44) de u nous dit que $u(x,y) = w(S(x,y), T(x,y))$ où $(S(x,y), T(x,y))$ est la fonction réciproque de $(X(s,t), Y(s,t))$. Cela nous permet d'écrire, en se servant de (50-51), que :

$$\begin{aligned}
u_x(x,y) &= \frac{\partial w}{\partial s}_{|s=S(x,y), t=T(x,y)} S_x(x,y) + \frac{\partial w}{\partial t}_{|s=S(x,y), t=T(x,y)} T_x(x,y) \\
&= \Big(v_1 \frac{\partial X}{\partial s} + v_2 \frac{\partial Y}{\partial s} \Big)_{|s=S(x,y), t=T(x,y)} S_x(x,y) \\
&\qquad + \Big(v_1 \frac{\partial X}{\partial t} + v_2 \frac{\partial Y}{\partial t} \Big)_{|s=S(x,y), t=T(x,y)} T_x(x,y) .
\end{aligned}$$

Par différentiation d'une fonction réciproque, on a :

$$\begin{aligned}
\begin{pmatrix} S_x(x,y) & T_x(x,y) \\ S_y(x,y) & T_y(x,y) \end{pmatrix} &= \begin{pmatrix} \frac{\partial X}{\partial s} & \frac{\partial Y}{\partial s} \\ \frac{\partial X}{\partial t} & \frac{\partial Y}{\partial t} \end{pmatrix}^{-1}_{|s=S(x,y), t=T(x,y)} \\
&= \frac{1}{\frac{\partial X}{\partial s}\frac{\partial Y}{\partial t} - \frac{\partial Y}{\partial s}\frac{\partial X}{\partial t}} \begin{pmatrix} \frac{\partial Y}{\partial t} & -\frac{\partial Y}{\partial s} \\ -\frac{\partial X}{\partial t} & \frac{\partial X}{\partial s} \end{pmatrix}_{|s=S(x,y), t=T(x,y)} .
\end{aligned}$$

En substituant dans l'expression de u_x trouvée ci-dessus, on obtient :

$$\begin{aligned} u_x(x,y) &= \frac{(v_1\frac{\partial X}{\partial s}+v_2\frac{\partial Y}{\partial s})\frac{\partial Y}{\partial t}-(v_1\frac{\partial X}{\partial t}+v_2\frac{\partial Y}{\partial t})\frac{\partial Y}{\partial s}}{\frac{\partial X}{\partial s}\frac{\partial Y}{\partial t}-\frac{\partial Y}{\partial s}\frac{\partial X}{\partial t}}\Big|_{s=S(x,y),t=T(x,y)} \\ &= v_1(S(x,y),T(x,y))\,. \end{aligned}$$

Le calcul de u_y s'effectue de la même façon, ce qui donne le résultat annoncé. ■

L'étape suivante établit l'existence locale de la solution de l'EDP (37).

Étape 7 : Soit u définie par (44). On a pour tout $\mathbf{x} \in W$:

$$u(\mathbf{x}) = a(\mathbf{x})u_x(\mathbf{x}) + b(\mathbf{x})u_y(\mathbf{x}) + d(\mathbf{x})u(\mathbf{x}) + e(\mathbf{x})\,. \tag{54}$$

Démonstration. Pour tout $(s,t) \in U$, on a d'après (49), (52-53), et la définition (44) de u :

$$\begin{aligned} a(X(s,t),Y(s,t))\frac{\partial u}{\partial x}(X(s,t),Y(s,t)) + b(X(s,t),Y(s,t))\frac{\partial u}{\partial y}(X(s,t),Y(s,t)) \\ = -d(X(s,t),Y(s,t))u(X(s,t),Y(s,t)) - e(X(s,t),Y(s,t))\,. \end{aligned}$$

Donc, pour tout $(x,y) \in W$, u vérifie (54). ■

Cela achève la démonstration de la proposition 6.

3 Équations linéaires d'ordre 2

On considère une équation linéaire d'ordre 2 en dimension 2, c'est-à-dire une EDP de la forme :

$$a(x,y)u_{xx}+2b(x,y)u_{xy}+c(x,y)u_{yy} = d(x,y)u_x+e(x,y)u_y+r(x,y)u+s(x,y)\,, \tag{55}$$

pour tout $(x,y) \in \Omega$, où Ω est un ouvert de $\mathbb{R}^2$ et les fonctions $a,b,c,d,e,r,s : \Omega \to \mathbb{R}$ sont suffisamment régulières. On supposera aussi que les trois fonctions a,b,c ne s'annulent pas simultanénement en un point quelconque de Ω.

Le premier objectif de cette section est de montrer qu'il n'y a en fait que trois grands types d'équations du type (55). Puis on analysera en détail quelques EDP du type (55) qui interviennent en physique.

3.1 Caractéristiques et classification des équations

On considère une courbe $\gamma \subset \Omega$ et sa représentation paramétrique $(x(t),y(t))_{t\in]0,1[}$: $\gamma = \left\{(x,y) \in \mathbb{R}^2 \mid x = x(t), y = y(t) \text{ pour } t \in]0,1[\right\}$. On suppose que cette courbe est régulière, plus exactement, on suppose que les fonctions x et y sont de classe $\mathcal{C}^1$ et vérifient $\forall t \in]0,1[,\ x'(t)^2 + y'(t)^2 \neq 0$. La question abordée dans cette section est de savoir s'il existe une unique solution de l'EDP (55) avec des conditions spécifiées sur la courbe γ. On va identifier le type de conditions à spécifier et le type de courbes pour lequel un tel résultat existe. Avant cela, on commence par rappeler quelques propriétés élémentaires des courbes régulières.

- L'abscisse curviligne σ est la fonction de $]0,1[$ dans $\mathbb{R}$ définie par

$$\sigma(t) = \int_0^t [x'(r)^2 + y'(r)^2]^{1/2}\,\mathrm{d}r\,.$$

Cette fonction est un $\mathcal{C}^1$-difféomorphisme strictement croissant. On peut donc définir la fonction réciproque qui à s associe $\tau(s)$: $\sigma(t) = s$ si, et seulement si, $\tau(s) = t$. Enfin on peut définir les fonctions $\widetilde{x}$ et $\widetilde{y}$ qui à s associent $\widetilde{x}(s) = x(\tau(s))$ et $\widetilde{y}(s) = x(\tau(s))$.

- Soit $M \in \gamma$. Il existe un unique s tel que $M = (\widetilde{x}(s), \widetilde{y}(s))$. Le vecteur tangent à γ au point M est défini comme le vecteur :

$$\mathbf{T} := \frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}\mathbf{e}_x + \frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}\mathbf{e}_y\,. \tag{56}$$

Ici les vecteurs $\mathbf{e}_x$ et $\mathbf{e}_y$ sont les vecteurs unitaires pointant dans la direction de l'axe des x et y, respectivement. Le vecteur $\mathbf{T}$ est de norme 1 car :

$$\|\mathbf{T}\|^2 = \left(\frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}\right)^2 + \left(\frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}\right)^2 = \left[\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2\right]_{|t=\tau(s)} \left(\frac{\mathrm{d}\tau}{\mathrm{d}s}\right)^2 = 1\,.$$

- Soit $M \in \gamma$. Il existe un unique s tel que $M = (\widetilde{x}(s), \widetilde{y}(s))$. Le vecteur normal à γ au point M est le vecteur $\mathbf{N}$ obtenu par rotation de $\pi/2$ de $\mathbf{T}$. Il est donné par :

$$\mathbf{N} := -\frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}\mathbf{e}_x + \frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}\mathbf{e}_y\,. \tag{57}$$

La dérivée normale d'un champ $u \in \mathcal{C}^1(\Omega, \mathbb{R})$ au point $M \in \gamma$ est définie par $u_n := \nabla u(M) \cdot \mathbf{N}$.

Lemme 7. Soient $\Omega \subset \mathbb{R}^2$ un voisinage ouvert de γ et $u \in \mathcal{C}^1(\Omega, \mathbb{R})$. Il est équivalent de connaître u et u_n sur γ ou de connaître u, u_x et u_y sur γ.

Il est clair que, si l'on connaît u, u_x et u_y, alors on connaît u et u_n car u_n s'exprime en termes de u_x et u_y par définition. La réciproque est moins claire, mais juste cependant, l'idée sous-jacente est que la dérivée tangentielle $\nabla u \cdot \mathbf{T}$ peut se retrouver à partir des valeurs de u le long de la courbe, et que la dérivée normale et la dérivée tangentielle de u donnent u_x et u_y car la matrice qui apparaît dans la relation :

$$\begin{pmatrix} \mathbf{N} \cdot \nabla u \\ \mathbf{T} \cdot \nabla u \end{pmatrix} = \begin{pmatrix} -\frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s} & \frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s} \\ \frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s} & \frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s} \end{pmatrix} \begin{pmatrix} u_x \\ u_y \end{pmatrix}$$

est inversible (le déterminant de la matrice vaut -1).

Démonstration. On note $\forall t \in \,]0,1[$:

$$w(t) := u(x(t), y(t))\,, \qquad p(t) := u_x(x(t), y(t))\,, \tag{58}$$

$$q(t) := u_y(x(t), y(t))\,, \qquad z(t) := u_n(x(t), y(t))\,. \tag{59}$$

L'énoncé du lemme signifie qu'il est équivalent de connaître les fonctions $(w(t), z(t))_{t\in]0,1[}$ ou les fonctions $(w(t), p(t), q(t))_{t\in]0,1[}$. Or (57) montre que :

$$\begin{aligned} z(t) &= u_x(x(t), y(t)) \left(-\frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}(\sigma(t))\right) + u_y(x(t), y(t)) \left(\frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}(\sigma(t))\right) \\ &= -p(t)\frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}(\sigma(t)) + q(t)\frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}(\sigma(t))\,. \end{aligned} \tag{60}$$

Par conséquent, si l'on connaît $(w(t), p(t), q(t))_{t\in]0,1[}$, alors on connaît $(w(t), z(t))_{t\in]0,1[}$. Mais on a aussi :

$$\begin{aligned}\frac{\mathrm{d}w}{\mathrm{d}t}(t) &= u_x(x(t), y(t))\frac{\mathrm{d}x}{\mathrm{d}t}(t) + u_y(x(t), y(t))\frac{\mathrm{d}y}{\mathrm{d}t}(t)\\ &= p(t)\frac{\mathrm{d}x}{\mathrm{d}t}(t) + q(t)\frac{\mathrm{d}y}{\mathrm{d}t}(t)\\ &= p(t)\frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}(\sigma(t))\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t) + q(t)\frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}(\sigma(t))\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t)\,, \qquad (61)\end{aligned}$$

car $x(t) = \widetilde{x}(\sigma(t))$ et $y(t) = \widetilde{y}(\sigma(t))$. D'autre part, si l'on connaît $(w(t), z(t))_{t\in]0,1[}$, alors on connaît $(w(t), w'(t), z(t))_{t\in]0,1[}$. À t fixé, les équations (60-61) forment alors un système linéaire 2×2 d'inconnues $(p(t), q(t))$:

$$\begin{pmatrix} -\frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}(\sigma(t)) & \frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}(\sigma(t)) \\ \frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}(\sigma(t))\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t) & \frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}(\sigma(t))\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t)\end{pmatrix}\begin{pmatrix} p(t)\\ q(t)\end{pmatrix} = \begin{pmatrix} z(t)\\ \frac{\mathrm{d}w}{\mathrm{d}t}(t)\end{pmatrix}.$$

Le déterminant de ce système est

$$-\left(\frac{\mathrm{d}\widetilde{x}}{\mathrm{d}s}(\sigma(t))\right)^2\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t) - \left(\frac{\mathrm{d}\widetilde{y}}{\mathrm{d}s}(\sigma(t))\right)^2\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t) = -\frac{\mathrm{d}\sigma}{\mathrm{d}t}(t) \neq 0\,.$$

Donc on peut déterminer $(w(t), p(t), q(t))_{t\in]0,1[}$ à partir de $(w(t), z(t))_{t\in]0,1[}$. ■

Définition 5. La recherche d'une solution de l'EDP (55), connaissant u et u_n sur γ, s'appelle le *problème de Cauchy* relatif à γ.

Supposons qu'il existe une unique solution sur un voisinage ouvert de γ vérifiant le problème de Cauchy. Alors, nécessairement, les dérivées secondes de u doivent être complètement déterminées sur γ. Pour les trouver, on dérive $p(t) = u_x(x(t), y(t))$ et $q(t) = u_y(x(t), y(t))$ en t :

$$\begin{aligned}p'(t) &= u_{xx}(x(t), y(t))x'(t) + u_{xy}(x(t), y(t))y'(t)\,,\\ q'(t) &= u_{xy}(x(t), y(t))x'(t) + u_{yy}(x(t), y(t))y'(t)\,,\end{aligned}$$

et l'on écrit l'EDP que doit vérifier u au point $(x(t), y(t))$:

$$u_{xx}(x(t), y(t))a(t) + 2u_{xy}(x(t), y(t))b(t) + u_{yy}(x(t), y(t))c(t) = h(t)\,,$$

avec $a(t) = a(x(t), y(t))$, $b(t) = b(x(t), y(t))$, $c(t) = c(x(t), y(t))$, et :

$$h(t) = d(x(t), y(t))p(t) + e(x(t), y(t))q(t) + r(x(t), y(t))w(t) + s(x(t), y(t))\,,$$

en utilisant les notations (58-59). D'après le lemme précédent, comme on connaît u et u_n sur γ, on connaît la fonction $h(t)$. À t fixé, le vecteur :

$$\mathbf{Y}(t) := \begin{pmatrix} u_{xx}(x(t), y(t))\\ u_{xy}(x(t), y(t))\\ u_{yy}(x(t), y(t))\end{pmatrix}$$

est donc solution d'un système linéaire 3×3 :

$$\mathbf{A}(t)\mathbf{Y}(t) = \begin{pmatrix} p'(t)\\ q'(t)\\ h(t)\end{pmatrix}, \quad \mathbf{A}(t) := \begin{pmatrix} x'(t) & y'(t) & 0\\ 0 & x'(t) & y'(t)\\ a(t) & 2b(t) & c(t)\end{pmatrix}.$$

Le déterminant de ce système est :

$$\Delta(t) = c(x(t), y(t)))x'(t)^2 - 2b(x(t), y(t))x'(t)y'(t) + a(x(t), y(t))y'(t)^2 . \qquad (62)$$

Si $\Delta(t) \neq 0$, les dérivées secondes de u sont uniquement déterminées sur γ.
Si $\Delta(t) = 0$, alors le système précédent a ou bien zéro solution, ou bien une infinité.

Définition 6. Les *courbes caractéristiques* sont les courbes qui annulent $\Delta(t)$ $\forall t$.
On dit que γ n'est *caractéristique en aucun point* si $\Delta(t) \neq 0$ $\forall t$.

Le résultat suivant caractérise les courbes pour lesquelles le problème de Cauchy est bien posé.

Proposition 8.
Si γ n'est caractéristique en aucun point, si les fonctions a, b, c, d, e, r, s sont de classe $\mathcal{C}^2$, alors il existe une unique fonction u de classe $\mathcal{C}^2$ définie dans un voisinage de γ solution du problème de Cauchy relatif à γ.

Nous ne donnerons pas la preuve complète de ce résultat dans ce cours. On remarquera cependant que les calculs précédents sont tout à fait cohérents avec le résultat annoncé, car ils montrent que les dérivées secondes d'une solution du problème sont uniquement déterminées sur la courbe γ si, et seulement si, la courbe n'est caractéristique en aucun point.
Ce résultat a des conséquences importantes sur la régularité des solutions. Supposons que γ est une courbe qui sépare un domaine ouvert Ω en deux parties Ω_j, $j = 1, 2$, et supposons que, pour $j = 1, 2$, il existe une solution u_j de l'EDP de classe $\mathcal{C}^2$ définie sur Ω_j telle que $u_1 = u_2$ et $u_{1n} = u_{2n}$ sur γ. Alors, si γ n'est pas tangente à une courbe caractéristique, la fonction u définie sur Ω égale à u_j sur Ω_j pour $j = 1, 2$ est de classe $\mathcal{C}^2$.
Si γ est tangente à une courbe caractéristique, il se peut que les dérivées secondes de u_1 et u_2 soient différentes. C'est le phénomène de *propagation des singularités* que l'on va rencontrer dans certains exemples qui suivent.
On est capable de décrire les courbes caractéristiques dans le cas des EDP linéaires du second ordre.

Lemme 9. Soit Ω un domaine ouvert de $\mathbb{R}^2$.
1) Si $\forall (x, y) \in \Omega$, $a(x, y) \neq 0$, alors les courbes caractéristiques sont les courbes de la forme $y = y(x)$ avec $y(x)$ solution de l'équation différentielle :

$$a(x, y(x)) \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 - 2b(x, y(x)) \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) + c(x, y(x)) = 0 . \qquad (63)$$

2) Si $\forall (x, y) \in \Omega$, $c(x, y) \neq 0$, alors les courbes caractéristiques sont les courbes de la forme $x = x(y)$ avec $x(y)$ solution de l'équation différentielle :

$$a(x(y), y) \left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)^2 - 2b(x(y), y) \left(\frac{\mathrm{d}x}{\mathrm{d}y}\right) + c(x(y), y) = 0 . \qquad (64)$$

3) Si $\forall (x, y) \in \Omega$, $a(x, y) = 0$ et $c(x, y) = 0$, alors les courbes caractéristiques sont les droites verticales ou horizontales de la forme $y =$ constante ou $x =$ constante.

Démonstration.
Soit $\gamma = (x(t), y(t))_{t\in]0,1[}$ une courbe caractéristique (telle que $\forall t,\ \Delta(t) = 0$). On considère le premier cas. Si $x'(t) = 0$ pour un certain t, alors le fait que $a \neq 0$ implique que $y'(t) = 0$ d'après l'expression (62) de $\Delta(t)$, ce qui est impossible car γ est régulière. Donc $\forall t,\ x'(t) \neq 0$, la courbe γ n'a pas de tangente verticale, elle peut être définie par une courbe de la forme $y = f(x)$ avec $y'(t)/x'(t) = f'(x(t))$. L'équation $\Delta(t) = 0\ \forall t$ est alors équivalente à (63). Le second cas se traite de même en échangeant les rôles de x et y. Dans le troisième cas, l'équation $\Delta(t) = 0$ devient $b(x(t), y(t))x'(t)y'(t) = 0$. Comme on a affaire à une EDP du second ordre, on doit avoir $b \neq 0$ sur Ω, et donc $x'(t) = 0$ ou $y'(t) = 0$. ∎

Si $a \neq 0$ (dans le sens où $\forall (x, y) \in \Omega,\ a(x, y) \neq 0$), alors les courbes caractéristiques sont de la forme $y = y(x)$ où $y(x)$ est solution de (63). Cette équation différentielle exprime une relation entre $\mathrm{d}y/\mathrm{d}x$ et $y(x)$. Cette relation, vue comme une équation d'inconnue $\mathrm{d}y/\,\mathrm{d}x$, peut posséder deux, une, ou zéro solution réelle, selon le signe du discriminant de l'équation du second degré $aX^2 - 2bX + c = 0$.
Si $b^2 > ac$ dans Ω, alors $\mathrm{d}y/\mathrm{d}x$ peut être égal à deux fonctions de $y(x)$:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b(x, y(x)) \pm \sqrt{b^2(x, y(x)) - ac(x, y(x))}}{a(x, y(x))}, \tag{65}$$

et il y a donc deux équations différentielles ordinaires associées à l'équation (63) selon le signe choisi, ce qui donne deux familles de solutions de l'équation (63). Ce sont effectivement des familles car il faut encore spécifier les constantes d'intégration.
Si $b^2 = ac$ dans Ω, alors $\mathrm{d}y/\mathrm{d}x$ doit être égal à :

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b(x, y(x))}{a(x, y(x))}, \tag{66}$$

et il y a donc une unique famille de solutions de l'équation (63).
Si $b^2 < ac$ dans Ω, alors il n'y a pas de solution (réelle).

Définition 7. 1) Une EDP telle que $b^2(x, y) > a(x, y)c(x, y)$ dans un domaine Ω est dite *hyperbolique* dans Ω. Elle admet deux familles de courbes caractéristiques.
2) Une EDP telle que $b^2(x, y) = a(x, y)c(x, y)$ dans un domaine Ω est dite *parabolique* dans Ω. Elle admet une seule famille de courbes caractéristiques.
3) Une EDP telle que $b^2(x, y) < a(x, y)c(x, y)$ dans un domaine Ω est dite *elliptique* dans Ω. Elle n'admet pas de courbes caractéristiques.

Le caractère hyperbolique, parabolique ou elliptique d'une EDP s'appelle son type. Il ne dépend que des termes de plus haut degré, et nous allons voir dans la prochaine section que le type de l'équation est intrinsèque et ne dépend pas du système de coordonnées (x, y) choisi.

Exemples. 1) L'équation des ondes $u_{tt} - u_{xx} = 0$ est telle que $a = 1$, $b = 0$ et $c = -1$, c'est une EDP hyperbolique.

2) L'équation de Laplace $u_{xx} + u_{yy} = 0$ est telle que $a = 1$, $b = 0$ et $c = 1$, c'est une EDP elliptique.
3) L'équation de la chaleur $u_t = u_{xx}$ est telle que $a = 0$, $b = 0$ et $c = 1$, c'est une EDP parabolique.

On peut trouver dans la littérature une définition équivalente. Si l'on note

$$\mathbf{A}(x,y) := \begin{pmatrix} a(x,y) & b(x,y) \\ b(x,y) & c(x,y) \end{pmatrix},$$

alors on a les caractérisations suivantes :
1) Une EDP telle que la matrice $\mathbf{A}(x,y)$ a deux valeurs propres de signes opposés dans un domaine Ω est dite hyperbolique dans Ω.
2) Une EDP telle que la matrice $\mathbf{A}(x,y)$ a une valeur propre nulle dans un domaine Ω est dite parabolique dans Ω.
3) Une EDP telle que la matrice $\mathbf{A}(x,y)$ a deux valeurs propres de même signe dans un domaine Ω est dite elliptique dans Ω.
Le cas où la matrice $\mathbf{A}(x,y)$ a deux valeurs propres nulles est exclu, car cela veut dire que la matrice $\mathbf{A}(x,y)$ est nulle, et nous avons justement exclu le cas où les trois fonctions a, b, c s'annulent simultanément.
Il se trouve que cette caractérisation permet d'étendre la définition des types hyperbolique, parabolique et elliptique à des EDP en dimension plus grande que 2, mais nous n'aborderons pas ce cas dans ce module.
On peut enfin expliquer le choix des qualificatifs pour les trois types : pour une constante réelle r fixée, la courbe

$$\gamma_r = \left\{ (\xi,\eta) \in \mathbb{R}^2 \mid \begin{pmatrix} \xi \\ \eta \end{pmatrix}^t \mathbf{A}(x,y) \begin{pmatrix} \xi \\ \eta \end{pmatrix} = r \right\}$$

est une hyperbole lorsque la matrice $\mathbf{A}(x,y)$ a deux valeurs propres de signes opposés, une parabole lorsque la matrice $\mathbf{A}(x,y)$ a une valeur propre nulle, et une ellipse lorsque la matrice $\mathbf{A}(x,y)$ a deux valeurs propres de mêmes signes.
Le classement proposé ici des EDP linéaires du second ordre est intéressant car il reflète un ensemble de caractéristiques et résultats communs sur le plan des phénomènes physiques modélisés, sur le plan mathématique (existence, unicité, régularité,...) et sur le plan numérique (schéma de résolution propre à chaque type).

3.2 Réduction aux formes standards

Le but de cette section est double. On veut montrer que le type d'une équation est intrinsèque et ne dépend pas du système de coordonnées choisi. On veut aussi montrer que l'on peut ramener l'EDP (55) à une forme plus simple, et qu'il y a trois formes simples possibles, selon le type hyperbolique, parabolique ou elliptique de l'EDP.
On considère le changement de variables :

$$\xi := \varphi(x,y), \quad \eta := \psi(x,y). \tag{67}$$

On suppose que ce changement de variables constitue un $\mathcal{C}^2$-difféomorphisme, c'est-à-dire

que $\varphi, \psi \in \mathcal{C}^2(\Omega, \mathbb{R})$ et le jacobien de la transformation est non nul :

$$J(x,y) := \varphi_x \psi_y(x,y) - \varphi_y \psi_x(x,y) \neq 0\,, \quad \forall (x,y) \in \Omega. \tag{68}$$

Il existe alors une transformation réciproque de classe $\mathcal{C}^2$ telle que $x = \widetilde{\varphi}(\xi,\eta)$ et $y = \widetilde{\psi}(\xi,\eta)$. On peut donc associer de manière unique à une fonction u de (x,y) une fonction $\widetilde{u}$ de (ξ,η) en posant

$$\widetilde{u}(\xi,\eta) := u(\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))\,.$$

On retrouve u à partir de $\widetilde{u}$ par $u(x,y) = \widetilde{u}(\varphi(x,y), \psi(x,y))$. Si $u \in \mathcal{C}^2$, alors $\widetilde{u} \in \mathcal{C}^2$ car c'est une composée de fonctions de classe $\mathcal{C}^2$. En dérivant par rapport à x, et en appliquant les règles usuelles sur les dérivées des fonctions composées, on obtient :

$$u_x(x,y) = \widetilde{u}_\xi(\varphi(x,y), \psi(x,y))\varphi_x(x,y) + \widetilde{u}_\eta(\varphi(x,y), \psi(x,y))\psi_x(x,y)\,.$$

En répétant ces calculs, on trouve que les dérivées partielles de u et $\widetilde{u}$ sont liées par les relations :

$$\begin{aligned}
u_x &= \varphi_x \widetilde{u}_\xi + \psi_x \widetilde{u}_\eta\,,\\
u_y &= \varphi_y \widetilde{u}_\xi + \psi_y \widetilde{u}_\eta\,,\\
u_{xx} &= \varphi_x^2 \widetilde{u}_{\xi\xi} + 2\varphi_x\psi_x \widetilde{u}_{\xi\eta} + \psi_x^2 \widetilde{u}_{\eta\eta} + \varphi_{xx}\widetilde{u}_\xi + \psi_{xx}\widetilde{u}_\eta\,,\\
u_{yy} &= \varphi_y^2 \widetilde{u}_{\xi\xi} + 2\varphi_y\psi_y \widetilde{u}_{\xi\eta} + \psi_y^2 \widetilde{u}_{\eta\eta} + \varphi_{yy}\widetilde{u}_\xi + \psi_{yy}\widetilde{u}_\eta\,,\\
u_{xy} &= \varphi_x\varphi_y \widetilde{u}_{\xi\xi} + (\varphi_x\psi_y + \varphi_y\psi_x)\widetilde{u}_{\xi\eta} + \psi_x\psi_y \widetilde{u}_{\eta\eta} + \varphi_{xy}\widetilde{u}_\xi + \psi_{xy}\widetilde{u}_\eta\,,
\end{aligned}$$

où $u_.$, $\varphi_.$ et $\psi_.$ sont évaluées en (x,y) et les $\widetilde{u}_.$ sont évaluées en $(\varphi(x,y), \psi(x,y))$. En considérant ces équations en $x = \widetilde{\varphi}(\xi,\eta)$ et $y = \widetilde{\psi}(\xi,\eta)$, on trouve que les identités précédentes sont aussi vraies lorsque les $u_.$, $\varphi_.$ et $\psi_.$ sont évaluées en $(\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))$ et les $\widetilde{u}_.$ sont évaluées en (ξ,η). Cela implique que, si u est solution de l'EDP (55), alors $\widetilde{u}$ vérifie :

$$\widetilde{a}(\xi,\eta)\widetilde{u}_{\xi\xi} + 2\widetilde{b}(\xi,\eta)\widetilde{u}_{\xi\eta} + \widetilde{c}(\xi,\eta)\widetilde{u}_{\eta\eta} = \widetilde{d}(\xi,\eta)\widetilde{u}_\xi + \widetilde{e}(\xi,\eta)\widetilde{u}_\eta + \widetilde{r}(\xi,\eta)\widetilde{u} + \widetilde{s}(\xi,\eta)\,, \tag{69}$$

où l'on définit les fonctions :

$$\widetilde{a}(\xi,\eta) := [\varphi_x^2 a + 2\varphi_x\varphi_y b + \varphi_y^2 c](\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))\,, \tag{70}$$

$$\widetilde{b}(\xi,\eta) := [\varphi_x\psi_x a + (\varphi_x\psi_y + \varphi_y\psi_x)b + \varphi_y\psi_y c](\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))\,, \tag{71}$$

$$\widetilde{c}(\xi,\eta) := [\psi_x^2 a + 2\psi_x\psi_y b + \psi_y^2 c](\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))\,, \tag{72}$$

$$\widetilde{d}(\xi,\eta) := [\varphi_{xx} a + 2\varphi_{xy} b + \varphi_{yy} c + \varphi_x d + \varphi_y e](\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))\,,$$

$$\widetilde{e}(\xi,\eta) := [\psi_{xx} a + 2\psi_{xy} b + \psi_{yy} c + \psi_x d + \psi_y e](\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))\,,$$

$$\widetilde{r}(\xi,\eta) := r(\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))\,,$$

$$\widetilde{s}(\xi,\eta) := s(\widetilde{\varphi}(\xi,\eta), \widetilde{\psi}(\xi,\eta))\,.$$

La réciproque est vraie bien entendu, à savoir, si $\widetilde{u}$ satisfait (69), alors u satisfait (55). On peut maintenant remarquer que :

$$\widetilde{a}\widetilde{c} - \widetilde{b}^2 = J^2[ac - b^2]\,. \tag{73}$$

Comme $J^2 > 0$, cela montre que le signe de $ac - b^2$ est conservé par n'importe quel changement de variables. Par conséquent, le type d'une EDP linéaire du second ordre est intrinsèque à l'EDP et ne dépend pas du choix de coordonnées.

Exercice 3.
Vérifier (73).
Solution. On substitue les expressions (70-72) dans le membre de gauche de (73), on trouve que l'on peut mettre J^2 défini par (68) en facteur et l'on obtient le résultat voulu.

Nous allons montrer dans les trois sous-sections qui suivent que l'on peut choisir φ et ψ de telle manière qu'au moins un (voire deux) des trois termes $\widetilde{a}$, $\widetilde{b}$ ou $\widetilde{c}$ est nul. Pour fixer les choses, on suppose dorénavant que $a \neq 0$ dans Ω. Le lemme suivant éclaire la manière dont on va procéder, en soulignant que les changements de variables appropriés sont liés aux courbes caractéristiques.

Lemme 10. Supposons que $a \neq 0$ dans Ω. Un changement de variables de la forme (67) est tel que $\widetilde{a} = 0$ si, et seulement si, les lignes de niveau de φ sont des courbes caractéristiques.

La ligne de niveau de φ au niveau $\varphi_0 \in \mathbb{R}$ est définie comme l'ensemble des points :

$$\gamma := \{(x,y) \in \Omega \mid \varphi(x,y) = \varphi_0\}\,.$$

Démonstration. Supposons qu'il existe un changement de variables de la forme (67) tel que $\widetilde{a} = 0$, i.e. :

$$\forall (x,y) \in \Omega\,, \quad a\varphi_x^2 + 2b\varphi_x\varphi_y + c\varphi_y^2 = 0\,. \tag{74}$$

Clairement, si $(x,y) \in \Omega$, alors $\varphi_y(x,y)$ n'est pas nul, car si $\varphi_y(x,y) = 0$, alors on aurait $a(x,y)\varphi_x^2(x,y) = 0$, donc $\varphi_x(x,y) = 0$ et cela contredit $J(x,y) \neq 0$. En divisant (74) par $\varphi_y^2(x,y)$, on trouve que $\varphi_x/\varphi_y(x,y)$ est solution de l'équation du second degré :

$$a(x,y)\left(\frac{\varphi_x}{\varphi_y}\right)^2 + 2b(x,y)\left(\frac{\varphi_x}{\varphi_y}\right) + c(x,y) = 0\,. \tag{75}$$

Montrons que les lignes de niveau de φ sont des courbes caractéristiques. Soient $\varphi_0 \in \mathbb{R}$ et γ la ligne de niveau de φ au niveau φ_0. Si $\gamma \neq \varnothing$, alors il existe $(x_0, y_0) \in \Omega$ tel que $\varphi(x_0, y_0) = \varphi_0$. D'après le théorème des fonctions implicites (*cf.* le livre [L2]), comme $\varphi_y(x_0, y_0) \neq 0$, il existe une unique fonction $y(x)$ telle que $\varphi(x, y(x)) = \varphi_0$ pour tout x dans un voisinage I de x_0. Donc l'ensemble γ est une courbe $(x, y(x))_{x \in I}$. La fonction $y(x)$ vérifie $y(x_0) = y_0$. Elle est de classe $\mathcal{C}^1$ et, en dérivant la relation $\varphi(x, y(x)) = \varphi_0$, on voit qu'elle vérifie l'équation différentielle :

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{\varphi_x}{\varphi_y}(x, y(x))\,.$$

En substituant dans (75), on trouve que $y(x)$ vérifie (63), ce qui montre que γ est une courbe caractéristique.

Réciproquement, supposons qu'il existe un $\mathcal{C}^2$-difféomorphisme φ dont les lignes de niveau sont des courbes caractéristiques. On considère le changement de variables $\xi := \varphi(x,y)$ et $\eta := \psi(x,y)$ où le $\mathcal{C}^2$-difféomorphisme ψ est choisi arbitrairement, du moment que le jacobien de la transformation $(x,y) \mapsto (\varphi(x,y), \psi(x,y))$ ne s'annule pas. Pour tout $(x_0, y_0) \in \Omega$, la ligne de niveau de φ au niveau $\varphi(x_0, y_0)$ est une courbe caractéristique par hypothèse, donc elle est de la forme $y = y(x)$ avec $y(x)$ solution de l'équation différentielle (63) d'après le lemme 9.

On a donc $\varphi(x, y(x)) = \varphi(x_0, y_0)$ pour tout x dans un voisinage de x_0, ce qui implique en particulier $\varphi_x(x_0, y_0) + \varphi_y(x_0, y_0)(\mathrm{d}y/\,\mathrm{d}x)(x_0) = 0$. Comme $\mathrm{d}y/\,\mathrm{d}x$ vérifie (63), on trouve que φ_x/φ_y vérifie (75), et donc $\widetilde{a}$ est nul d'après sa définition (70). On a donc montré que le changement de variables $\xi = \varphi(x,y)$ et $\eta = \psi(x,y)$ produit une EDP de la forme (69) dont le coefficient $\widetilde{a}$ devant $\widetilde{u}_{\xi\xi}$ est nul. ■

3.2.1 EDP hyperbolique $b^2 > ac$

Il existe deux familles de courbes caractéristiques, décrites par les équations $\varphi(x,y) = \varphi_0$ et $\psi(x,y) = \psi_0$ obtenues en intégrant (65) avec le signe $+$ et avec le signe $-$, respectivement. Le changement de variables $\xi := \varphi(x,y)$ et $\eta := \psi(x,y)$ produit alors une EDP dont les coefficients $\widetilde{a}$ et $\widetilde{c}$ devant $u_{\xi\xi}$ et $u_{\eta\eta}$ sont nuls. De plus, on a $\widetilde{b} \neq 0$, car $\widetilde{b}^2 = J^2(b^2 - ac) > 0$ d'après (73). L'équation (69) se simplifie alors en :

$$\widetilde{u}_{\xi\eta} = H_1(\widetilde{u}_\xi, \widetilde{u}_\eta, \widetilde{u}, \xi, \eta)\,, \qquad H_1 := \frac{\widetilde{d}}{2\widetilde{b}}\widetilde{u}_\xi + \frac{\widetilde{e}}{2\widetilde{b}}\widetilde{u}_\eta + \frac{\widetilde{r}}{2\widetilde{b}}\widetilde{u} + \frac{\widetilde{s}}{2\widetilde{b}}\,. \tag{76}$$

C'est la première forme canonique d'une EDP hyperbolique. Il en existe une autre, que l'on obtient après le nouveau changement de variables :

$$\alpha := \xi + \eta\,, \quad \beta := \xi - \eta\,, \quad \widehat{u}(\alpha,\beta) := \widetilde{u}\left(\frac{\alpha+\beta}{2}, \frac{\alpha-\beta}{2}\right). \tag{77}$$

On trouve :

$$(\widehat{u}_{\alpha\alpha} - \widehat{u}_{\beta\beta})(\alpha,\beta) = \widetilde{u}_{\eta\xi}\left(\frac{\alpha+\beta}{2}, \frac{\alpha-\beta}{2}\right),$$

et l'on obtient ainsi la seconde forme canonique d'une EDP hyperbolique :

$$\widehat{u}_{\alpha\alpha} - \widehat{u}_{\beta\beta} = H_2(\widehat{u}_\alpha, \widehat{u}_\beta, \widehat{u}, \alpha, \beta)\,, \tag{78}$$

où H_2 s'obtient à partir de H_1 par le changement de variables (77).

Cas particulier. Supposons les coefficients a, b, c constants, avec $b^2 > ac$.

Les deux familles de courbes caractéristiques sont des droites décrites par les équations $\varphi(x,y) = \varphi_0$ et $\psi(x,y) = \psi_0$ avec

$$\varphi(x,y) = y - \frac{b + \sqrt{b^2 - ac}}{a}x\,, \qquad \psi(x,y) = y - \frac{b - \sqrt{b^2 - ac}}{a}x\,.$$

Le changement de variables approprié pour obtenir la première forme canonique est :

$$\xi := y - \frac{b + \sqrt{b^2 - ac}}{a}x\,, \qquad \eta := y - \frac{b - \sqrt{b^2 - ac}}{a}x\,.$$

Le changement de variables approprié pour obtenir la seconde forme canonique est :

$$\alpha := 2y - 2\frac{b}{a}x\,, \qquad \beta := -2\frac{\sqrt{b^2-ac}}{a}x\,.$$

Exemple. Considérons l'EDP $x^2u_{xx} - u_{yy} = u$, $(x,y) \in \,]0,+\infty[\, \times \mathbb{R}$.
Les courbes caractéristiques sont les courbes $(x, y(x))$ avec $x^2(\mathrm{d}y/\,\mathrm{d}x)^2 - 1 = 0$, c'est-à-dire $\mathrm{d}y/\,\mathrm{d}x = \pm 1/x$. Il y a deux familles de courbes caractéristiques : $y(x) = \ln x +$constante et $y(x) = -\ln x +$constante. Le changement de variables à adopter pour réduire l'EDP à la première forme canonique est donc $\xi := y - \ln x$ et $\eta := y + \ln x$. La transformation réciproque est $x = \exp((\eta-\xi)/2)$ et $y = (\xi+\eta)/2$. En posant :

$$\widetilde{u}(\xi,\eta) := u\left(\exp(\frac{\eta-\xi}{2}), \frac{\eta+\xi}{2}\right), \qquad u(x,y) = \widetilde{u}(y - \ln x, y + \ln x)\,,$$

et en calculant $\widetilde{b} = -2$, $\widetilde{d} = -1$, $\widetilde{e} = 1$, $\widetilde{r} = 1$, et $\widetilde{s} = 0$, on obtient que u satisfait $x^2u_{xx} - u_{yy} = u$ si, et seulement si, $\widetilde{u}$ satisfait :

$$\widetilde{u}_{\xi\eta} = \frac{1}{4}\widetilde{u}_\xi - \frac{1}{4}\widetilde{u}_\eta - \frac{1}{2}\widetilde{u}\,.$$

3.2.2 EDP parabolique $b^2 = ac$

Il existe une unique famille de courbes caractéristiques, décrite par l'équation $\varphi(x,y) = \varphi_0$ obtenue en intégrant (66). On peut prendre n'importe quelle fonction ψ, du moment que le jacobien J soit non nul. Par exemple, $\psi(x,y) = y$ convient. Le coefficient $\widetilde{a}$ est alors nul. En notant δ_b le signe de b et δ_a le signe de a (qui est aussi celui de c, car $ac = b^2 > 0$), on en déduit que :

$$\begin{aligned} 0 = \widetilde{a} &= a\varphi_x^2 + 2b\varphi_x\varphi_y + c\varphi_y^2 = a\varphi_x^2 + 2\delta_b\sqrt{ac}\varphi_x\varphi_y + c\varphi_y^2 \\ &= 2\delta_a\big[\varphi_x\sqrt{|a|} + (\delta_b/\delta_a)\varphi_y\sqrt{|c|}\big]^2\,, \end{aligned}$$

et donc $\varphi_x\sqrt{|a|} + (\delta_b/\delta_a)\varphi_y\sqrt{|c|} = 0$. Cela implique que $\widetilde{b}$ est nul aussi, comme le montre le calcul suivant :

$$\begin{aligned} \widetilde{b} &= 2[\varphi_x\psi_x a + (\varphi_x\psi_y + \varphi_y\psi_x)\delta_b\sqrt{ac} + \varphi_y\psi_y c] \\ &= 2\delta_a[\varphi_x\sqrt{|a|} + (\delta_b/\delta_a)\varphi_y\sqrt{|c|}][\psi_x\sqrt{|a|} + (\delta_b/\delta_a)\psi_y\sqrt{|c|}] = 0\,. \end{aligned}$$

On obtient ainsi la forme canonique d'une EDP parabolique :

$$\widetilde{u}_{\eta\eta} = H_3(\widetilde{u}_\xi, \widetilde{u}_\eta, \widetilde{u}, \xi, \eta)\,, \qquad H_3 := \frac{\widetilde{d}}{\widetilde{c}}\widetilde{u}_\xi + \frac{\widetilde{e}}{\widetilde{c}}\widetilde{u}_\eta + \frac{\widetilde{r}}{\widetilde{c}}\widetilde{u} + \frac{\widetilde{s}}{\widetilde{c}}\cdot \tag{79}$$

Cas particulier. Supposons les coefficients a, b, c constants, avec $b^2 = ac$.
La famille de courbes caractéristiques est constituée des droites décrites par les équations $\varphi(x,y) = \varphi_0$ avec :

$$\varphi(x,y) = y - \frac{b}{a}x\,.$$

Un changement de variables approprié pour obtenir la forme canonique est :

$$\xi := y - \frac{b}{a}x\,, \qquad \eta := y\,.$$

Exemple. On considère l'EDP $u_{xx} + 2u_{xy} + u_{yy} = u_x + u_y$ dans $\mathbb{R}^2$. Il y a une unique famille de courbes caractéristiques, d'équation $y - x =$constante.
On effectue donc le changement de variables $\xi := y - x$ et $\eta := y$. En posant :

$$\widetilde{u}(\xi,\eta) := u(\eta - \xi, \eta)\,, \qquad u(x,y) = \widetilde{u}(y - x, y)\,,$$

on obtient que u satisfait $u_{xx} + 2u_{xy} + u_{yy} = u_x + u_y$ si, et seulement si, $\widetilde{u}$ satisfait :

$$\widetilde{u}_{\eta\eta} = \widetilde{u}_\eta\,.$$

Exercice 4.
Déterminer dans l'exemple précédent la forme générale de la solution.
Solution. On commence par déterminer la forme générale de la solution de l'équation canonique. Il s'agit ici d'une équation différentielle ordinaire du second ordre, dont la solution est $\widetilde{u}(\xi,\eta) = f(\xi) + g(\xi)\exp(\eta)$, où f et g sont des fonctions arbitraires de classe $\mathcal{C}^2$. Donc la forme générale de la solution de l'EDP parabolique est :

$$u(x,y) = f(y - x) + g(y - x)\exp(y)\,.$$

3.2.3 EDP elliptique $b^2 < ac$

Il n'existe alors pas de courbe caractéristique. La conséquence est que l'on n'arrive pas à ramener l'EDP à une forme aussi simple que (76) ou (79) dans laquelle il n'y a qu'un terme avec une dérivée seconde. Mais on peut quand même se ramener à une forme dans laquelle il n'y a que deux termes avec une dérivée seconde, à savoir :

$$\widehat{u}_{\alpha\alpha} + \widehat{u}_{\beta\beta} = H_4(\widehat{u}_\alpha, \widehat{u}_\beta, u, \alpha, \beta)\,, \tag{80}$$

que l'on appelle forme canonique d'une EDP elliptique. On n'en fera pas la preuve complète, qui consiste à manipuler des complexes, car l'équation du second degré (63) d'inconnue $\mathrm{d}y/\mathrm{d}x$ admet deux solutions complexes conjuguées, qui permettent de définir deux familles de courbes à valeurs complexes solutions de (65), l'une étant conjuguée de l'autre. Formellement, les calculs faits dans le paragraphe sur le cas hyperbolique sont encore valables. Les calculs de changement de variables peuvent être menés en prenant les parties réelle et imaginaire de $\xi = \alpha + i\beta$ et $\eta = \alpha - i\beta$ en posant $\alpha = (\xi + \eta)/2$ et $\beta = (\xi - \eta)/(2i)$. La fonction :

$$\widehat{u}(\alpha,\beta) := \widetilde{u}(\alpha + i\beta, \alpha - i\beta)\,, \qquad \widetilde{u}(\xi,\xi^*) := \widehat{u}((\xi + \xi^*)/2, (\xi - \xi^*)/(2i))\,,$$

est alors solution d'une EDP de la forme :

$$\widehat{a}(\alpha,\beta)\widehat{u}_{\alpha\alpha} + 2\widehat{b}(\alpha,\beta)u_{\alpha\beta} + \widehat{c}(\alpha,\beta)u_{\beta\beta} = \widehat{d}(\alpha,\beta)\widehat{u}_\alpha + \widehat{e}(\alpha,\beta)\widehat{u}_\beta + \widehat{r}(\alpha,\beta)\widehat{u} + \widehat{s}(\alpha,\beta)\,,$$

où les fonctions $\widehat{a}, \ldots, \widehat{s}$ vont de $\mathbb{R}^2$ dans $\mathbb{R}$ avec en particulier :

$$\widehat{a}(\alpha,\beta) := \frac{1}{4}[\widetilde{a} + \widetilde{c} + 2\widetilde{b}](\alpha + i\beta, \alpha - i\beta)\,,$$
$$\widehat{b}(\alpha,\beta) := \frac{1}{2i}[\widetilde{a} - \widetilde{c}](\alpha + i\beta, \alpha - i\beta)\,,$$
$$\widehat{c}(\alpha,\beta) := \frac{1}{4}[-\widetilde{a} - \widetilde{c} + 2\widetilde{b}](\alpha + i\beta, \alpha - i\beta)\,.$$

Or $\widetilde{a} = \widetilde{c} = 0$, donc $\widehat{b} = 0$ et $\widehat{a} = \widehat{c} \neq 0$. En divisant l'EDP par $\widehat{a}$, on trouve la forme canonique (80) d'une EDP elliptique. Ces arguments sont formels mais peuvent être justifiés. Dans le cadre de ce module, il suffit de savoir appliquer la méthode de réduction à la forme standard telle que mise en oeuvre dans le cas de coefficients constants ou dans l'exemple qui suit.

Cas particulier. Supposons les coefficients a, b, c constants, avec $b^2 < ac$.
Les deux familles de courbes caractéristiques complexes sont des droites décrites par les équations $\varphi(x,y) = \varphi_0$ et $\psi(x,y) = \psi_0$ avec

$$\varphi(x,y) = y - \frac{b + i\sqrt{ac - b^2}}{a}x\,, \qquad \varphi(x,y) = y - \frac{b - i\sqrt{ac - b^2}}{a}x\,.$$

Le changement de variables approprié pour obtenir la forme canonique est :

$$\alpha := y - \frac{b}{a}x\,, \qquad \beta := \frac{\sqrt{ac - b^2}}{a}x\,.$$

Exemple. Considérons l'EDP $x^2 u_{xx} + u_{yy} = u$ pour $(x,y) \in \,]0,+\infty[\, \times \mathbb{R}$. Les courbes caractéristiques sont les courbes complexes $(x, y(x))$ avec $x^2(\mathrm{d}y/\,\mathrm{d}x)^2 + 1 = 0$, c'est-à-dire $\mathrm{d}y/\,\mathrm{d}x = \pm i/x$. Il y a deux familles de courbes caractéristiques : $y(x) = i\ln x +$constante et $y(x) = -i\ln x +$constante. On a donc $\xi := y - i\ln x$ et $\eta := y + i\ln x$, et le changement de variables à adopter pour réduire l'EDP à sa forme canonique est $\alpha := y$ et $\beta := -\ln x$. La transformation réciproque est $x = \exp(-\beta)$ et $y = \alpha$. En posant :

$$\widehat{u}(\alpha,\beta) := u(\exp(-\beta), \alpha)\,, \qquad u(x,y) = \widehat{u}(y, -\ln x)\,,$$

on obtient que u satisfait $x^2 u_{xx} + u_{yy} = u$ si, et seulement si, $\widehat{u}$ satisfait :

$$\widehat{u}_{\alpha\alpha} + \widehat{u}_{\beta\beta} = -\widehat{u}_\beta - \widehat{u}\,.$$

4 Équations linéaires d'ordre 2 à coefficients constants

Dans cette section, on va étudier en détail quelques exemples d'EDP linéaires d'ordre 2 à coefficients constants. Ces exemples sont issus de la physique.

4.1 L'équation des ondes uni-dimensionnelle

On considère ici l'équation des ondes uni-dimensionnelle :

$$u_{tt} - c^2 u_{xx} = 0\,, \qquad (t,x) \in \,]0,+\infty[\, \times \mathbb{R}\,, \tag{81}$$

où c est une constante strictement positive. Cette équation décrit la propagation d'une onde. Elle modélise par exemple la propagation d'une vague à la surface de l'océan ($u(t,x)$ est la hauteur de la vague au point x à l'instant t) ou la propagation d'un signal électrique ou électromagnétique dans un câble ($u(t,x)$ est la valeur du champ électrique au point x à l'instant t).

On va commencer par rechercher la solution générale de cette EDP. Il s'agit d'une EDP hyperbolique, ses courbes caractéristiques sont les droites $x + ct$ =constante et $x - ct$ =constante. Supposons que u soit une solution de classe $\mathcal{C}^2$ et appliquons le changement de variables adéquat :

$$\xi = x + ct\,, \quad \eta = x - ct\,.$$

La fonction $\widetilde{u}(\xi,\eta) := u((\xi-\eta)/(2c),(\xi+\eta)/2)$ vérifie :

$$\widetilde{u}_{\xi\eta} = 0\,.$$

En intégrant par rapport à η cette équation, on trouve que $\widetilde{u}_\xi$ est une constante en η, c'est-à-dire que c'est une fonction qui ne dépend que de ξ :

$$\widetilde{u}_\xi(\xi,\eta) = \varphi(\xi)\,.$$

En intégrant par rapport à ξ, on trouve que :

$$\widetilde{u}(\xi,\eta) = \Phi(\xi) + \Psi(\eta)\,,$$

où Φ est une primitive de φ, et Ψ est une constante d'intégration, qui ne dépend pas de ξ, et qui est donc une fonction de η seulement. En revenant aux variables originales (t,x), on obtient que u est de la forme :

$$u(t,x) = \Phi(x+ct) + \Psi(x-ct)\,. \tag{82}$$

Réciproquement, si u est de la forme précédente, avec Φ et Ψ deux fonctions de classe $\mathcal{C}^2$, alors :

$$u_{tt}(t,x) = c^2\Phi''(x+ct) + c^2\Psi''(x-ct)\,,$$
$$u_{xx}(t,x) = \Phi''(x+ct) + \Psi''(x-ct)\,,$$

ce qui montre que u est solution de l'EDP (81).

On va maintenant examiner le problème de Cauchy pour l'équation des ondes (81) dans le domaine $(t,x) \in \,]0,+\infty[\, \times \mathbb{R}$ avec les conditions initiales sur le bord $\{0\} \times \mathbb{R}$ de ce domaine :

$$u(0,x) = f(x)\,, \quad u_t(0,x) = g(x)\,, \quad x \in \mathbb{R}\,. \tag{83}$$

Proposition 11.

L'équation des ondes uni-dimensionnelle (81) pour $(t,x) \in \,]0,+\infty[\, \times \mathbb{R}$ avec les conditions initiales (83) à $t = 0$ admet une unique solution dans $\mathcal{C}^2(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R})$ donnée par :

$$u(t,x) = \frac{1}{2}\left[f(x-ct) + f(x+ct)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(s)\,\mathrm{d}s\,, \tag{84}$$

si $f \in \mathcal{C}^2(\mathbb{R},\mathbb{R})$ et $g \in \mathcal{C}^1(\mathbb{R},\mathbb{R})$.

Noter que l'on peut aussi écrire la solution sous la forme (82) avec :

$$\Phi(x) = \frac{1}{2}f(x) + \frac{1}{2c}\int_0^x g(s)\,\mathrm{d}s\,, \tag{85}$$

$$\Psi(x) = \frac{1}{2}f(x) - \frac{1}{2c}\int_0^x g(s)\,\mathrm{d}s\,. \tag{86}$$

Démonstration. Soit u une solution de classe $\mathcal{C}^2$, alors c'est une fonction de la forme (82). Les conditions initiales (83) vont nous permettre d'identifier les fonctions Φ et Ψ. En effet, ces fonctions doivent vérifier :

$$u(0,x) = f(x) \Longrightarrow \Phi(x) + \Psi(x) = f(x) \;\; \forall x \in \mathbb{R}\,, \tag{87}$$

$$u_t(0,x) = g(x) \Longrightarrow c\Phi'(x) - c\Psi'(x) = g(x) \;\; \forall x \in \mathbb{R}\,. \tag{88}$$

On intègre la seconde équation différentielle :

$$\Phi(x) - \Psi(x) = \frac{1}{c}\int_0^x g(s)\,\mathrm{d}s + C_0\,, \tag{89}$$

où C_0 est une constante d'intégration. En sommant (87) avec (89), on obtient :

$$\Phi(x) = \frac{1}{2}f(x) + \frac{1}{2c}\int_0^x g(s)\,\mathrm{d}s + \frac{1}{2}C_0\,.$$

En soustrayant (89) de (87), on obtient :

$$\Psi(x) = \frac{1}{2}f(x) - \frac{1}{2c}\int_0^x g(s)\,\mathrm{d}s - \frac{1}{2}C_0\,.$$

La décomposition (82) n'est clairement pas unique, car on peut ajouter une constante à Ψ et la retrancher à Φ sans changer le résultat. En choisissant cette constante égale à $C_0/2$, on obtient (85) et (86).

Réciproquement, la fonction u définie par (84) est la somme d'une fonction de $x + ct$ et d'une autre fonction de $x - ct$, qui sont toutes deux de classe $\mathcal{C}^2$. Elle satisfait donc l'équation des ondes (81). De plus, on vérifie immédiatement que $u(0,x) = f(x)$, et :

$$u_t(t,x) = \frac{c}{2}\left[-f'(x-ct) + f'(x+ct)\right] + \frac{1}{2}\left[g(x+ct) + g(x-ct)\right]\,, \tag{90}$$

ce qui montre que $u_t(0,x) = g(x)$. La fonction (84) est donc solution du problème de Cauchy. ■

Si f et g sont analytiques, alors la solution (84) de l'équation des ondes est analytique, comme le prédit le théorème de Cauchy-Kowalewski.

Si $f \in \mathcal{C}^{2+k}$ et $g \in \mathcal{C}^{1+k}$, $k \geqslant 0$, alors la solution de l'équation des ondes appartient à $\mathcal{C}^{2+k}$, mais pas plus. Ainsi, l'équation des ondes ne provoque aucune régularisation de la solution, dans le sens où la solution à un instant $t > 0$ n'est pas plus régulière que la condition initiale à $t = 0$. Cette propriété est caractéristique des EDP hyperboliques, et l'on verra en étudiant l'équation de la chaleur page 422 que la situation est très différente dans le cas des EDP paraboliques.

En examinant l'expression complète de la solution (84), appelée formule de d'Alembert, on constate que la solution u à l'instant t et au point x ne dépend que des valeurs des fonctions initiales f et g dans l'intervalle $[x - ct, x + ct]$. On obtient ainsi le corollaire suivant.

Corollaire 12. Si les conditions initiales f et g sont à support compact dans $[x_0, x_1]$, alors, à tout instant $t > 0$, la solution de l'équation des ondes est à support compact dans $[x_0 - ct, x_1 + ct]$.

Cette proposition se généralise à l'équation des ondes en plusieurs dimensions, et est en fait spécifique des systèmes hyperboliques. On dit que les conditions initiales se propagent à vitesse finie (ici, c).

Corollaire 13. Sous les mêmes hypothèses que pour la proposition 11, en supposant de plus que les conditions initiales sont telles que f' et $g \in L^2(\mathbb{R})$, alors l'énergie de l'onde :

$$E(t) := \frac{1}{2}\int_{\mathbb{R}} u_x(t,x)^2 + \frac{1}{c^2}u_t(t,x)^2 \,\mathrm{d}x$$

est une quantité conservée au cours du temps égale à E_0 :

$$E_0 := \frac{1}{2}\int_{\mathbb{R}} f'(x)^2 + \frac{1}{c^2}g(x)^2 \,\mathrm{d}x\,.$$

Démonstration. En utilisant la représentation de d'Alembert de la solution, on trouve que u_t est donnée par (90), et u_x est donnée par :

$$u_x(t,x) = \frac{1}{2}\left[f'(x-ct) + f'(x+ct)\right] + \frac{1}{2c}\left[g(x+ct) - g(x-ct)\right]\,. \tag{91}$$

La densité d'énergie est définie par $e(t,x) := (1/2)[u_x(t,x)^2 + c^{-2}u_t(t,x)^2]$. En utilisant les expressions (90) et (91), en développant les carrés et en simplifiant, on trouve que la densité d'énergie est de la forme :

$$e(t,x) = e_1(x-ct) + e_2(x+ct)\,,$$

$$e_1(x) = \frac{1}{4}\left[f'(x) - \frac{1}{c}g(x)\right]^2, \qquad e_2(x) = \frac{1}{4}\left[f'(x) + \frac{1}{c}g(x)\right]^2.$$

Comme f' et g sont supposées appartenir à $L^2(\mathbb{R})$, la densité d'énergie est intégrable en x, et son intégrale vaut :

$$E(t) = \int_{\mathbb{R}} e_1(x-ct)\,\mathrm{d}x + \int_{\mathbb{R}} e_2(x+ct)\,\mathrm{d}x = \int_{\mathbb{R}} e_1(x) + e_2(x)\,\mathrm{d}x = E_0\,. \qquad \blacksquare$$

Examinons la solution de d'Alembert d'un point de vue un peu plus physique. On a vu que la solution pouvait s'écrire sous la forme (82) :

$$u(t,x) = \Phi(x-ct) + \Psi(x+ct)\,.$$

Cela veut dire qu'elle comprend un « mode » Φ qui se déplace vers la droite à la vitesse c, et un « mode » Ψ qui se déplace vers la gauche à la vitesse $-c$. Considérons tout d'abord

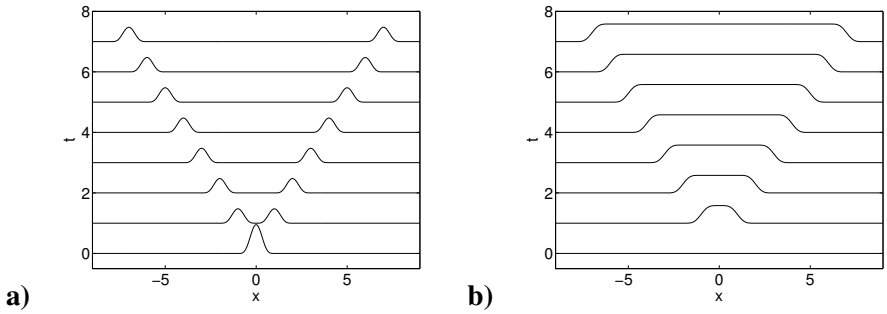


FIGURE III.2.3 – Propagation d'une onde uni-dimensionnelle. On trace les profils spatiaux de la solution aux instants $t = 0, 1, 2, \ldots, 7$, en prenant pour conditions initiales à $t = 0$: $f(x) = f_0(x)$, $g(x) = 0$ (figure a) et $f(x) = 0$, $g(x) = f_0(x)$ (figure b). On a pris ici $c = 1$ et la fonction $f_0(x) = \cos^4(\pi x/2)\mathbf{1}_{[-1,1]}(x)$ est à support compact et de classe $\mathcal{C}^2$ sur $\mathbb{R}$.

le cas $g = 0$. Alors la solution est simplement donnée par :

$$u(t,x) = \frac{1}{2}\left[f(x-ct) + f(x+ct)\right] .$$

Si la fonction f est localisée, disons à support compact, alors on constate (figure III.2.3a) que la condition initiale f se décompose en deux ondes qui ont la même forme que f, avec chacune la moitié de l'amplitude initiale. De plus, la solution u revient à son état original, c'est-à-dire :

$$u(t,x) \overset{t\to+\infty}{\longrightarrow} 0 ,$$

quel que soit $x \in \mathbb{R}$ fixé.
Si l'on suppose maintenant que $f = 0$, alors :

$$u(t,x) = \frac{1}{2c}\int_{x-ct}^{x+ct} g(s)\,\mathrm{d}s .$$

Ici aussi, on constate (figure III.2.3b) que la condition initiale g se décompose en deux ondes qui se propagent vers la droite et vers la gauche.
Cependant, si $g \in L^1$ et $\displaystyle\int_{\mathbb{R}} g(s)\,\mathrm{d}s \neq 0$, alors la solution ne revient pas à son état original après le passage de l'onde, mais converge comme suit :

$$u(t,x) \overset{t\to+\infty}{\longrightarrow} \frac{1}{2c}\int_{\mathbb{R}} g(s)\,\mathrm{d}s ,$$

pour tout $x \in \mathbb{R}$ fixé (voir la figure III.2.3b).

La formule de d'Alembert est utile pour trouver la solution de l'équation des ondes dans d'autres domaines que $]0,+\infty[\times\mathbb{R}$. Considérons le problème de Dirichlet :

$$u_{tt}-c^2u_{xx}=0\,,\quad (t,x)\in\,]0,+\infty[\,\times\,]0,+\infty[\,, \tag{92}$$

$$u(0,x)=f(x)\,,\quad u_t(0,x)=g(x)\,,\quad x\in\mathbb{R}_+\,, \tag{93}$$

$$u(t,0)=0\,,\quad t\in\mathbb{R}_+\,, \tag{94}$$

où $f\in\mathcal{C}^2(\mathbb{R}_+,\mathbb{R})$ et $g\in\mathcal{C}^1(\mathbb{R}_+,\mathbb{R})$ avec $f(0)=g(0)=0$.
Cette équation modélise la propagation d'une onde avec une paroi réfléchissante en $x=0$, qui empêche tout mouvement $u(t,x=0)=0$. La condition $f(0)=g(0)=0$ est une condition de compatibilité nécessaire pour qu'il existe une solution u. En effet, si une solution u existe, alors $0=u(0,0)=f(0)$ et $0=u_t(0,0)=g(0)$.
On peut convertir le problème (92-94) en le mettant sous la forme (81-83) en prolongeant $f(\cdot)$, $g(\cdot)$ et $u(t,\cdot)$ sur $\mathbb{R}$ tout entier par réflexion impaire. Plus précisément, on introduit les fonctions :

$$\widetilde{f}(x)=\begin{cases} f(x) & \text{si } x\geqslant 0\,,\\ -f(-x) & \text{si } x<0\,,\end{cases}\qquad \widetilde{g}(x)=\begin{cases} g(x) & \text{si } x\geqslant 0\,,\\ -g(-x) & \text{si } x<0\,.\end{cases}$$

La fonction $\widetilde{f}$ est définie sur $\mathbb{R}$, elle est de classe $\mathcal{C}^2$ sur $]-\infty,0[$ et sur $]0,+\infty[$. De plus, $\widetilde{f}(0^+)=\widetilde{f}(0^-)=0$, $\widetilde{f}'(0^+)=f'(0)=\widetilde{f}'(0^-)$, ce qui montre que $\widetilde{f}$ est de classe $\mathcal{C}^1$ sur $\mathbb{R}$. De même, $\widetilde{g}\in\mathcal{C}^1(\mathbb{R},\mathbb{R})$. Cependant, $\widetilde{f}''(0^+)=f''(0)$ et $\widetilde{f}''(0^-)=-f''(0)$, donc $\widetilde{f}$ est de classe $\mathcal{C}^2$ sur $\mathbb{R}$ si, et seulement si, $f''(0)=0$.
Supposons maintenant que $f''(0)=0$ (on envisagera le cas $f''(0)\neq 0$ à la fin de la section). Si $u\in\mathcal{C}^2(\mathbb{R}_+\times\mathbb{R}_+,\mathbb{R})$ est solution de (92-94), alors $\widetilde{u}$ définie par :

$$\widetilde{u}(t,x)=\begin{cases} u(t,x) & \text{si } x\geqslant 0\,,\\ -u(t,-x) & \text{si } x<0\,,\end{cases}$$

est de classe $\mathcal{C}^2$ sur $\mathbb{R}_+\times\mathbb{R}$, et vérifie l'EDP :

$$\widetilde{u}_{tt}-c^2\widetilde{u}_{xx}=0\,,\quad (t,x)\in\,]0,+\infty[\,\times\mathbb{R}\,,$$

$$\widetilde{u}(0,x)=\widetilde{f}(x)\,,\quad \widetilde{u}_t(0,x)=\widetilde{g}(x)\,,\quad x\in\mathbb{R}\,.$$

La formule de d'Alembert (84) montre alors que, pour tout $(t,x)\in\mathbb{R}_+\times\mathbb{R}$:

$$\widetilde{u}(t,x)=\frac{1}{2}\left[\widetilde{f}(x+ct)+\widetilde{f}(x-ct)\right]+\frac{1}{2c}\int_{x-ct}^{x+ct}\widetilde{g}(s)\,\mathrm{d}s\,.$$

Pour tout $(t,x)\in\mathbb{R}_+\times\mathbb{R}_+$, cette formule s'écrit aussi :

$$u(t,x)=\begin{cases} \dfrac{1}{2}\left[f(x+ct)+f(x-ct)\right]+\dfrac{1}{2c}\displaystyle\int_{x-ct}^{x+ct}g(s)\,\mathrm{d}s & \text{si}\quad x\geqslant ct\geqslant 0\,,\\ \dfrac{1}{2}\left[f(x+ct)-f(ct-x)\right]+\dfrac{1}{2c}\displaystyle\int_{ct-x}^{x+ct}g(s)\,\mathrm{d}s & \text{si}\quad ct\geqslant x\geqslant 0\,.\end{cases} \tag{95}$$

Réciproquement, la fonction u définie par (95) satisfait le problème (92-94). On a donc montré l'existence et l'unicité de la solution du problème de Dirichlet, et obtenu aussi une

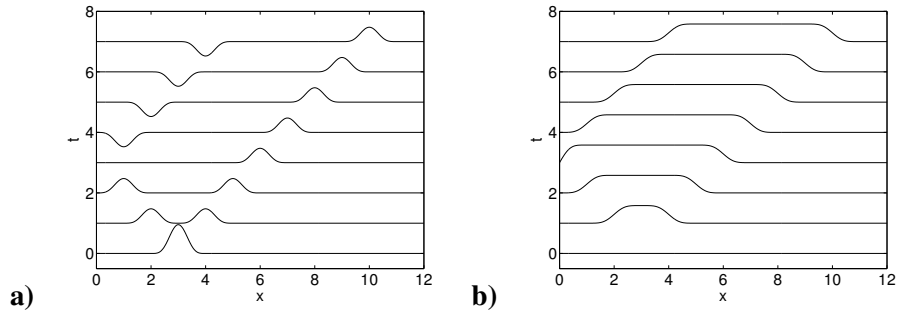


FIGURE III.2.4 – Propagation d'une onde uni-dimensionnelle avec paroi réfléchissante en $x = 0$. On trace les profils spatiaux de la solution aux instants $t = 0, 1, 2, \ldots, 7$, en prenant pour conditions initiales à $t = 0$: $f(x) = f_0(x-3)$, $g(x) = 0$ (figure a) et $f(x) = 0$, $g(x) = f_0(x-3)$ (figure b). On a pris ici $c = 1$ et la fonction $f_0(x) = \cos^4(\pi x/2)\mathbf{1}_{[-1,1]}(x)$ est à support compact et de classe $\mathcal{C}^2$ sur $\mathbb{R}$.

représentation explicite de la solution. Si $g = 0$, alors on peut facilement donner une interprétation de la solution. On trouve que la condition initiale f se divise en deux parties, l'une se propageant vers la droite avec la vitesse c, l'autre se propageant vers la gauche avec la vitesse $-c$. Cette seconde partie est réfléchie par la condition au bord $x = 0$, et elle donne lieu finalement à une onde allant vers la droite à la vitesse c (voir la figure III.2.4a).

Remarque. Si l'on ne suppose pas que $f''(0) = 0$, alors les formules précédentes sont encore valables partout sauf aux points (t, x) où $x = ct$. En effet, la fonction u définie par (95) vérifie les conditions aux bords (93-94) et elle vérifie l'équation des ondes (92) sur le domaine $]0, +\infty[\times]0, +\infty[$ privé de cette droite. Sur cette droite, la fonction u possède une singularité dans le sens où la dérivée seconde (en x et en t) présente un saut. En fait, cette singularité est issue du point $(t = 0, x = 0)$, qui est un « coin » du bord du domaine, et elle se propage le long de la droite $x = ct$, qui est une courbe caractéristique. Comme on l'a déjà noté page 400, c'est un phénomène habituel pour les EDP hyperboliques. On dira juste que l'on peut donner un sens rigoureux à l'EDP et à la solution u dans ce cadre, même si l'on n'étudiera pas ce genre de situations dans ce module.

4.2 Méthode de séparation de variables

Cette méthode consiste à rechercher des solutions particulières d'une EDP sous une forme séparable. En utilisant la linéarité de l'EDP, on peut ainsi obtenir une famille assez grande de solutions par combinaison linéaire de ces solutions séparables. Plus précisément, dans le cas d'une EDP linéaire en dimension 2 de la forme :

$$a(x,y)u_{xx} + 2b(x,y)u_{xy} + c(x,y)u_{yy} = d(x,y)u_x + e(x,y)u_y + r(x,y)u\,,$$

on recherche les solutions de l'EDP qui ont la forme $u(x,y) = X(x)Y(y)$, où X est une fonction de x seulement et Y est une fonction de y. Dans de nombreux cas, l'EDP

sur u peut alors être réduite à deux équations différentielles ordinaires sur les fonctions X et Y. Bien qu'il n'y ait en général pas de garantie que l'EDP admette des solutions séparables, la méthode est souvent utilisée pour essayer de trouver de telles solutions particulières. Dans les cas favorables, on obtient une famille de solutions de la forme $u_n(x,y) = X_n(x)Y_n(y)$, $n \geqslant 1$, et la linéarité de l'EDP nous assure que toute combinaison linéaire est encore solution. On peut ainsi construire une grande famille de solutions. Si l'on arrive à déterminer une combinaison linéaire des fonctions u_n qui vérifie les conditions initiales et les conditions aux bords du problème auquel on a affaire, alors on aura obtenu une solution du problème. On va appliquer la méthode de séparation de variables dans les paragraphes suivants sur des EDP linéaires du second ordre. Cette méthode peut se généraliser à des EDP en dimension plus grande que 2.

4.3 L'équation des cordes

On considère l'équation des ondes uni-dimensionnelle dans le domaine $]0,+\infty[\,\times\,]0,L[$, où L est un nombre réel strictement positif donné :

$$u_{tt} - c^2 u_{xx} = 0, \qquad (t,x) \in \,]0,+\infty[\,\times\,]0,L[\,. \tag{96}$$

On va examiner le problème de Cauchy associé à ce problème, pour lequel les conditions initiales (à $t=0$) et aux bords (en $x=0$ et $x=L$) sont spécifiées :

$$u(0,x) = f(x)\,, \quad u_t(0,x) = g(x)\,, \quad x \in [0,L]\,, \tag{97}$$

$$u(t,0) = 0\,, \quad u(t,L) = 0\,, \quad t \in \mathbb{R}_+\,, \tag{98}$$

où f et g sont des fonctions de classe $\mathcal{C}^2$ et de classe $\mathcal{C}^1$ sur $[0,L]$, respectivement.
Ce problème est appelé équation des cordes, car il modélise la vibration d'une corde de longueur L, tendue entre deux points ($u(t,x)$ est le déplacement transverse de la corde au point x à l'instant t). Dans ce modèle, $c^2 = T/m$, où T est une constante de tension (plus la corde est tendue, plus T est grand), et m est la densité linéaire de la corde (la masse par unité de longueur, supposée constante dans le cas d'une corde homogène).
Pour que les conditions (97) et (98) soient compatibles entre elles, il faut aussi réclamer :

$$f(0) = f(L) = 0\,, \tag{99}$$

(si une solution u existe, alors $f(0) = u(0,0) = 0$ et $f(L) = u(0,L) = 0$), et :

$$g(0) = g(L) = 0\,, \tag{100}$$

(si une solution u existe, alors $g(0) = u_t(0,0) = 0$ et $g(L) = u_t(0,L) = 0$). Si ces conditions supplémentaires ne sont pas satisfaites, alors on peut tout de suite dire que le problème n'a pas de solution dans $\mathcal{C}^2$.
On va aussi supposer les deux conditions suivantes :

$$f''(0) = f''(L) = 0\,, \qquad g'(0) = g'(L) = 0\,. \tag{101}$$

Ces deux conditions vont nous assurer que la solution est bien régulière sur le domaine tout entier. On pourrait déterminer une solution de l'EDP sans la condition (101), mais on observerait alors dans la solution des singularités dans la dérivée seconde qui se propagent le long des caractéristiques $x = kL \pm ct$, $k \in \mathbb{Z}$. On verra cela dans la démonstration.

On applique la méthode de séparation de variables pour rechercher des solutions de l'EDP (96) avec les conditions (97-98). Supposons donc qu'il existe une solution de classe $\mathcal{C}^2$ de la forme :

$$u(t,x) = T(t)X(x)\,. \tag{102}$$

En substituant cette forme dans l'EDP (96), on obtient :

$$X(x)\frac{\mathrm{d}^2T}{\mathrm{d}t^2}(t) - c^2\frac{\mathrm{d}^2X}{\mathrm{d}x^2}(x)T(t) = 0\,.$$

En divisant formellement par $c^2u(t,x)$:

$$\frac{1}{c^2T(t)}\frac{\mathrm{d}^2T}{\mathrm{d}t^2}(t) = \frac{1}{X(x)}\frac{\mathrm{d}^2X}{\mathrm{d}x^2}(x)\,,$$

on obtient une équation qui, dans le membre de droite, est une fonction de x seulement, et, dans le membre de gauche, est une fonction de t seulement. Ce ne peut donc être qu'une constante indépendante de x et t. Il existe donc $\lambda \in \mathbb{R}$ tel que :

$$\frac{\mathrm{d}^2T}{\mathrm{d}t^2}(t) = \lambda c^2T(t)\,, \tag{103}$$

$$\frac{\mathrm{d}^2X}{\mathrm{d}x^2}(x) = \lambda X(x)\,. \tag{104}$$

On résout ces équations différentielles ordinaires avec les conditions aux bords :

$$u(t,0) = 0 \Longrightarrow X(0)T(t) = 0\,, \quad \forall t \in \mathbb{R}_+\,,$$
$$u(t,L) = 0 \Longrightarrow X(L)T(t) = 0\,, \quad \forall t \in \mathbb{R}_+\,.$$

Si l'on avait $X(0) \neq 0$ ou $X(L) \neq 0$, alors on devrait avoir $T(t) = 0\ \forall t \in \mathbb{R}_+$, et donc $u(t,x) = 0\ \forall (t,x) \in \mathbb{R}_+ \times [0,L]$. Or on recherche des solutions non triviales (non identiquement nulles), donc on doit avoir :

$$X(0) = X(L) = 0\,. \tag{105}$$

Les solutions possibles sont de trois types, selon que $\lambda < 0$, $\lambda = 0$, $\lambda > 0$.
Si $\lambda > 0$, alors X est de la forme :

$$X(x) = Ae^{\sqrt{\lambda}x} + Be^{-\sqrt{\lambda}x}\,,$$

où A et B sont deux constantes d'intégration. En tenant compte des conditions aux bords, A et B doivent vérifier le système linéaire 2×2 :

$$\begin{cases} A + B = 0\,, \\ Ae^{\sqrt{\lambda}L} + Be^{-\sqrt{\lambda}L} = 0\,. \end{cases}$$

Le déterminant de ce système est $e^{-\sqrt{\lambda}L} - e^{\sqrt{\lambda}L} \neq 0$, donc ce système admet une unique solution, qui est $(A,B) = (0,0)$. Donc on obtient la solution triviale $X = 0$, soit $u = 0$.
Si $\lambda = 0$, alors X est de la forme :

$$X(x) = Ax + B\,,$$

où A et B sont deux constantes d'intégration. En tenant compte des conditions aux bords, A et B doivent vérifier le système linéaire 2×2 :

$$\begin{cases} B = 0\,, \\ AL + B = 0\,, \end{cases}$$

dont l'unique solution est $(A, B) = (0, 0)$. Donc $X = 0$ et $u = 0$.
Si $\lambda < 0$, alors X est de la forme :

$$X(x) = A\cos(\alpha x) + B\sin(\alpha x)\,,$$

où $\alpha = \sqrt{-\lambda}$, A et B étant deux constantes d'intégration. En tenant compte des conditions aux bords, A et B doivent vérifier le système linéaire 2×2 :

$$\begin{cases} A = 0\,, \\ A\cos(\alpha L) + B\sin(\alpha L) = 0\,. \end{cases}$$

On obtient donc une famille de solutions non triviales $X(x) = B\sin(\alpha x)$ si, et seulement si, α vérifie $\sin(\alpha L) = 0$. Les solutions strictement positives de cette équation sont $\alpha_n = n\pi/L$, $n \in \mathbb{N}^*$. Il y a donc une infinité (dénombrable) de valeurs $(\lambda_n)_{n\geqslant 1}$:

$$\lambda_n = -\frac{n^2\pi^2}{L^2}\,,$$

pour lesquelles l'équation différentielle ordinaire (104) avec les conditions aux bords (105) a une solution non triviale :

$$X_n(x) = B_n \sin\left(\frac{n\pi x}{L}\right)\,.$$

Pour ces valeurs de λ_n, on peut résoudre l'équation différentielle ordinaire (103) :

$$T_n(t) = C_n \cos\left(\frac{\pi n c}{L}t\right) + D_n \sin\left(\frac{\pi n c}{L}t\right)\,.$$

Pour chaque $n \in \mathbb{N}^*$, il existe donc une famille de solutions séparables de l'EDP (96) :

$$u_n(t,x) = \left[a_n \cos\left(\frac{\pi n c}{L}t\right) + b_n \sin\left(\frac{\pi n c}{L}t\right)\right] \sin\left(\frac{n\pi x}{L}\right)\,, \tag{106}$$

où a_n et b_n sont arbitraires. En fait, certaines des opérations que l'on a faites, par exemple diviser par $u(t,x)$, étaient formelles. Cependant, on peut vérifier directement en calculant les dérivées secondes que la fonction définie dans (106) est bien une solution de (96). En utilisant la linéarité de l'EDP, on obtient que toute combinaison linéaire de ces solutions est encore solution, et plus généralement, toute somme infinie de la forme :

$$u(t,x) = \sum_{n=1}^{+\infty} \left[a_n \cos\left(\frac{\pi n c}{L}t\right) + b_n \sin\left(\frac{\pi n c}{L}t\right)\right] \sin\left(\frac{n\pi x}{L}\right) \tag{107}$$

est encore solution, sous réserve que la série converge, que la fonction soit de classe $\mathcal{C}^2$ et que l'on puisse échanger les signes somme et dérivée.
Avant d'énoncer le résultat, nous faisons un aparté sur la décomposition en série de Fourier de la solution. La condition initiale $x \mapsto f(x)$ et la solution cherchée $x \mapsto u(t,x)$ sont des

fonctions de classe $\mathcal{C}^2([0,L],\mathbb{R})$ s'annulant aux bords 0 et L. De telles fonctions admettent des décompositions en séries de Fourier particulières.

Lemme 14.
Soit $x \mapsto v(x)$ une fonction de classe $\mathcal{C}^2$ sur $[0,L]$, vérifiant $v(0) = v(L) = 0$. Alors, pour tout $x \in [0,L]$:

$$v(x) = \sum_{n=1}^{+\infty} a_n \sin\left(\frac{n\pi x}{L}\right), \tag{108}$$

où a_n, $n \geqslant 1$, est donné par :

$$a_n = \frac{2}{L}\int_0^L v(x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,. \tag{109}$$

Démonstration. Définissons la fonction $\widetilde{v}$:

$$\widetilde{v}(x) := \begin{cases} v(x) \text{ si } x \in [0,L]\,, \\ -v(-x) \text{ si } x \in [-L,0]\,. \end{cases} \tag{110}$$

La fonction $\widetilde{v}$ peut être prolongée en une fonction périodique de période $2L$, impaire, de classe $\mathcal{C}^1$ sur $\mathbb{R}$, de classe $\mathcal{C}^2$ sur $\bigcup_{k\in\mathbb{Z}}]kL,(k+1)L[$, et admettant des dérivées secondes à droite et à gauche aux points kL, $k \in \mathbb{Z}$. Ses coefficients de Fourier sont :

$$a_n = \frac{1}{L}\int_{-L}^L \widetilde{v}(x)\sin\left(\frac{2n\pi x}{2L}\right)\mathrm{d}x = \frac{2}{L}\int_0^L v(x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,,$$

$$b_n = \frac{1}{L}\int_{-L}^L \widetilde{v}(x)\cos\left(\frac{2n\pi x}{2L}\right)\mathrm{d}x = 0\,,$$

où l'on a utilisé le fait que $\widetilde{v}$ est impaire. D'après le théorème de Dirichlet (*cf.* le livre [L2]), cette fonction est égale à la somme de sa série de Fourier (108) en tout point x. Si l'on suppose de plus les conditions $v''(0) = v''(L) = 0$, alors $\widetilde{v}$ est de classe $\mathcal{C}^2$ sur $\mathbb{R}$. ■

Proposition 15. Soient $f \in \mathcal{C}^2([0,L],\mathbb{R})$ et $g \in \mathcal{C}^1([0,L],\mathbb{R})$ vérifiant les trois conditions de compatibilité (99-101). Il existe une unique solution u de classe $\mathcal{C}^2$ sur $\mathbb{R}_+ \times [0,L]$ au problème de Cauchy (96) avec les conditions (97-98). Cette solution s'écrit sous la forme d'une série de Fourier (107) avec :

$$a_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,, \tag{111}$$

$$b_n = \frac{2}{n\pi c}\int_0^L g(x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,. \tag{112}$$

Démonstration. Considérons u une solution de classe $\mathcal{C}^2$ et montrons que cette solution est nécessairement de la forme indiquée. Pour $n \geqslant 1$, on introduit le coefficient de Fourier :

$$A_n(t) = \frac{2}{L}\int_0^L u(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,.$$

La fonction A_n est de classe $\mathcal{C}^2$ sur $\mathbb{R}_+$, et sa dérivée temporelle seconde est (en utilisant le théorème de dérivation sous le signe somme) :

$$\frac{\mathrm{d}^2 A_n}{\mathrm{d}t^2}(t) = \frac{2}{L}\int_0^L u_{tt}(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,.$$

On utilise l'EDP (96) qui donne pour tout $t > 0$:

$$\frac{\mathrm{d}^2 A_n}{\mathrm{d}t^2}(t) = \frac{2c^2}{L}\int_0^L u_{xx}(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,.$$

On intègre une fois par parties :

$$\begin{aligned}\frac{\mathrm{d}^2 A_n}{\mathrm{d}t^2}(t) &= \frac{2c^2}{L}\left[u_x(t,x)\sin\left(\frac{n\pi x}{L}\right)\right]_0^L - \frac{2c^2 n\pi}{L^2}\int_0^L u_x(t,x)\cos\left(\frac{n\pi x}{L}\right)\mathrm{d}x\\ &= -\frac{2c^2 n\pi}{L^2}\int_0^L u_x(t,x)\cos\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,,\end{aligned}$$

puis une seconde fois :

$$\begin{aligned}\frac{\mathrm{d}^2 A_n}{\mathrm{d}t^2}(t) &= -\frac{2c^2 n\pi}{L^2}\left[u(t,x)\cos\left(\frac{n\pi x}{L}\right)\right]_0^L - \frac{2c^2 n^2\pi^2}{L^3}\int_0^L u(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\\ &= -\frac{2c^2 n^2\pi^2}{L^3}\int_0^L u(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x\\ &= -\frac{c^2 n^2\pi^2}{L^2}A_n(t)\,.\end{aligned}$$

Pour passer de la première à la deuxième ligne, on s'est servi du fait que u satisfait les conditions aux bords (97). On obtient ainsi une équation différentielle ordinaire vérifiée par A_n, dont la solution générale est :

$$A_n(t) = a_n\cos\left(\frac{n\pi c}{L}t\right) + b_n\sin\left(\frac{n\pi c}{L}t\right)\,. \tag{113}$$

On peut identifier les coefficients a_n et b_n en utilisant les conditions initiales (97) :

$$\begin{aligned}f(x) = u(0,x) &\Longrightarrow \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x = A_n(0) = a_n\,,\\ g(x) = u_t(0,x) &\Longrightarrow \frac{2}{L}\int_0^L g(x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x = A_n'(0) = \frac{n\pi c}{L}b_n\,,\end{aligned}$$

ce qui donne les expressions (111-112). En utilisant le lemme 14, on obtient que u est donnée par :

$$u(t,x) = \sum_{n=1}^{+\infty} A_n(t)\sin\left(\frac{\pi n x}{L}\right)\,. \tag{114}$$

Réciproquement, on peut vérifier directement que la fonction u définie par (107) avec a_n et b_n donnés par (111-112) est solution de l'EDP (96) avec les conditions (97-98). On

écrit $u(t,x) = v(t,x) + w(t,x)$ avec :

$$v(t,x) := \sum_{n=1}^{+\infty} a_n \cos\left(\frac{\pi n c}{L} t\right) \sin\left(\frac{n\pi x}{L}\right) ,$$

$$w(t,x) := \sum_{n=1}^{+\infty} b_n \sin\left(\frac{\pi n c}{L} t\right) \sin\left(\frac{n\pi x}{L}\right) ,$$

et l'on va traiter séparément les fonctions v et w.

La fonction $\widetilde{f}$ définie sur $[-L, L]$ par :

$$\widetilde{f}(x) = \begin{cases} f(x) & \text{si } x \in [0, L] , \\ -f(-x) & \text{si } x \in [-L, 0] , \end{cases}$$

est prolongeable sur $\mathbb{R}$ en une fonction impaire, périodique de période $2L$ et de classe $\mathcal{C}^2$ (voir la preuve du lemme 14). Cette fonction est égale à la somme de sa série de Fourier d'après le théorème de Dirichlet (*cf.* le livre [L2]). Or, en utilisant des relations de trigonométrie usuelles, on trouve que la fonction $v(t,x)$ peut s'écrire sous la forme :

$$\begin{aligned} v(t,x) &= \sum_{n=1}^{+\infty} \frac{a_n}{2} \left[\sin\left(\frac{n\pi}{L}(x+ct)\right) + \sin\left(\frac{n\pi}{L}(x-ct)\right)\right] \\ &= \frac{1}{2}\left[\widetilde{f}(x+ct) + \widetilde{f}(x-ct)\right] . \end{aligned}$$

Comme $\widetilde{f}$ est de classe $\mathcal{C}^2$, la fonction v est de classe $\mathcal{C}^2$ et vérifie l'équation des ondes. Elle vérifie aussi $v(0,x) = f(x)$ et $v_t(0,x) = 0$.

Noter que, si f ne vérifie pas la condition (101), alors tous les arguments précédents sont encore valables, mais la fonction $\widetilde{f}$ a des dérivées secondes à droite et à gauche en kL différentes. La fonction v vérifie alors l'équation des ondes dans le domaine $(t,x) \in]0, +\infty[\times]0, L[$ sauf le long des lignes caractéristiques $x \pm ct = kL$, $k \in \mathbb{Z}$. Cela illustre le phénomène de propagation des singularités mentionné page 400.

La fonction G définie sur $[-L, L]$ par :

$$G(x) = \begin{cases} \displaystyle\int_0^x g(y)\,\mathrm{d}y & \text{si } x \in [0, L] , \\ \displaystyle\int_0^{-x} g(y)\,\mathrm{d}y & \text{si } x \in [-L, 0] , \end{cases}$$

est prolongeable sur $\mathbb{R}$ en une fonction périodique de période $2L$. Cette fonction est paire. Elle est de classe $\mathcal{C}^2$ sur $]kL, (k+1)L[$, car g est de classe $\mathcal{C}^1$. De plus $G(0^+) = G(0^-)$, $G'(0^+) = g(0) = 0 = -g(0) = G'(0^-)$, $G''(0^+) = g'(0) = 0 = G''(0^-)$, ce qui montre que G est deux fois continûment différentiable en 0. On montre de même que G est deux fois continûment différentiable en kL, $k \in \mathbb{Z}$. Donc la fonction G est de classe $\mathcal{C}^2$. Ses

coefficients de Fourier sont :

$$\frac{1}{L}\int_{-L}^{L} G(x)\sin\left(\frac{2n\pi x}{2L}\right)\mathrm{d}x = 0\,,$$
$$\frac{1}{L}\int_{-L}^{L} G(x)\cos\left(\frac{2n\pi x}{2L}\right)\mathrm{d}x = \frac{2}{L}\int_{0}^{L} G(x)\cos\left(\frac{n\pi x}{L}\right)\mathrm{d}x\,,$$

où l'on a utilisé le fait que G est paire. Le coefficient b_n peut s'écrire :

$$\begin{aligned} b_n &= \frac{2}{n\pi c}\int_0^L g(x)\sin\left(\frac{\pi n x}{L}\right)\mathrm{d}x \\ &= \frac{2}{n\pi c}\left[G(x)\sin\left(\frac{\pi n x}{L}\right)\right]_0^L - \frac{2}{cL}\int_0^L G(x)\cos\left(\frac{\pi n x}{L}\right)\mathrm{d}x \\ &= -\frac{1}{c}\frac{2}{L}\int_0^L G(x)\cos\left(\frac{\pi n x}{L}\right)\mathrm{d}x\,, \end{aligned}$$

où l'on a utilisé le fait que $\sin(0) = \sin(n\pi) = 0$. Cela montre que b_n est, au facteur $-1/c$ près, le coefficient de Fourier de la fonction G périodique de classe $\mathcal{C}^2$. D'après le théorème de Dirichlet, pour tout $x \in \mathbb{R}$:

$$\sum_{n=1}^{+\infty} b_n \cos\left(\frac{n\pi x}{L}\right) = -\frac{1}{c}G(x)\,.$$

Or la fonction $w(t,x)$ peut s'écrire sous la forme :

$$\begin{aligned} w(t,x) &= \sum_{n=1}^{+\infty}\frac{b_n}{2}\left[-\cos\left(\frac{n\pi}{L}(x+ct)\right) + \cos\left(\frac{n\pi}{L}(x-ct)\right)\right] \\ &= \frac{1}{2c}\left[G(x+ct) - G(x-ct)\right]\,, \end{aligned}$$

ce qui montre que la fonction w satisfait l'équation des ondes. Elle vérifie aussi $w(0,x) = 0$ et $w_t(0,x) = G'(x) = g(x)$.

Comme u est la somme de v et w, la fonction u est de classe $\mathcal{C}^2$ et vérifie (96-98). ■

L'examen de la solution (106) permet de tirer des informations qui peuvent être d'importance physique. Ainsi, on peut vérifier que la solution est une superposition de modes de vibration, dont le $n^{ème}$ est de la forme (106). Le $n^{ème}$ mode, encore appelé $n^{ème}$ harmonique, oscille temporellement avec la période $T_n = 2L/(nc)$ qui correspond à la fréquence $1/T_n = n\nu_1$, où $\nu_1 = c/(2L)$ est la fréquence fondamentale. Toutes les fréquences de vibration sont des multiples de cette fréquence.

Dans le cas de l'équation des cordes, il est fréquent que la condition initiale corresponde à un déplacement imposé avec une vitesse nulle (on tire sur la corde, puis on lâche). Dans ce cas la condition initiale g est nulle, et les coefficients b_n sont nuls. Si l'on suppose que l'on excite un mode pur, alors on trouve la solution :

$$u_n(t,x) = a_n \cos\left(\frac{n\pi c}{L}t\right)\sin\left(\frac{n\pi x}{L}\right)\,, \tag{115}$$

qui est une onde stationnaire dans le sens où son amplitude $a_n \sin\left(\frac{n\pi x}{L}\right)$ oscille dans son ensemble. La solution possède $n+1$ nœuds (i.e. zéros) situés en $x_k = kL/n$, $k \in [\![0, n]\!]$. Elle possède aussi n ventres (i.e. extrema) aux points $x_l = (l - 1/2)L/n$, $l \in [\![1, n]\!]$.
Pour finir, on peut remarquer que la solution (106) peut s'écrire sous une forme de solution de d'Alembert. En effet, en faisant appel à des formules de trigonométrie élémentaire, la solution (115) s'écrit simplement :

$$u_n(t,x) = \frac{1}{2}a_n \sin\left(\frac{n\pi}{L}(x - ct)\right) + \frac{1}{2}a_n \sin\left(\frac{n\pi}{L}(x + ct)\right) .$$

Exercice 5.
On suppose les mêmes conditions que dans la proposition 15 de la page 418.

1. Montrer que l'énergie de l'onde définie par :

$$E(t) := \frac{1}{2}\int_0^L u_x(t,x)^2 + c^{-2}u_t(t,x)^2 \,\mathrm{d}x$$

   est constante et égale à E_0.

2. Montrer que :

$$E_0 = \frac{1}{2}\int_0^L f'(x)^2 + c^{-2}g(x)^2 \,\mathrm{d}x .$$

Solution. 1. En utilisant le théorème de dérivation sous le signe somme :

$$\frac{\mathrm{d}E}{\mathrm{d}t}(t) = \int_0^L u_x u_{xt}(t,x) + c^{-2}u_t u_{tt}(t,x) \,\mathrm{d}x .$$

On utilise l'EDP (96) qui montre que $u_{tt} = c^2 u_{xx}$ pour tout $t > 0$:

$$\frac{\mathrm{d}E}{\mathrm{d}t}(t) = \int_0^L u_x u_{xt}(t,x) + u_t u_{xx}(t,x) \,\mathrm{d}x .$$

On intègre par parties le second terme :

$$\frac{\mathrm{d}E}{\mathrm{d}t}(t) = \int_0^L u_x u_{xt}(t,x) \,\mathrm{d}x + [u_t u_x(t,x)]_0^L - \int_0^L u_{tx} u_x(t,x) \,\mathrm{d}x = 0 ,$$

où l'on s'est servi du fait que $u_t(t,0) = u_t(t,L) = 0$ pour tout t grâce à la condition aux bords (98).
2. Il suffit de voir que $u_x(0,x) = f'(x)$ et $u_t(0,x) = g(x)$ $\forall x \in [0,L]$.

4.4 L'équation de la chaleur uni-dimensionnelle

On considère l'équation de la chaleur uni-dimensionnelle dans le domaine $]0,+\infty[\times]0,L[$:

$$u_t = \kappa u_{xx} , \quad \forall (t,x) \in]0,+\infty[\times]0,L[, \tag{116}$$

où L et κ sont des nombres réels strictement positifs. On va examiner le problème de Cauchy associé à cette EDP, pour lequel les conditions initiales (à $t = 0$) et aux bords

($x = 0$ et $x = L$) sont spécifiées :

$$u(0,x) = f(x)\,, \quad x \in [0,L]\,, \tag{117}$$

$$u(t,0) = 0\,, \quad u(t,L) = 0, \quad t \in \mathbb{R}_+\,, \tag{118}$$

où f est une fonction de classe $\mathcal{C}^2$.
Ce problème modélise par exemple l'évolution de la distribution de température dans un cable de longueur L ($u(t,x)$ est la valeur de la température au point x à l'instant t). La constante κ est appelée coefficient de conductivité (plus κ est grand, et plus vite la chaleur se diffuse, comme on va le voir).
La fonction f doit vérifier la condition :

$$f(0) = f(L) = 0\,, \tag{119}$$

pour que l'on ait une chance de trouver une solution. En effet, si u est solution, alors on doit avoir $f(0) = u(0,0) = 0$ et $f(L) = u(0,L) = 0$.
Appliquons la méthode de séparation de variables pour rechercher des solutions de l'EDP (116). Supposons qu'il existe une solution de classe $\mathcal{C}^2$ de la forme :

$$u(t,x) = T(t)X(x)\,. \tag{120}$$

En substituant cette forme dans l'EDP (116), on obtient :

$$X(x)\frac{\mathrm{d}T}{\mathrm{d}t}(t) = \kappa\frac{\mathrm{d}^2X}{\mathrm{d}x^2}(x)T(t)\,.$$

En divisant formellement par $\kappa u(t,x)$:

$$\frac{1}{\kappa T(t)}\frac{\mathrm{d}T}{\mathrm{d}t}(t) = \frac{1}{X(x)}\frac{\mathrm{d}^2X}{\mathrm{d}x^2}(x)\,,$$

on obtient une équation qui, dans le membre de droite, est une fonction de x seulement, et, dans le membre de gauche, est une fonction de t seulement. Ce ne peut donc être qu'une constante indépendante de x et t. Il existe donc $\lambda \in \mathbb{R}$ tel que :

$$\frac{\mathrm{d}T}{\mathrm{d}t}(t) = \lambda\kappa T(t)\,, \tag{121}$$

$$\frac{\mathrm{d}^2X}{\mathrm{d}x^2}(x) = \lambda X(x)\,. \tag{122}$$

On résout ces équations différentielles ordinaires avec les conditions aux bords :

$$u(t,0) = 0 \Longrightarrow X(0)T(t) = 0\,, \quad \forall t \in \mathbb{R}_+\,,$$

$$u(t,L) = 0 \Longrightarrow X(L)T(t) = 0\,, \quad \forall t \in \mathbb{R}_+\,.$$

On recherche des solutions non triviales (non identiquement nulles), donc on doit avoir :

$$X(0) = X(L) = 0\,. \tag{123}$$

En suivant exactement le raisonnement fait dans le cas de l'équation des cordes, page 416, on montre qu'il n'y a pas de solution non identiquement nulle de forme (120) si $\lambda \geqslant 0$.

Si $\lambda < 0$, alors $\lambda = -\alpha^2$, $\alpha > 0$ et il existe une famille de solutions non triviales $X(x) = B\sin(\alpha x)$, si, et seulement si, α vérifie $\sin(\alpha L) = 0$. Les solutions strictement positives de cette équation sont $\alpha_n = n\pi/L$, $n \in \mathbb{N}^*$. Il y a donc une infinité (dénombrable) de valeurs $(\lambda_n)_{n\geqslant 1}$, avec $\lambda_n = -n^2\pi^2/L^2$, pour lesquelles l'équation différentielle ordinaire (122) avec les conditions aux bords (123) a une solution non triviale :

$$X_n(x) = B_n \sin\left(\frac{n\pi x}{L}\right) .$$

Pour ces valeurs de λ_n, on peut résoudre l'équation différentielle ordinaire (121) :

$$T_n(t) = C_n \exp\left(-\frac{\kappa\pi^2 n^2}{L^2}t\right) .$$

Pour chaque $n \in \mathbb{N}^*$, il existe donc une famille de solutions séparables de l'EDP (116) :

$$u_n(t,x) = a_n \exp\left(-\frac{\kappa\pi^2 n^2}{L^2}t\right)\sin\left(\frac{\pi n x}{L}\right) , \tag{124}$$

où $a_n \in \mathbb{R}$ est arbitraire. Par linéarité de l'EDP, on obtient que toute combinaison linéaire de ces solutions est encore solution, et plus généralement, toute somme infinie de la forme :

$$u(t,x) = \sum_{n=1}^{+\infty} a_n \exp\left(-\frac{\kappa\pi^2 n^2}{L^2}t\right)\sin\left(\frac{\pi n x}{L}\right) \tag{125}$$

est encore solution, sous réserve que la série converge, que la fonction soit de classe $\mathcal{C}^2$ et que l'on puisse échanger les signes somme et dérivée.

Proposition 16. Soit $f \in \mathcal{C}^2([0,L],\mathbb{R})$ vérifiant la condition de compatibilité (119). Il existe une unique fonction u telle que u, u_t, u_x, u_{xx} sont continues sur $\mathbb{R}_+ \times [0,L]$ et u est solution du problème de Cauchy (116) avec les conditions (117-118). Cette solution s'écrit sous la forme d'une série de Fourier (125) avec :

$$a_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x . \tag{126}$$

De plus, cette solution est de classe $\mathcal{C}^\infty$ dans $]0,+\infty[\times[0,L]$.

Démonstration.

Considérons u une solution telle que u, u_t, u_x, u_{xx} sont continues sur $\mathbb{R}_+ \times [0,L]$. Pour $n \geqslant 1$, on introduit le coefficient de Fourier :

$$A_n(t) = \frac{2}{L}\int_0^L u(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x .$$

Sa dérivée est :

$$\frac{\mathrm{d}A_n}{\mathrm{d}t}(t) = \frac{2}{L}\int_0^L u_t(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x = \frac{2\kappa}{L}\int_0^L u_{xx}(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x .$$

En intégrant deux fois par parties :

$$\frac{\mathrm{d}A_n}{\mathrm{d}t}(t) = -\frac{2n^2\pi^2\kappa}{L^3}\int_0^L u(t,x)\sin\left(\frac{n\pi x}{L}\right)\mathrm{d}x = -\frac{\kappa n^2\pi^2}{L^2}A_n(t) .$$

En appliquant le lemme 14 de la page 418 on trouve que la fonction $x \mapsto u(t,x)$ est égale à la somme :

$$u(t,x) = \sum_{n=1}^{+\infty} A_n(t) \sin\left(\frac{n\pi x}{L}\right) ,$$

en tout point $x \in [0,L]$. Cela montre que $u(t,x)$ est bien donnée par (125).

Réciproquement, la fonction u définie par (125) est bien solution. En effet, les coefficients a_n sont bornés :

$$\sup_{n\in\mathbb{N}^*} |a_n| \leqslant 2 \sup_{x\in[0,L]} |f(x)| ,$$

et donc les coefficients de Fourier de $u(t,x)$ décroissent exponentiellement en n dès que $t > 0$, ainsi que ceux des dérivées temporelles de $u(t,x)$. La fonction u est donc de classe $\mathcal{C}^\infty$, on peut dériver sous le signe somme et vérifier que u est bien solution. On peut aussi vérifier que u satisfait la conditions initiale (117) et les conditions aux bords (118). ■

Noter que l'équation de la chaleur est régularisante, contrairement à l'équation des ondes. En effet, la solution a la régularité de la fonction f à $t = 0$, mais elle devient de classe $\mathcal{C}^\infty$ sitôt que $t > 0$. Il s'agit en fait d'une propriété générale des EDP paraboliques.

Exercice type corrigé

III.2.0 Le but de l'exercice est de donner une preuve de l'unicité de la solution de l'équation des cordes par une méthode dite d'énergie. On considère l'EDP (96) avec les conditions aux bords (97-98). Soient u et v deux solutions de classe $\mathcal{C}^2$.

1. Montrer que $w := u - v$ est solution de l'EDP avec des conditions aux bords que l'on identifiera.
2. Montrer que l'énergie de w définie par :

$$E(t) := \frac{1}{2}\int_0^L w_x(t,x)^2 + \frac{1}{c^2} w_t(t,x)^2 \,\mathrm{d}x$$

   est constante en t : $E(t) = E(0)\ \forall t \geqslant 0$.
3. Calculer $E(0)$ et conclure.

Solution.
1. La fonction w est de classe $\mathcal{C}^2$ car c'est la somme de deux fonctions de classe $\mathcal{C}^2$. Par linéarité de l'EDP (96), la fonction w est solution de l'EDP. Elle vérifie de plus les conditions aux bords $w(0,x) = u(0,x) - v(0,x) = f(x) - f(x) = 0$ et $w_t(0,x) = u_t(0,x) - v_t(0,x) = g(x) - g(x) = 0$.
2. Nous alllons dériver $E(t)$ par rapport à t en utilisant le théorème de dérivation sous le signe somme (voir le théorème 124 de la page 306 de ce livre ou le corollaire 131 de la page 720 du livre [L2]). Nous devons tout d'abord vérifier les hypothèses du théorème. La fonction w est de classe $\mathcal{C}^2$ sur $\mathbb{R}_+ \times [0,L]$. Par conséquent, la fonction $\widetilde{f}(t,x) := w_x(t,x)^2 + c^{-2} w_t(t,x)^2$ est continue sur $\mathbb{R}_+ \times [0,L]$ et admet en tout point une dérivée partielle en t qui est elle-même continue

sur $\mathbb{R}_+ \times [0, L]$. Cela suffit pour assurer que $\widetilde{F}(t) := \displaystyle\int_0^L \widetilde{f}(t,x)\,\mathrm{d}x$ est de classe $\mathcal{C}^1$ sur $\mathbb{R}_+$ et :

$$\frac{\mathrm{d}\widetilde{F}}{\mathrm{d}t}(t) = \int_0^L \frac{\partial \widetilde{f}}{\partial t}(t,x)\,\mathrm{d}x\,,$$

ce qui s'écrit aussi :

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_0^L w_x(t,x)^2 + \frac{1}{c^2}w_t(t,x)^2\,\mathrm{d}x\right) &= \int_0^L \frac{\partial}{\partial t}\Big(w_x(t,x)^2 + \frac{1}{c^2}w_t(t,x)^2\Big)\,\mathrm{d}x \\ &= \int_0^L 2w_x w_{xt}(t,x) + \frac{2}{c^2}w_t w_{tt}(t,x)\,\mathrm{d}x\,,\end{aligned}$$

et l'on obtient :

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \int_0^L w_x w_{xt}(t,x) + \frac{1}{c^2}w_t w_{tt}(t,x)\,\mathrm{d}x = \int_0^L w_x w_{xt}(t,x) + w_t w_{xx}(t,x)\,\mathrm{d}x\,.$$

En intégrant par parties le second terme du membre de droite, on trouve :

$$\int_0^L w_t w_{xx}(t,x)\,\mathrm{d}x = \Big[w_t w_x(t,x)\Big]_0^L - \int_0^L w_{tx}w_x(t,x)\,\mathrm{d}x\,.$$

Comme $w(t,0) = 0$ et $w(t,L) = 0$ pour tout $t \geqslant 0$, on a aussi $w_t(t,0) = 0$ et $w_t(t,L) = 0$, ce qui montre que le crochet est nul. On trouve donc :

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \int_0^L w_x w_{xt}(t,x) - w_{tx}w_x(t,x)\,\mathrm{d}x = 0\,.$$

3. Comme $w(0,x) = 0$ et $w_t(0,x) = 0$ pour tout $x \in [0,L]$, l'énergie initiale de w est nulle : $E(0) = 0$.
Comme l'énergie est conservée, cela implique que l'énergie de w est nulle à tout instant t. Vu la définition de l'énergie, cela impose que $\displaystyle\int_0^L w_t^2\,\mathrm{d}x = 0$. Or w_t^2 est une fonction continue et positive, donc w_t est identiquement nulle. Comme $w(0,x) = 0$ et $w(t,x) = \displaystyle\int_0^t w_t(s,x)\,\mathrm{d}s$, on trouve que w est nulle. On peut donc conclure que $u = v$ et qu'il y a unicité de la solution de l'EDP (96) avec les conditions aux bords (97-98).

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`

III.2.1 On considère l'équation des ondes (81) avec les conditions initiales (83). Supposons que $f \in \mathcal{C}^2(\mathbb{R},\mathbb{R})$ et $g \in \mathcal{C}^1(\mathbb{R},\mathbb{R})$ soient à support compact. L'énergie cinétique de l'onde est $k(t) = \frac{1}{2}\displaystyle\int_{\mathbb{R}} u_x(t,x)^2\,\mathrm{d}x$ et son énergie potentielle est $p(t) = \frac{1}{2c^2}\displaystyle\int_{\mathbb{R}} u_t(t,x)^2\,\mathrm{d}x$.

1. Montrer que $p(t) + k(t)$ est constante.

2. Montrer qu'il existe $t_0 \in \mathbb{R}_+$ tel que $p(t) = k(t)$ pour tout $t \geqslant t_0$.

On a donc montré la propriété d'équipartition de l'énergie. Cette propriété est encore vraie asymptotiquement (i.e. pour $t \to +\infty$) lorsque les données initiales ne sont pas à support compact.

III.2.2 Soient $n \in \mathbb{R}\backslash\{0\}$ et $f \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$. Le but de l'exercice est de résoudre par la méthode des caractéristiques l'EDP avec conditions de Dirichlet :

$$\begin{aligned} &u_x - yu_y = y^n\,, \quad (x,y) \in \,]0,+\infty[\, \times \mathbb{R}\,, \\ &u(x=0,y) = f(y)\,, \quad y \in \mathbb{R}\,. \end{aligned}$$

1. Déterminer les équations caractéristiques de cette EDP.
2. Résoudre les équations caractéristiques et identifier la solution de l'EDP.

III.2.3 Résoudre par la méthode des caractéristiques l'EDP avec conditions de Dirichlet :

$$\begin{aligned} &u_x - yu_y = u\,, \quad (x,y) \in \,]0,+\infty[\, \times \mathbb{R}\,, \\ &u(x=0,y) = f(y)\,, \quad y \in \mathbb{R}\,, \end{aligned}$$

où $f \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$.

III.2.4 Soit $n \in \mathbb{R}\backslash\{1\}$. On considère l'EDP $x^{2n}u_{xx} - u_{yy} = 0$ dans $]0,+\infty[\, \times \mathbb{R}$.

1. Montrer qu'il s'agit d'une EDP hyperbolique.
2. Déterminer les courbes caractéristiques et réduire l'EDP à sa première forme canonique.
3. On suppose ici que $n = 2$. Montrer que la forme canonique devient $[(\xi - \eta)\widetilde{u}]_{\xi\eta} = 0$, et que la forme générale de la solution est :

$$\widetilde{u}(\xi,\eta) = \frac{f(\xi) - g(\eta)}{\xi - \eta}\,,$$

où f et g sont deux fonctions de classe $\mathcal{C}^2$.

III.2.5 On considère l'EDP $x^2u_{xx} + 2xyu_{xy} + y^2u_{yy} = 0$ dans $]0,+\infty[\, \times \mathbb{R}$.

1. Montrer qu'il s'agit d'une EDP parabolique.
2. Trouver les caractéristiques et réduire l'EDP à sa forme canonique.
3. En résolvant la forme canonique de l'EDP, montrer que la forme générale de la solution est $u(x,y) = yf(y/x) + g(y/x)$, où $f, g \in \mathcal{C}^2(\mathbb{R}, \mathbb{R})$.

III.2.6 On considère l'EDP $4u_{xx} + 5u_{xy} + u_{yy} = 2 - u_x - u_y$ dans $\mathbb{R}^2$.

1. Montrer qu'il s'agit d'une EDP hyperbolique.
2. Trouver les caractéristiques et réduire l'EDP à sa première forme canonique, dont on notera $\widetilde{u}(\xi,\eta)$ une solution.

3. En posant $v(\xi,\eta) = \widetilde{u}(\xi,\eta)\exp(m\xi + n\eta)$, montrer que l'on peut choisir $m, n \in \mathbb{R}$ de telle sorte que l'on se ramène à l'EDP :

$$v_{\xi\eta} = -\frac{8}{9}\exp(-\frac{\xi}{3})\,.$$

Résoudre cette équation.

4. Montrer que la forme générale de la solution est :

$$u(x,y) = \frac{8}{3}(y - \frac{x}{4}) + \exp(\frac{y-x}{3})\left[f(y-x) + g(y - \frac{x}{4})\right]\,,$$

où $f, g \in \mathcal{C}^2(\mathbb{R},\mathbb{R})$.

III.2.7 Soit $n \in \mathbb{R}\backslash\{1\}$. On considère l'EDP $u_{xx} - y^{2n}u_{yy} = ny^{2n-1}u_y$ dans $\mathbb{R} \times]0,+\infty[$.

1. Montrer que l'EDP est hyperbolique et déterminer ses courbes caractéristiques.
2. Écrire l'EDP sous sa première forme canonique.
3. Montrer que la solution générale est de la forme :

$$u(x,y) = f(y^{-n+1} + (n-1)x) + g(y^{-n+1} - (n-1)x)\,,$$

où f et g sont des fonctions de classe $\mathcal{C}^2$.

III.2.8 On considère l'équation des cordes dissipative :

$$\begin{aligned}
&u_{tt} - c^2u_{xx} + \alpha u_t = 0\,, \quad (t,x) \in]0,+\infty[\times]0,L[\,,\\
&u(t,0) = u(t,L) = 0\,, \quad t \in \mathbb{R}_+\,,\\
&u(0,x) = f(x)\,, \quad u_t(0,x) = g(x)\,, \quad x \in [0,L]\,,
\end{aligned}$$

où $c, \alpha, L > 0$, $f \in \mathcal{C}^2([0,L],\mathbb{R})$ et $g \in \mathcal{C}^1([0,L],\mathbb{R})$. On suppose que f et g et leurs dérivées s'annulent aux bords. On applique la méthode de séparation de variables et l'on recherche une solution de la forme $u(t,x) = T(t)X(x)$.

1. Montrer que X et T doivent vérifier $X'' + \lambda X = 0$ et $T'' + \alpha T' + \lambda c^2 T = 0$ où $\lambda > 0$ est une constante.
2. Montrer qu'il existe une suite de solutions $\lambda = \lambda_n$ et $X(x) = X_n(x)$, $n \geqslant 1$, où $\lambda_n = n^2\pi^2/L^2$ et $X_n(x) = A_n\sin(n\pi x/L)$.
3. Supposons $4\pi^2c^2/L^2 > \alpha^2$. Montrer que la solution est de la forme :

$$u(t,x) = \sum_{n=1}^{+\infty} e^{-\frac{\alpha t}{2}}\left[a_n\cos\Big(\frac{\delta_n t}{2}\Big) + b_n\sin\Big(\frac{\delta_n t}{2}\Big)\right]\sin\Big(\frac{n\pi x}{L}\Big)\,,$$

où $\delta_n = \sqrt{4c^2\lambda_n - \alpha^2}$, et identifier a_n et b_n.

4. Étudier le cas où $4\pi^2c^2/L^2 = \alpha^2$, puis le cas où $4\pi^2c^2/L^2 < \alpha^2$.

Constructions de $\mathbb{R}$ et introduction aux nombres p-adiques

III.3

Jean Dieudonné disait en 1960[1] que les constructions de $\mathbb{R}$ « n'ont pas grand intérêt en analyse » et qu'il vaut mieux adopter une démarche axiomatique. Pour des débutants, la présentation axiomatique est sans doute la plus raisonnable et c'est le point de vue que nous avons adopté dans le livre [L1], en le tempérant toutefois par l'exposé des grandes lignes d'une construction de $\mathbb{R}$ par les écritures décimales[2]. Mais les choses ont changé depuis 1960 et même depuis la première édition du livre [L1]. Les algorithmes ont envahi le monde et les étudiants disposent dès le début du cycle universitaire d'outils de calcul numérique et formel. Ils peuvent ainsi soupçonner que les réels vus par une machine ne sont pas vraiment les réels des axiomes... Pour cette raison et pour bien d'autres, il est intéressant de voir « ce que l'on a caché sous le tapis » en employant la « boite noire » des axiomes. Si le point de vue axiomatique reste à privilégier pour le débutant, il nous semble important d'aller plus loin en troisième année d'études universitaires. L'histoire montre que la conception des nombres pose des problèmes profonds et difficiles, qu'il n'est pas raisonnable d'ignorer totalement (en particulier pour les étudiants qui se destinent à l'enseignement). Nous pensons, par ailleurs, que l'on n'a pas une connaissance « technique » vraiment satisfaisante de $\mathbb{R}$ pour être à l'aise en analyse si l'on se limite à une approche axiomatique. Nous pensons aussi que le sujet ne relève pas nécessairement de ce que Dieudonné appelait des *mathématiques ennuyeuses* et nous espérons en convaincre le lecteur.

Plusieurs parties de ce module sont consacrées aux constructions « classiques » du corps $\mathbb{R}$ des nombres réels (plus précisément d'un corps satisfaisant les axiomes des nombres réels). Nous verrons aussi que des variantes « naturelles » de *deux* de ces constructions conduisent à d'autres « corps de nombres », les corps p-adiques $\mathbb{Q}_p$, dont nous donnerons les propriétés de base. Ainsi, tout en approfondissant l'étude de $\mathbb{R}$, nous proposerons au lecteur une première promenade dans le monde p-adique. Celle-ci propose des vues sur les divers as-

[1] Dans la première édition, en anglais, de [D1].

[2] Dans [Ru], W. Rudin choisit une approche similaire à la notre en adoptant la démarche axiomatique tout en donnant en appendice une construction, celle par les coupures de Dedekind.

pects du paysage *tout en restant à un niveau élémentaire*[3], En étudiant ce module, le lecteur devrait voir apparaître « en filigrane » une idée simple[4] (bien qu'a priori non intuitive...) : le corps $\mathbb{R}$ des nombres réels et les corps $\mathbb{Q}_p$ des nombres p-adiques « permettent d'observer un même monde à des endroits différents ». Du point de vue de l'arithmétique, il n'y a aucune raison de n'observer le vaste monde qu'à un seul endroit et l'utilisation des corps $\mathbb{Q}_p$ est aussi légitime que celle de $\mathbb{R}$.

Dans le livre [L1], nous avons construit le corps $\mathbb{R}$ des nombres réels à partir de la notion *d'écriture décimale*. Dans la présentation de cette construction, nous avons admis *trois* résultats importants dont la preuve est un peu délicate. Nous démontrerons ces résultats dans ce module, ce qui complétera la construction.

Nous donnerons aussi deux autres constructions de $\mathbb{R}$. L'une est basée sur la notion de *section commençante* (une variante de la notion de *coupure de Dedekind*), l'autre sur la notion de *suite de Cauchy*, introduite dans le livre [L2].

Pour l'historique des diverses constructions, nous renvoyons à l'introduction de la partie IV.1.3 du livre [L1, page 577]. Nous suggérons au lecteur de la (re)lire avant de commencer l'étude de ce module.

Nous terminerons le module par un bref historique de la notion de nombre et une discussion des relations entre les nombres et le monde réel. Ceci nous conduira à quelques réflexions sur la notion de *calculabilité* pour les nombres réels (due à Alan Turing), en relation avec les fondements des mathématiques et la naissance de l'informatique. Nous préciserons en particulier les différences (considérables...) entre les réels définis par les axiomes et les réels vus par une machine.

Convention de lecture. Dans chacune des parties 1, 3, 4, l'on part du corps $\mathbb{Q}$ et l'on construit un corps satisfaisant aux axiomes du corps des réels, que nous allons rappeler. La règle du jeu est naturellement que l'on *oublie* chaque fois le corps $\mathbb{R}$ utilisé dans les deux livres [L1, L2] (excepté, bien sûr, pour ce qui concerne la construction de $\mathbb{R}$ par les écritures décimales, pour la partie 1, *cf.* [L1, IV.1.3]). Quand on parlera de $\mathbb{R}$ dans l'une de ces trois parties, il s'agira donc exclusivement de l'ensemble $\mathbb{R}$ *qui a été défini dans cette partie*. Nos prouverons que les corps obtenus dans chaque cas sont deux à deux « naturellement isomorphes ». Ce résultat « d'unicité de $\mathbb{R}$ », annoncé dans le livre [L1], sera démontré dans ce module . Nous donnerons deux preuves très différentes.

Pour chacune les deux premières constructions, la définition de *l'ensemble* $\mathbb{R}$ est simple et intuitive, ce qui est plus difficile et de munir cet ensemble d'une structure de corps ordonné et de vérifier les axiomes des réels. Pour la troisième, la définition de *l'ensemble* $\mathbb{R}$ est plus technique, mais la définition et l'étude des opérations est, par contre, beaucoup plus simple[5].

[3] Le lecteur intéressé pourra approfondir en consultant, par exemple, les monographies : Yvette Amice, *Les nombres p-adiques*, PUF, Fernando Q. Gouvêa, *p-adic numbers, An Introduction*, Universitext, Springer, Alain M. Robert, *A Course in p-adic Analysis*, Graduate texts in Mathematics, Springer, et [BCh]

[4] Cette idée est née en Allemagne vers la fin du XIXe siècle. Elle est essentiellement due à Leopolod Kronecker $(1823 - 1891)$ et à son élève Kurt Hensel. Kronecker a dit (en version française) : *« Dieu a fait les nombres entiers, tout le reste est l'œuvre de l'homme »*. Diverses constructions algorithmiques, à partir des écritures à droite et à gauche, présentées dans ce module, illustrent bien ce point de vue, mais nous utilisons *aussi* de façon essentielle la théorie des ensembles de Cantor, que rejetait vigoureusement Kronecker.

[5] Il y a toutefois quelques difficultés pour la relation d'ordre.

Nous rappelons les axiomes des nombres réels. Pour les définitions nécessaires nous renvoyons le lecteur à la partie IV.1, 1.2 du livre [L1].

Définition 1 (Axiomes des nombres réels). On dit qu'un ensemble vérifie les axiomes des nombres réels s'il est muni d'une structure de corps totalement ordonné contenant $\mathbb{Q}$ comme sous-corps et possédant la propriété de la borne supérieure.

Il y a de nombreuses variantes équivalentes des axiomes des nombres réels, *cf.* l'exercice III.3.36 de la page 528.
Dans ce module on considérera parfois un corps $\mathcal{R}$ satisfaisant les axiomes des nombres réels. Les propriétés établies pour $\mathbb{R}$, dans la partie IV.1, 1.2 du livre [L1], l'ont été *à partir de ces axiomes*, elles sont donc valables pour un tel corps $\mathcal{R}$ et nous pourrons les utiliser dans ce cadre.

1 Constructions à partir d'écritures décimales

Le propos de cette partie est double. D'une part nous compléterons la construction de $\mathbb{R}$ à partir des écritures décimales présentée dans le livre [L2]. D'autre part nous donnerons une « imitation » de cette construction (une vue par un « miroir déformant ») où l'on remplace les développements décimaux *illimités à droite* par des développement décimaux *illimités à gauche* (on autorise un nombre fini de chiffres *après* la virgule, au lieu d'un nombre fini de chiffres *avant* celle-ci). Cette variante conduit a de nouveaux nombres, les nombres décadiques (ou 10-adiques). En remplaçant la base de la numération 10 par $b \in \mathbb{N}$ ($b \geqslant 2$), on obtient les nombres b-adiques. Mais, contrairement au cas des écritures illimitées à droite[6], le résultat *dépend de* b. Si b est premier, on le note traditionnellement p et l'on parle de nombres p-adiques. Si b n'est pas premier (par exemple $b = 10$), nous verrons qu'il y a quelques « pathologies ».
Historiquement, la première construction rigoureuse des réels par les écritures décimales semble due au mathématicien allemand Paul Du Bois-Reymond (*Théorie générale des fonctions*[7]). Cette construction est aussi celle proposée plus tard par Lebesgue[8] (*Sur la mesure des grandeurs*, 1956).

1.1 Écritures décimales

Nous faisons quelques rappels sur les écritures décimales et la représentation décimale des rationnels. Pour plus de détails, on se reportera à [L1, IV.1, 3.1 et IV.1, 3.2.]

Définition 2. On appelle écriture décimale toute « suite » de la forme $\pm n_0{,}\alpha_1\alpha_2\ldots\alpha_s\ldots$, où $\pm$ est $+$ ou $-$, $n_0 \in \mathbb{N}$, et $\alpha_i \in [\![0,9]\!]$ pour tout $i \in \mathbb{N}^*$.

Plus formellement, une écriture décimale est un élément de l'ensemble $\{\pm\} \times \mathbb{N} \times [\![0,9]\!]^{\mathbb{N}}$. On peut convenir que $n_0 \in \mathbb{N}$ est écrit en écriture décimale. Les α_i sont appelés *chiffres*.

[6] Dans ce cas le résultat est unique à isomorphisme unique près.
[7] *Allgemeine Funktionentheorie*, 1882.
[8] Lebesgue, pensant en particulier à la présentation des réels au lycée ou en début d'études universitaires, repart de la version grecque de nombre par mesure des grandeurs.

Par définition, on obtient l'écriture décimale $dec(\delta)$, ou développement décimal, d'un nombre décimal δ en complétant la représentation décimale usuelle avec un nombre *infini* de zéros *à droite*. Ceci permet d'identifier l'ensemble $\mathcal{D}$ des décimaux à un sous-ensemble de l'ensemble des écritures décimales.
Une écriture décimale est dite *impropre* si ses décimales sont toutes des 9 à partir d'un certain rang ou bien si elle est de la forme $-0,000\ldots$. Dans le cas contraire l'écriture est dite *propre* (*cf.* [L2, IV.1, 3.1.1, définition 29]).

1.2 Construction de $\mathbb{R}$ à partir des écritures décimales

Pour cette partie, nous renvoyons le lecteur à la partie IV.1, 3 du livre [L1]. Notre propos est de donner la démonstration du théorème 1, du théorème et définition 3 et du théorème et définition 5 énoncés ci-dessous, qui avaient été admis dans [L1, IV.1, 3]. L'ensemble $\mathbb{R}$ des nombres réels est *défini* comme l'ensemble des écritures décimales *propres* (*cf.* [L2, IV.1, 3.2.1, définition 31]). L'écriture décimale (ou développement décimal) d'un rationnel $\tau \in \mathbb{Q}$ est notée $dec(\tau) \in \mathbb{R}$. L'application $dec : \mathbb{Q} \to \mathbb{R}$ est injective (*cf.* [L1, IV.1, 3.2.1, proposition 61]).

Méthode de lecture de cette partie.
Lire la partie IV.1, 3 du livre [L1] en se reportant au présent texte pour les preuves des trois énoncés admis et quelques compléments (*cf.* 1.2.2).

1.2.1 Théorème de la borne supérieure

C'est *le* point fondamental de la construction de $\mathbb{R}$ à partir des écritures décimales.
La preuve du résultat suivant avait été admise dans le livre [L1]. Nous la donnons ici.

Théorème 1 (Borne supérieure). Tout sous-ensemble non vide et majoré de $\mathbb{R}$ admet une borne supérieure.
De même, tout sous-ensemble non vide et minoré admet une borne inférieure.

Avant de donner la démonstration, commençons par nous interroger sur ce que signifie être la borne supérieure ou inférieure d'un ensemble de réels, quand ceux-cis sont définis par des écritures décimales propres.

Lemme 2. Soit A une partie non vide de $\mathbb{R}_+$.

- Si un réel positif $x = n_0, \alpha_1 \ldots \alpha_h \ldots$ est la borne inférieure de A alors, il vérifie la propriété suivante, notée (P_A) : l'ensemble des minorants de A contient l'entier n_0 ainsi que toutes les troncatures $x_k = n_0, \alpha_1 \ldots \alpha_k$ (pour tout entier $k \geqslant 0$), mais ne contient pas $\tilde{x}_0 := n_0 + 1$ ni aucun nombre de la forme $\tilde{x}_k = x_k + 10^{-k}$.
- Réciproquement, si une écriture décimale $n_0, \alpha_1 \ldots \alpha_h \ldots$ $(n_0 \geqslant 0)$ vérifie la propriété (P_A), alors elle est propre et le réel $x = n_0, \alpha_1 \ldots \alpha_h \ldots$ est la borne inférieure de A.

On peut énoncer un lemme similaire à propos de la borne supérieure. On le prouverait de façon analogue

Démonstration du lemme

- Montrons d'abord que si $x = \inf A$, alors les propriétés annoncées sont vraies. Tous les nombres x_k, ainsi que n_0 sont inférieurs ou égaux à x. La borne infèrieure étant un minorant, tous ces autres nombres sont donc aussi minorant. Par contre, $x < n_0+1$ et $x < x_k + 10^{-k}$ (cf l'exercice 2). Donc, si l'un de ces nombres était minorant, l'on aurait un minorant strictement plus grand que x, en contradiction de la définition de la borne inférieure, le plus grand des minorants.
- Montrons maintenant la réciproque. Commençons par vérifier que l'écriture est propre.

  Supposons d'abord qu'il existe $k \in \mathbb{N}^*$ tel que $\alpha_k \neq 9$; posons $\alpha'_k := \alpha_k + 1$. Alors $x_k = n_0, \alpha_1 \dots \alpha_k$ est un minorant de A tandis que $\tilde{x}_k = x_k + 10^{-k} = n_0, \alpha_1 \dots \alpha'_k$ ne l'est pas. Il existe donc un élément $y = m_0, \beta_1 \dots \beta_k \dots$ de A tel que $x_k \leqslant y < x_k + 10^{-k}$. On a alors $m_0 = n_0$ ainsi que $\beta_1 = \alpha_1, \dots, \beta_k = \alpha_k$ et, comme y est une écriture décimale propre, il existe $p > k$ tel que $\beta_p < 9$ et l'on ne peut donc pas avoir :

  $$\alpha_{k+1} = \dots = \alpha_p = 9 \text{ sinon } x_p := n_0, \alpha_1 \dots \alpha_p > y.$$

  Donc, pour tout k tel que $\alpha_k < 9$, il existe $p > k$ tel que $\alpha_p < 9$.

  En utilisant que $n_0 + 1$ n'est pas un minorant de A, On démontre de la même façon qu'il existe k tel que $\alpha_k \neq 9$. Ainsi, la représentation décimale $n_0, \alpha_1 \dots \alpha_i \dots$ est *propre*.

  Vérifions maintenant que $x = n_0, \alpha_1 \dots \alpha_i \dots$ est bien la borne inférieure de A. Montrons d'abord que c'est un minorant. Supposons, par contradiction, l'existence d'un réel $y \in A$ tel que $y < x$. Si $y = m_0, \beta_1 \beta_2 \dots$, alors ou bien $m_0 < n_0$, ou bien $m_0 = n_0$ et il existe un indice $k \geqslant 1$ tel que $\alpha_k > \beta_k$ et $\alpha_i = \beta_i$ pour les indices $i < k$. Dans le premier cas, on a $y < n_0$. Dans le second, on a $y < x_k$. Ainsi, dans les deux cas, le nombre $y \in A$ est strictement plus petit qu'un minorant de A, ce qui est absurde.

  Supposons à présent que x ne soit pas le plus grand minorant. Il existe donc un minorant $z = p_0, \gamma_1 \gamma_2 \dots > x$. Encore une fois, les règles de comparaison nous disent que, ou bien $p_0 > n_0$, ou bien $p_0 = n_0$ et il existe un indice $k \geqslant 1$ tel que $\gamma_k > \alpha_k$, avec $\gamma_i = \alpha_i$ pour les indices $i < k$. Dans le premier cas, $p_0 > n_0$ implique $p_0 \geqslant n_0 + 1$, car ce sont des nombres entiers, et donc $z \geqslant p_0 \geqslant n_0 + 1$. Dans le second cas, $z \geqslant p_0, \gamma_1 \dots \gamma_k \geqslant x_k + 10^{-k}$. Ainsi z est plus grand qu'un nombre qui, par hypothèse, n'est pas un minorant et l'on aboutit à une contradiction. ■

Démonstration du théorème 1 Il est plus simple de regarder le cas d'un ensemble de réels positifs. Il est un peu plus facile aussi de considérer d'abord le cas de la borne *inférieure*.

Soit A un ensemble non vide et minoré de réels positifs (c'est-à-dire $\forall x \in A,\ x \geqslant 0$). Si $0 \in A$, le problème est réglé : $\inf A = 0$ et A a même un élément minimal. Sinon, comme A est non vide, il possède au moins un élément $\tilde{a} > 0$.

Notons $\mathcal{M}$ l'ensemble des minorants de A. Nous cherchons à montrer qu'il admet un élément maximal. Pour cela, on va établir l'existence d'un nombre x satisfaisant les conditions du lemme.

Notons d'abord que $0 \in \mathcal{M}$, puisque tous les éléments de A sont positifs. En revanche, il existe un nombre naturel n qui est strictement supérieur à $\tilde{a}$. Ce nombre n'est pas un minorant, ni a fortiori aucun autre plus grand que lui, donc $\mathcal{M}$ ne contient qu'un nombre fini de nombres entiers naturels. Notons n_0 le plus grand de tous : ce sera la partie entière de notre future borne inférieure x, puisque par définition $n_0 + 1 \notin \mathcal{M}$, comme requis par le lemme.

Nous construisons maintenant successivement les décimales α_k de x . Supposons avoir déjà construit les $k-1$ premières, pour un certain $k \geqslant 1$ (aucune donc, si $k = 1$). D'après ce qui est demandé dans le lemme, nous voulons trouver une décimale α_k telle que :

$$x_k := n_0, \alpha_1 \dots \alpha_k \in \mathcal{M} \qquad \text{et} \qquad x_k + 10^{-k} \notin \mathcal{M} \tag{1}$$

(l'hypothèse de récurrence est que nous avons ces relations à l'indice $k-1$).

Il n'y a que dix possibilités pour α_k, à savoir les dix décimales $0, \dots, 9$. Considérons donc les dix nombres $n_0, \alpha_1 \dots \alpha_{k-1}0$, $\dots$, $n_0, \alpha_1 \dots \alpha_{k-1}9$. Par hypothèse de récurrence, le premier au moins est dans $\mathcal{M}$, parce qu'il est égal à x_{k-1}. Choisissons α_k de sorte que $n_0, \alpha_1 \dots \alpha_k$ soit le plus grand de ces dix nombres appartenant à l'ensemble de minorants $\mathcal{M}$. Par définition même, on a donc $x_k \in \mathcal{M}$.

Si $\alpha_k < 9$, alors $x_k + 10^{-k}$ est simplement le suivant des dix nombres mentionnés ci-dessus, et ; par hypothèse, il n'est pas dans $\mathcal{M}$: ainsi (1) est vraie.

Si $\alpha_k = 9$, alors $x_k + 10^{-k}$ est égal à $x_{k-1} + 10^{-k+1}$ parce que l'addition sur la dernière décimale propage une retenue (ou à $n_0 + 1$ dans le cas particulier $k = 1$). Donc l'hypothèse de récurrence (ou le fait que $n_0 + 1 \notin \mathcal{M}$) nous dit que la propriété (1) est aussi vraie au rang k.

Le lemme nous montre que l'écriture décimale $x = n_0, \alpha_1 \alpha_2 \dots$ est un réel et que c'est la borne inférieure de A. On a donc fini de montrer que tout ensemble non vide de réels positifs a une borne infèrieure.

Soit maintenant A un ensemble non vide et majoré de réels positifs. L'ensemble $\mathcal{M}$ des majorants de A est donc non vide et minoré par n'importe quel élément de A. Il admet donc une borne inférieure x. Montrons que x est dans $\mathcal{M}$. Si ce n'était pas le cas, il existerait $y \in A$ tel que $x < y$. Mais alors y serait un minorant de $\mathcal{M}$ ce qui contredirait le fait que x est le plus grand des minorants de A. Ainsi x est le plus petit élément de $\mathcal{M}$, donc en particulier la borne supérieure de A.

Pour des ensembles de réels négatifs, on peut se ramener aux cas précédents en posant $-A := \{x \in \mathbb{R} \mid -x \in A\}$ et en remarquant que $\sup A = -\inf(-A)$ et $\inf A = -\sup(-A)$.

Enfin, pour un ensemble A non vide contenant des réels positifs et des négatifs, on se ramène aux cas précédents en considèrant $A \cap \mathbb{R}_+$ et $A \cap \mathbb{R}_-$. ■

1.2.2 Exercices sur les bornes supérieures et les bornes inférieures

Notation. Soient $x := m_0, \alpha_1, \alpha_2 \dots \alpha_s \dots$ un nombre réel *positif* et, pour tout $h \in \mathbb{N}$, $x_h := m_0, \alpha_1, \alpha_2 \dots \alpha_h$ la troncature de x à l'ordre h (*cf.* [L1, IV.1, 3.1, définition 30]). On note $\tilde{x}_h := x_h + 10^{-h}$. On prendra garde au fait que x_h est une troncature mais que $\tilde{x}_h$ n'en est pas une. C'est un abus de notation.

Exercice 1.

Soit $d \in \mathcal{D}$ un décimal. On note x son écriture impropre. Montrer que d est la borne supérieure de l'ensemble $\{x_n\}_{n\in\mathbb{N}}$.

Solution.

On note $d := n_0, \alpha_1 \dots \alpha_p$, avec $\alpha_p > 0$. On a $x = n_0, \alpha_1 \dots (\alpha_p - 1)999\dots 9\dots$. On a, pour tout $m \in \mathbb{N}$, $x_m = n_0, \alpha_1 \dots (\alpha_p - 1)999\dots 9 < d$. Ainsi d est un majorant de $X := \{x_m\}_{m\in\mathbb{N}}$ et c'est le plus petit, donc la borne supérieure. Supposons en effet qu'il existe un majorant M de X, de développement décimal propre $n_0, \beta_1 \dots \beta_k \beta_{k+1} \dots, \beta_n \dots$ tel que $M < d$. On a $\beta_h = \alpha_h$ pour tous les indices h inférieurs ou égaux à un certain entier $N - 1$, puis $\beta_N < \alpha_N$. Mais alors $M < x_N$ et M ne majore pas X d'où une contradiction.

Exercice 2.

Soit $x := m_0, \alpha_1, \alpha_2 \dots \alpha_s \dots$ un nombre réel *positif*

(i) Montrer que la suite $(\tilde{x}_h = x_h + 10^{-h})_{h\in\mathbb{N}^*}$ est décroissante et que, pour tout $h \in \mathbb{N}^*$, $x < \tilde{x}_h$.

(ii) Soit $y \in \mathbb{R}$, tel que $x < y$. Montrer qu'il existe $k \in \mathbb{N}$, tel que $\tilde{x}_k < y$.

Solution. On a $\tilde{x}_h \in \mathcal{D}$ et $x_h < \tilde{x}_h$.

(i) On a $x_{h+1} \leqslant x_h + 9 \times 10^{-h-1}$, $\tilde{x}_{h+1} = x_{h+1} + 10^{-h-1} \leqslant x_h + 10^{-h} = \tilde{x}_h$.
La suite $(\tilde{x}_h)_{h\in\mathbb{N}^*}$ est donc décroissante.
On convient de $\alpha_0 := m_0$ et $\beta_0 := n_0$. Soit $h \in \mathbb{N}*$. Notons $\tilde{x}_h = n_0, \beta_1, \beta_2 \dots \beta_h$. Alors :
   - ou bien $m_0 < n_0$, ce qui entraîne $x < \tilde{x}_h$;
   - ou bien $m_0 = n_0$ et, en utilisant $x_h < \tilde{x}_h$, on montre qu'il existe $h' \in \mathbb{N}^*$, $h' \leqslant h$, tel que $\alpha_1 = \beta_1, \dots, \alpha_{h'-1} = \beta_{h'-1}$, $\alpha_{h'} < \beta_{h'}$; on a alors, d'après la définition de la relation d'ordre :
   $$x = m_0, \alpha_1, \alpha_2 \dots \alpha_{h'-1}\alpha_h \dots < m_0, \alpha_1, \alpha_2 \dots \alpha_{h'-1}\beta_{h'} \dots \beta_h = \tilde{x}_{h'} \leqslant \tilde{x}_h.$$

Ainsi $x < \tilde{x}_h$ dans tous les cas.

(ii) Il existe $u \in \mathbb{R}$ tel que $x < u < y$. Soient $u := m_0, \beta_1, \beta_2 \dots \beta_s \dots$ et, pour tout $h \in \mathbb{N}$, u_h la troncature de u à l'ordre h. Puisque $x < u$, il existe $k \in \mathbb{N}$ *minimal* tel que $\alpha_k < \beta_k$ (en convenant $\alpha_0 := m_0$ et $\beta_0 := n_0$). Alors $m_0 < n_0$ ou $k \geqslant 1$ et $x_{k-1} = u_{k-1}$, $x_k < y_k$, $\alpha_k < \beta_k$. Par suite : $x_k < \tilde{x}_k \leqslant u < y$.

Exercice 3.

Soit $x > 0$ un réel, écrit en décimal sous la forme $n_0, \alpha_1\alpha_2 \dots$. Notons x_k sa troncature à la $k^{ème}$ décimale et $X \subset \mathcal{D}$ l'ensemble de toutes ces troncatures ($k \in \mathbb{N}$).

(i) Montrer que x est la borne supérieure de X. Dans quels cas est-il le maximum ? Que peut-on dire dans le cas où $x < 0$?

(ii) On note $\tilde{x}_k := x_k + 10^{-k}$ et $\tilde{X} := \{\tilde{x}_k\}_{\in\mathbb{N}}$. Montrer que $x < \tilde{x}_k$, que la suite $(\tilde{x}_k)_{k\in\mathbb{N}}$ est décroissante et que x est la borne inférieure de $\tilde{X}$.

(iii) Montrer que les suites de décimaux $(x_k)_{k\in\mathbb{N}}$ et $(\tilde{x}_k)_{k\in\mathbb{N}}$ sont *adjacentes* (*cf.* [L1, IV.1, 2.3.3, définition 24]).

Solution.

(i) On procède comme dans l'exercice 1 ci-dessus. Les chiffres α_h sont les mêmes pour x et x_k pour $h \leqslant k$ et vérifient $\alpha_h \geqslant 0$ pour $h > k$, on peut donc écrire $x_k \leqslant x$ et x est un majorant de X et c'est le plus petit, donc la borne supérieure. Supposons en effet qu'il existe M de développement décimal propre $n_0, \beta_1 \ldots \beta_k \beta_{k+1} \ldots, \beta_n \ldots$ tel que $M < x$. On a $\beta_h = \alpha_h$ pour tous les indices h inférieurs ou égaux à un certain entier $N-1$, puis $\beta_N < \alpha_N$. Mais alors $M < x_N$ et M ne majore pas X.

Pour que $x = \max X$, il faut et il suffit que $x \in X$, ce qui exige que les α_h soient nuls au delà d'un certain rang, donc que x soit un décimal. la réciproque est évidente.

De la même façon, on montre que si $x < 0$, c'est la borne inférieure de X.

(ii) On a vu dans l'exercice 2 que $x < \tilde{x}_k$ et que la suite $(\tilde{x}_k)_{k\in\mathbb{N}}$ est décroissante.

On a, pour tout $k \in \mathbb{N}$, $x_0 \leqslant \tilde{x}_k$. L'ensemble $\tilde{X}$ est non vide et minoré, il admet donc une borne inférieure a. On a $x \leqslant a$. Pour montrer $x = a$, on raisonne par l'absurde. On suppose $x < a$. D'après l'exercice 2, il existe alors $k \in \mathbb{N}$ tel que $\tilde{x}_k < a$. Mais $a \leqslant \tilde{x}_k$. On a donc une contradiction.

(iii) Les suites sont adjacentes puisque la suite de décimaux $(10^{-k})_{k\in\mathbb{N}}$ tend vers 0 quand k tend vers $+\infty$

1.2.3 Somme de deux réels

La preuve du résultat suivant avait été admise dans le livre [L1]. Nous la donnons ici.

Théorème et définition 3. Soient $x = m_0{,}\alpha_1\alpha_2\ldots$ et $y = n_0{,}\beta_1\beta_2\ldots$ deux réels positifs. Pour tout $k \in \mathbb{N}^*$, notons $x_k := m_0{,}\alpha_1\ldots\alpha_k$ et $y_k := n_0{,}\beta_1\ldots\beta_k$, qui sont des nombres décimaux. Alors l'ensemble :

$$A := \{x_1 + y_1, x_2 + y_2, \ldots\} = \{x_k + y_k \mid k \in \mathbb{N}^*\}$$

est non vide et majoré. Sa borne supérieure (dans $\mathbb{R}$) est appelée la *somme* de x et y : on la note $x + y$.

Notons $\tilde{y}_k := y_k + 10^{-k}$; l'ensemble :

$$B := \{x_1 - \tilde{y}_1, x_2 - \tilde{y}_2, \ldots\} = \{x_k - \tilde{y}_k \mid k \in \mathbb{N}^*\}$$

est également majoré. Sa borne supérieure est appelée *différence* de x et y, on la note $x - y$.

Si x et y sont des décimaux la somme et la différence définies ci-dessus coïncident avec la somme et la différence usuelles.

Autrement dit, pour additionner, on *tronque* les nombres à la $k^{\text{ème}}$ décimale, on ajoute les décimaux obtenus, et finalement on prend le plus petit réel dépassant tous les résultats obtenus de cette façon.

Démonstration. Soient n et m deux entiers tels que $n > x$ et $m > y$. Comme pour tout k, on a $x_k \leqslant x < n$, $y_k \leqslant y < m$, on a $x_k + y_k < n + m$ et $n + m$ est donc un majorant de A. Par suite A admet une borne supérieure μ. On définit $x + y := \mu$.

Montrons que, lorsque x et y sont décimaux, on a bien $\mu = x + y$ (au sens de la somme ordinaire des décimaux). En effet, dans ce cas, toutes les décimales de x sont nulles à partir d'un certain rang k_1 et celles de y à partir d'un rang k_2. Pour tout $k \geqslant \max(k_1, k_2)$, $x_k + y_k = x + y$ et A est donc un ensemble fini. Comme par ailleurs $x_k \leqslant x$ et $y_k \leqslant y$ pour tout k, $x + y$ est en l'élément maximal de A, donc aussi sa borne supérieure.

La démonstration pour B est similaire. On utilise un résultat de l'exercice 3 : on a $y < \tilde{y}_k$, d'où l'on déduit $-\tilde{y}_k < -y$. Si m' est un entier tel que $-y < m'$, on a $x_k - \tilde{y}_k < n + m'$ et $n + m'$ est un majorant de B. ∎

Remarque. Comme nous l'avons expliqué dans le livre [L1] (*cf.* IV.1, 3.4, page 591), la somme de deux réels *n'est pas effective*. En général on ne peut pas la « calculer » (*cf.* l'exercice III.3.3 de la page 521) : pour $k \in \mathbb{N}$ donné, on ne peut pas en général calculer la troncature $(x + y)_k$ de la somme à l'aide de la donnée des deux troncatures x_m et y_m si grand que soit m. Dans certains cas, il faudrait « connaître toutes les décimales de x et y » et il y en a une infinité.[9]

1.2.4 Multiplication des réels. Inverses

Produit de réels

On rappelle la définition du produit dans $\mathbb{R}$ (cf. [L1, IV.1 3.4.2]). La multiplication de deux réels se définit de manière similaire à ce que nous avons fait pour l'addition. Il faut commencer par des nombres positifs. Les définitions s'étendent ensuite aux autres réels avec les règles de signes habituelles.

Théorème et définition 4. Soient x et y deux réels *positifs*. Pour tout $k \in \mathbb{N}^*$, notons x_k et y_k leurs troncatures respectives à l'ordre k. Alors l'ensemble de nombres rationnels :

$$C := \{x_1 y_1, x_2 y_2, \ldots\} = \{x_k y_k \mid k \in \mathbb{N}^*\}$$

est non vide et majoré. Sa borne supérieure est appelée le *produit* de x et y : on le note xy ou $x \times y$.

Inverse d'un réel pour la multiplication

Il reste à montrer que tout élément non nul de $\mathbb{R}$ admet un inverse pour la multiplication. La définition de l'inverse est plus délicate que celle de l'addition et de la multiplication car l'inverse d'un décimal est un rationnel mais n'est pas, en général, un décimal.

Soit $x \in \mathbb{R}$, $x > 0$. Pour tout $k \in \mathbb{N}$, on note, comme plus haut, $\tilde{x}_k = x_k + 10^{-k} > 0$. On a $\tilde{x}_k \in \mathbb{Q}^*$, donc $\frac{1}{\tilde{x}_k} \in \mathbb{Q}$. Pour allèger la notation on identifie $\frac{1}{\tilde{x}_k} \in \mathbb{Q}$ à $dec\left(\frac{1}{\tilde{x}_k}\right) \in \mathbb{R}$.

La preuve du résultat suivant avait été admise dans le livre [L1]. Nous la donnons ici.

[9] La différence entre « la définition de la somme » et « le calcul de la somme » correspond épistémologiquement à la différence aristotélicienne entre *l'infini actuel* (ou *en acte*) et *l'infini potentiel* (ou *en puissance*). L'infini actuel est le « cadre naturel » de la théorie des ensembles, l'infini potentiel est celui de la notion de calculabilité.

Théorème et définition 5. Soit $x \in \mathbb{R}$. Supposons $x > 0$ et notons, comme plus haut, $\tilde{x}_k := x_k + 10^{-k}$. Alors l'ensemble :

$$D := \{1/\tilde{x}_0, 1/\tilde{x}_1, \ldots\} = \left\{1/\tilde{x}_k \;\middle|\; k \in \mathbb{N}\right\}.$$

est non vide et majoré dans $\mathbb{R}$. Sa borne supérieure est appelée *l'inverse* de x, on le note $1/x$ ou $\frac{1}{x}$. On a $x\frac{1}{x} = 1$.
Les définitions des inverses dans $\mathbb{Q}$ et dans $\mathbb{R}$ sont compatibles.

Si $x < 0$, on définit son inverse par $\frac{1}{x} := -\frac{1}{-x}$.

Démonstration. On a, pour tout $k \in \mathbb{N}^*$, $x_k \leqslant x < \tilde{x}_k$. Les $\tilde{x}_k$ sont des décimaux, donc des rationnels et, puisqu'ils sont strictement positifs, ils admettent des inverses rationnels tels que $\frac{1}{\tilde{x}_k} < \frac{1}{x}$. Puisque $x > 0$, il existe $m \in \mathbb{N}^*$ tel que $10^{-m} < x$. Donc $\frac{1}{\tilde{x}_k} < 10^m$. Ainsi l'ensemble D est majoré.

Il existe $k_0 \in \mathbb{N}^*$ tel que, pour tout $k \geqslant k_0$, $x_k > 0$. Pour tout $k \geqslant k_0$, on a $\frac{1}{\tilde{x}_k} \leqslant \frac{1}{x} \leqslant \frac{1}{x_k}$. Par suite :

$$x_k\frac{1}{\tilde{x}_k} \leqslant x\frac{1}{x} \leqslant \tilde{x}_k\frac{1}{x_k} \quad \text{et} \quad \frac{\tilde{x}_k - 10^{-k}}{\tilde{x}_k} \leqslant x\frac{1}{x} \leqslant \frac{x_k + 10^k}{x_k},$$

donc $1 - \frac{10^{-k}}{\tilde{x}_k} \leqslant x\frac{1}{x} \leqslant 1 + \frac{10^k}{x_k}$ Quitte à augmenter m, on peut supposer que $10^{-m} < x_{k_0}$.

Alors, pour tout $k \geqslant k_0$: $1 - 10^{-k+m} \leqslant x\dfrac{1}{x} \leqslant 1 - 10^{-k+m}$. On a donc $x\frac{1}{x} = 1$.

Soient $\xi \in \mathbb{Q}^*$ et $\frac{1}{\xi}$ son inverse dans $\mathbb{Q}$. On note $x = dec(\xi)$. On a :

$x\ dec\left(\frac{1}{\xi}\right) = dec\left(\xi\,\frac{1}{\xi}\right) = 1$, donc : $dec\left(\frac{1}{\xi}\right) = \left(\frac{1}{x}\,x\right) dec\left(\frac{1}{\xi}\right) = \frac{1}{x}\left(x\,dec\left(\frac{1}{\xi}\right)\right) = \frac{1}{x}\cdot$

Les définitions des inverses dans $\mathbb{Q}$ et $\mathbb{R}$ sont donc compatibles. ∎

Fin de la construction de $\mathbb{R}$ par les écritures décimales

En *identifiant* $\mathbb{Q}$ à son image dans $\mathbb{R}$ par l'injection naturelle dec, on obtient le théorème suivant qui termine notre construction de $\mathbb{R}$.

Théorème 6. L'ensemble $\mathbb{R}$ muni de la structure de corps totalement ordonné définie ci-dessus satisfait les axiomes des nombres réels.

Remarques.

1. Dans toutes les constructions « classiques » de $\mathbb{R}$ (décimaux, coupures ou sections commençantes, suites de Cauchy de rationnels), on n'obtient pas en fait un corps *contenant* $\mathbb{Q}$ comme sous-corps. Il est nécessaire *d'identifier* $\mathbb{Q}$ à un sous-corps de $\mathbb{R}$ pour vérifier les axiomes des nombres réels. Ceci ne pose pas de problème grâce au corollaire 66 de la page 491 qui sera établi plus loin.

2. Au lieu de construire les nombres réels avec des écritures décimales, on peut utiliser des écritures en base b ($b \in \mathbb{N}$, $b \geqslant 2$). Les *chiffres* sont les éléments de $[\![0, b-1]\!]$. Les définitions et les preuves sont similaires, *mutatis mutandis*. Ainsi chaque choix de b permet de construite un corps $\mathcal{R}_b$ satisfaisant les axiomes des nombres réels (modulo l'identification à $\mathbb{Q}$ du sous-corps premier). Les résultats sont différents *mais* isomorphes à isomorphisme unique près.

1.3 Écritures décimales à gauche Nombres décadiques

Ce paragraphe et la partie 2 qui suivra forment un « long entracte » dans l'étude des constructions de $\mathbb{R}$. Nous reprendrons celles-ci dans la partie 3.

Nous présentons deux premières approches de « nouveaux nombres », les nombres p-adiques[10] (p-étant un nombre premier). Il y a trois présentations « classiques ».

1. Une présentation *algorithmique* par les écritures en base p illimitées à gauche. C'est la plus élémentaire (elle est essentiellement du niveau de la classe de terminale du lycée) et nous commencerons par elle (en choisissant au départ la numération à base 10, plutôt que p).
2. Une présentation *algébrique* par les *limites projectives* d'anneaux $\mathbb{Z}/p^n\mathbb{Z}$. C'est une reformulation « élégante » de la précédente.
3. Une présentation *métrique et topologique* par *complétion* du corps des rationnels $\mathbb{Q}$ muni d'une valeur absolue p-adique $|\ |_p$. Nous la décrirons dans la partie 4.1.

Il y a pour $\mathbb{R}$ des présentations très similaires à 1 et 3. Pour 1 c'est la description de $\mathbb{R}$ par écritures décimales (à droite), ou plus généralement en base b, que nous venons d'étudier en détails. Pour 3, nous la donnerons dans la partie 4.1. Dans le cas des réels, il n'y a pas d'analogue de 2 (ceci est lié à l'existence des développements impropres). Par contre il y a dans ce cas une autre présentation $2'$: par « complétion » de $\mathbb{Q}$ pour la relation d'ordre total, suivant la méthode de Dedekind que nous décrirons dans la partie 3. Il n'existe rien de tel pour les corps p-adiques qui ne sont pas totalement ordonnables.

Il y a une variante de la construction de $\mathbb{R}$ par les écritures décimales qui donne d'autres nombres que les nombres réels. Au lieu d'utiliser des écritures illimitées *à droite*, on peut utiliser des écritures illimitées *à gauche*. On obtient ainsi les nombres dits *décadiques* (ou 10-*adiques*). On autorise un nombre *fini* de chiffres après la virgule. S'il n'y en a pas, on obtient un *entier décadique*. On étend les opérations d'addition et de multiplication sur les entiers naturels aux nombres décadiques. Cette extension est *beaucoup plus simple* que dans le cas des représentations décimales classiques[11] : les algorithmes d'addition et de multiplication des décimaux en écriture décimale, utilisant les retenues, s'étendent sans problème en s'autorisant « à les poursuivre indéfiniment ». On obtient ainsi un anneau (nous le montrerons plus loin), *l'anneau* E_{10} des nombres décadiques. Nous verrons qu'il n'est pas intègre. On peut munir cet anneau d'une métrique "naturelle", pour laquelle l'ensemble $\mathbb{Z}_{10}$

[10] Et plus généralement b-adiques, b étant un entier supérieur ou égal à 2, une base de numération.

[11] La somme des écritures à gauche est effective. Celle des réels ne l'est pas, *cf.* la remarque de la page 437.

des entiers décadiques est *compact*. Pour cette métrique 10^m est « petit » quand $m \in \mathbb{N}$ est « grand ». C'est le contraire de ce qui se passe avec la métrique usuelle, pour laquelle 10^{-m} est « petit » quand $m \in \mathbb{N}$ est « grand ». Plus généralement un nombre entier décadique dont les m premiers chiffres (en ordonnant les chiffres vers la gauche) sont nuls est « petit » dans les mêmes conditions. C'est ce qui permet de faire disparaître légitimement les retenues sous le tapis par glissement vers la gauche. Par contre la relation d'ordre usuelle sur $\mathbb{N}$ ne se prolonge pas à $\mathbb{Z}_{10}$.
On a des constructions similaires avec une base de numération b arbitraire. *Mais,* contrairement à ce que nous avons vu pour les écritures décimales à droite (*cf.* la remarque 2 de la page 438), nous verrons que le résultat dépend (beaucoup...) de b : les anneaux de nombres b-adiques *ne sont pas isomorphes entre eux*.
Si la base est un nombre premier p (ou une puissance d'un tel nombre), les anneaux $\mathbb{Z}_p$ de nombres entiers p-adiques sont *intégres*. On note $\mathbb{Q}_p$ le corps des fractions de $\mathbb{Z}_p$, c'est le *corps des nombres p-adiques* (on a $E_p = \mathbb{Q}_p$). Comme pour le cas des réels, il y a d'autres constructions de $\mathbb{Q}_p$. Par exemple, la méthode de complétion que nous présenterons dans la partie 4.1, à partir des suites de Cauchy de $\mathbb{Q}$, permet de construire $\mathbb{R}$ *et* les corps $\mathbb{Q}_p$. Nous verrons que, comme dans le cas de $\mathbb{R}$ toutes les constructions donnent des résultats isomorphes quand p est fixé.

1.3.1 Définitions.

Comme nous l'avons fait dans la construction de $\mathbb{R}$, nous identifions un décimal (positif) $\delta \in \mathcal{D}^+$ à sa représentation décimale : $r = \alpha_n\alpha_{n-1}\dots\alpha_0,\beta$, où $\beta = \alpha_{-1}\dots\alpha_{-r}$ est une suite *finie*, de longueur arbitraire, de chiffres, éventuellement vide. On suppose $\alpha_n \neq 0$, mais par abus de notation nous nous affranchirons parfois de cette condition.

Définition 3. On appelle écriture décimale *à gauche* toute « suite » de la forme $\dots\alpha_n\alpha_{n-1}\dots\alpha_0,\beta$, où $\beta = \alpha_{-1}\dots\alpha_{-r}$ est une suite *finie*, de longueur arbitraire, de chiffres, éventuellement vide, et $(\alpha_i \in [\![0,9]\!])_{i\in\mathbb{N}}$ une suite infinie de chiffres.
Si β est vide, on dit que l'on a une écriture décimale à gauche *entière* (on supprime alors la virgule).

La *partie entière* de $\dots\alpha_n\alpha_{n-1}\dots\alpha_0,\beta$ est, par définition, $\dots\alpha_n\alpha_{n-1}\dots\alpha_0$.
Par définition, on obtient l'écriture décimale $\mathrm{decg}(\delta)$, ou développement décimal à gauche, d'un nombre décimal *positif* $\delta \in \mathcal{D}_+$ en complétant la représentation décimale usuelle avec un nombre *infini* de zéros *à gauche*. Ceci permet d'identifier l'ensemble $\mathcal{D}_+$ des décimaux positifs à un sous-ensemble de l'ensemble des écritures décimales à gauche. On identifie $\dots 0\dots 0,\beta$ au décimal dont l'écriture décimale classique est $0,\beta$. On note E_{10} l'ensemble des écritures décimales à gauche. On note $\mathbb{Z}_{10}$ l'ensemble des écritures décimales à gauche entières. On identifie un entier naturel à son écriture décimale. On identifie $\mathcal{D}_+$ à un sous-ensemble de E_{10}. Alors $\mathbb{Z}_{10}$ contient $\mathbb{N} \subset \mathcal{D}_+$.

Remarque. Si l'on compare la définition ci-dessus avec la définition 2 de la page 431, on constate que le *signe* a disparu. Nous verrons plus loin que le passage à des écritures *infinies* le rend inutile : en un certain sens $\mathbb{Z}$ est déjà dans $\mathbb{Z}_{10}$. Le prix à payer pour cette « magie » est la perte de la relation d'ordre...

Définition 4. L'ensemble des écritures décimales à gauche est aussi appelé ensemble des nombres décadiques. On le note E_{10}. L'ensemble des écritures décimales à gauche entiéres est aussi appelé ensemble des nombres entiers décadiques. On le note Z_{10}.

L'ensemble $\mathbb{Z}_{10}$ s'identifie à $[\![1,9]\!]^{\mathbb{N}}$. C'est un produit dénombrable d'ensembles finis de cardinal supérieur ou égal à 2. Il a donc *la puissance du continu*. Par ailleurs l'ensemble E_{10} des écritures décimales à gauche s'identifie à un sous-ensemble de $[\![1,9]\!]^{\mathbb{Z}}$ qui a la puissance du continu. Puisque $Z_{10} \subset E_{10}$, l'ensemble E_{10} a *la puissance du continu*.
Pour les écritures décimales, nous avions introduit la notion de *troncature* (à droite). Pour les écritures décimales à gauche on définit de façon analogue une notion de *troncature à gauche*. Dans un cas, l'on ne gardait que k chiffres *après* la virgule, dans l'autre on ne gardera que k chiffres *avant* la virgule.

Définition 5. Pour tout $k \in \mathbb{N}^*$, on appelle *troncature à gauche* à l'ordre k d'une écriture décimale à gauche $x := \ldots \alpha_n \alpha_{n-1} \ldots \alpha_0, \beta$ le *décimal* $x_k := \alpha_{k-1} \ldots \alpha_0, \beta$.

Autrement dit, on ne garde que k chiffres avant la virgule. On convient de définir la troncature à gauche à l'ordre 0 par $x_0 = 0, \beta$. Dans certains cas, nous interpréterons la troncature x_k comme une écriture finie $\alpha_{k-1} \ldots \alpha_0, \beta$ plutôt que comme un décimal.
Si $x := \ldots \alpha_{n-1} \ldots \alpha_0, \beta$, on a (en utilisant les opérations de $\mathcal{D}$) :

$$x_k = \alpha_{k-1} \ldots \alpha_0, \beta = \sum_{i=0}^{k-1} 10^i a_i + 0, \beta.$$

On note $x_k = \operatorname{trg}_k x$. Si $k' \geqslant k$, on vérifie $\operatorname{trg}_k \circ \operatorname{trg}_{k'} = \operatorname{trg}_k$. En particulier, si l'on tronque à l'ordre k la troncature de $x \in E_{10}$ à l'ordre $k+1$, on obtient la troncature à l'ordre k. On dit que la suite $(\operatorname{trg}_k x)_{k \in \mathbb{N}}$ est *une suite cohérente* de décimaux.
Plus généralement, une suite $(y_k)_{k \in \mathbb{N}^*}$ de décimaux est dite *cohérente* si, pour tout $k \in \mathbb{N}^*$, la représentation décimale de y_k a au plus k chiffres avant la virgule et si $\operatorname{trg}_k y_{k+1} = y_k$.

> ***Remarque***. A toute écriture à gauche $x \in E_{10}$, on peut associer la suite cohérente de ses troncatures à gauche (x_k). On vérifie que l'on obtient ainsi une bijection de E_{10} sur l'ensemble des suites cohérentes de décimaux. Nous utiliserons cette bijection de façon essentielle plus loin. Nous verrons qu'une suite cohérente de décimaux (ou de leurs représentations décimales) peut être interprétée de façon *algébrique* (un élément d'une limite projective d'ensembles) ou *topologique* (une suite convergente). Ceci conduit à deux autres définitions de E_{10} que nous expliciterons dans les parties 2.1 et 4.1.1.

La troncature trg_k induit une application de $\mathcal{D}_+$ dans $\mathcal{D}_+$, toujours notée trg_k par abus de langage. Si, pour tout $i \geqslant k$, $\alpha_i = 0$, alors $\operatorname{trg}_k x = x$.

Le résultat suivant, bien que presque trivial, est utile pour la suite. Il permet de justifier l'addition et la multiplication des écritures décimales à gauche par les méthodes élémentaires avec retenues, en partant de la droite

Lemme 7.

(i) Soient $k \in \mathbb{N}$ et $x \in \mathbb{N}$. Il existe $y \in \mathbb{N}$ (unique) tel que $x = \operatorname{trg}_k x + 10^k y$.

(ii) Soit $k \in \mathbb{N}$. Soient $u, v \in \mathbb{N}$. On a $\operatorname{trg}_k (u + 10^k v) = \operatorname{trg}_k u$.

(iii) Pour tout $k \in \mathbb{N}$ et tous $x, x' \in \mathbb{N}$, on a :

$$\operatorname{trg}_k (x + x') = \operatorname{trg}_k (\operatorname{trg}_k x + \operatorname{trg}_k x') \quad \text{et} \quad \operatorname{trg}_k (xx') = \operatorname{trg}_k \big((\operatorname{trg}_k x)(\operatorname{trg}_k x')\big).$$

Démonstration. Pour les entiers naturels, tronquer à gauche à l'ordre k « revient à calculer modulo 10^k ». Le lemme s'en déduit. ∎

Par suite, si x et x' sont des nombres entiers naturels, les k premiers chiffres de l'écriture décimale de $x + x'$ (resp. xx') ne dépendent que des k premiers chiffres de l'écriture décimale de x et x'.

À toute écriture décimale à gauche entière $x = \ldots \alpha_n \ldots \alpha_1 \ldots \alpha_0$ on associe la série *formelle* $\sum_{n \in \mathbb{N}} \alpha_n 10^n$, dite *série de Hensel* de x. On étend de façon évidente aux écritures décimales à gauche générales : $\sum_{n \geqslant n_0} \alpha_n 10^n$, avec $n_0 \in \mathbb{Z}$, pour tenir compte de la partie $0, \beta$. On prendra garde au fait que les séries de Hensel sont, comme les séries entières classiques, des « écritures à droite ».

Si l'écriture est celle d'un décimal $\delta \in \mathcal{D}_+$, la série n'a qu'un nombre fini de termes non nuls et la somme de ceux-ci redonne δ.

Par analogie avec l'ordre d'une série formelle[12] (*cf.* [L2, I.1, 4]), on définit *l'ordre* $\omega(x)$ d'une écriture décimale à gauche entière $x = \ldots \alpha_{n-1} \ldots \alpha_0$, non nulle, comme l'unique entier $n \in \mathbb{N}$ tel que $\alpha_n \neq 0$ et $\alpha_k = 0$, pour tout $k < n$. On convient de $\omega(0) = +\infty$. On étend de façon évidente cette définition aux écriture décimales à gauche quelconques. L'ordre d'une écriture est un élément de $\mathbb{Z} \cup \{+\infty\}$.

1.3.2 Sommes et produits d'écritures décimales à gauche

Nous allons voir que, pour tout $k \in \mathbb{Z}$, on peut facilement définir une multiplication des écritures décimales à gauche par 10^k, en généralisant le cas de $\mathcal{D}_+$. Pour les décimaux positifs écrits en écriture décimale, la multiplication par 10^k, avec $k \in \mathbb{N}^*$ se traduit par le décalage des chiffres de k crans vers la gauche. La multiplication par 10^k, avec $k \in -\mathbb{N}^*$ se traduit par le décalage des chiffres de k crans vers la droite. Dans les deux cas on peut « traverser la virgule ». La multiplication par $10^0 = 1$ est évidemment l'identité. On peut étendre de façon évidente cette définition aux écritures décimales à gauche.

Remarques.

1. Une différence essentielle entre les écritures à droite et les écritures à gauche est qu'il n'y a pas d'écriture à gauche impropre. Ceci élimine, quand on travaille à gauche, beaucoup de difficultés techniques. Mais cette facilité « se paye » d'une certaine façon, par la disparition d'un ordre total compatible avec les opérations.

[12] *Attention.* Pour les séries formelles on emploie aussi le terme « valuation » au lieu « d'ordre ». Il ne faut pas le faire pour les décadiques car l'ordre n'est pas une valuation. Il l'est pour les b-adiques, que nous définirons plus loin, quand b est un nombre premier (ou une puissance d'un nombre premier).

2. On peut utiliser une autre base b de numération. On définit alors de façon analogue des nombres b-adiques : ce sont les écritures à gauche en base b. On les note E_b et on note $\mathbb{Z}_b \subset E_b$ les écritures *entières*. Nous verrons que, *contrairement* à ce qui se passe pour les écritures classiques (à droite), les propriétés de ces nombres « dépendent de b ». En base b on note l'ordre ω_b.

   Pour des raisons qui apparaîtront plus loin, si $b = p$ est un nombre premier, alors E_b a « de meilleures propriétés » et on le note $\mathbb{Q}_p$.

Nous allons voir que la manipulation des écritures infinies à gauche est *beaucoup plus simple* que celle des écritures infinies à droite. Les règles de calcul apprises à l'école élémentaire pour les écritures décimales se généralisent en ce qui concerne l'addition, la multiplication et la soustraction. (Il y a par contre des problèmes avec la division.) On gagne même quelque chose en acceptant une *infinité de chiffres*, car il n'est plus nécessaire d'introduire des entiers *négatifs*, ils sont déjà dans $\mathbb{Z}_{10}$ comme nous allons le voir !

On additionne et multiplie les nombres décadiques en utilisant les algorithmes classiques de l'école élémentaire (en partant de la droite) avec report des retenues. On montre que, dans les deux cas la méthode donne une écriture infinie à gauche bien définie (les retenue éventuelles « disparaissent à l'infini à gauche) en se ramenant au cas où les nombres sont entiers (par mutiplication par des puissances de 10 convenables) et en utilisant le lemme 7.

Exemples.

1. On a : $\ldots 9999999994 + \ldots 0000000006 = \ldots 0000000000 = 0$. La retenue (qui est toujours 1) disparaît à l'infini vers la gauche ! On peut ainsi *définir* $-6 = \ldots 999994$.
2. On a : $\ldots 99999901 \times \ldots 01010101 = \ldots 00000001$ et $\ldots 00099 + \ldots 99901 = 0$. On peut ainsi *définir* $-1/99 = \ldots 01010101$. On vérifie :

$$(99 \times \ldots 01010101) + 1 = \ldots 99999999 + 1 = 0.$$

Du point de vue des séries de Hensel, si l'on somme, suivant la règle habituelle, deux séries formelles :

$$\sum_{n\in\mathbb{N}^*} \alpha_n 10^n \quad \text{et} \quad \sum_{n\in\mathbb{N}^*} \beta_n 10^n,$$

on obtient $\sum\limits_{n\in\mathbb{N}^*} (\alpha_n + \beta_n)10^n$ qui *n'est pas en général* une série de Hensel car l'entier $\alpha_n + \beta_n$ peut ne pas appartenir à $[\![0, 9]\!]$. La régle des retenues permet de transformer cette série en une série de Hensel.

L'ensemble des nombres décadiques peut s'écrire $E_{10} = \bigcup\limits_{k\in\mathbb{N}} 10^{-k} Z_{10}$.

Exercice 4.

Montrer que pour tout nombre entier décadique a il existe un unique nombre décadique c tel que $a + c = 0$. Généraliser aux nombres décadiques quelconques. Quel est l'analogue en base b ?

Solution. Soit $a = \ldots \alpha_n \alpha_{n-1} \ldots \alpha_1 \ldots \alpha_0$. On suppose d'abord que $\alpha_0 \neq 0$. On définit alors $c = \ldots \gamma_n \gamma_{n-1} \ldots \gamma_1 \ldots \gamma_0$ par $\gamma_0 := 10 - \alpha_0$ et $\gamma_k := 9 - \alpha_k$ pour $k \geqslant 1$. L'unicité est immédiate.

Par exemple :

$$a = \ldots 6237012, c = \ldots 7362988 \ ; \ a = \ldots 123456000, c = \ldots 876544000.$$

En particulier, si $a = 1$, alors $-1 := c = \ldots 99999999$.
Si $a = 0$, $c := 0$. Supposons $a \neq 0$. On l'écrit (de façon unique) $a = 10^{-k}a'$, où a' est un entier décadique dont le premier chiffre est non nul. Il existe c' tel que $a' + c' = 0$ et l'on pose $c := 10^{-k}c'$. On vérifie $a + c = 0$.
Si l'on remplace 10 par b, on remplace 9 par $b - 1$ et l'on définit c par $\gamma_0 := b - \alpha_0$ et $\gamma_k := b - 1 - \alpha_k$ pour $k \geqslant 1$.

Si $a + c = 0$, on note $c = -a$. Si $a \in \mathbb{N}$, on identifie $-a \in \mathbb{Z}_{10}$ à $-a \in \mathbb{Z}$. On identifie ainsi $\mathbb{Z}$ à un sous-ensemble de $\mathbb{Z}_{10}$ L'inclusion $\mathbb{Z} \subset \mathbb{Z}_{10}$ est stricte, puisque $\mathbb{Z}$ est dénombrable et que $\mathbb{Z}_{10}$ ne l'est pas.
Le lecteur vérifiera que l'on peut calculer $x + (-x')$ en calculant la différence $x - x'$ avec la règle élémentaire avec report des retenues.
Passons à la généralisation de la division. Elle pose un certain nombre de problèmes. Nous allons faire quelques remarques simples, nous retournerons à la question quand nous disposerons de plus d'outils. Nous verrons que la division est plus simple si b est un *nombre premier*.
Dans un premier temps, nous allons essayer de définir la division d'un nombre décadique par un entier naturel. La première étape est d'essayer de diviser 1 par $n \in \mathbb{N}^*$. On sait que dans $\mathbb{Z}$ cela n'est possible que si $n = 1$. Nous allons voir que, par contre, certains entiers de $\mathbb{N}^*$ deviennent inversibles dans $\mathbb{Z}_{10}$.
On peut calculer facilement $1/3 = \ldots 6666666\underline{6}7$ ou $1/7 = \ldots 142857142857\underline{142857}143$, en s'inspirant de l'algorithme de division élémentaire. Par exemple :

$$\frac{1}{3} = 10 \times \left(-\frac{2}{3}\right) + 7, \ \text{ puis } \ -\frac{2}{3} = 10 \times \left(-\frac{2}{3}\right) + 6,$$

puis la même chose une infinité de fois. Dans les deux cas on recopie une infinité de fois à gauche le bloc de chiffre souligné. Le lecteur sceptique vérifiera, en utilisant « la règle de multiplication des petits enfants », que :

$$(\ldots 6666666\underline{6}7) \times 3 = 1 \quad \text{et} \quad (\ldots 142857142857\underline{142857}143) \times 7 = 1.$$

Par contre, on ne peut pas calculer $1/2$ dans $\mathbb{Z}_{10}$ car il n'existe pas de chiffre $\alpha \in [\![0, 9]\!]$ tel que 2α se termine par 1. Les choses s'arrangent si l'on accepte des chiffres après la virgule, c'est-à-dire dans E_{10}. En effet l'ensemble $\mathcal{D}_+$ des décimaux positifs s'identifie à un sous-ensemble de E_{10} et l'on a $1/2 = 0,5$, $1/5 = 0,2$. Pour diviser par 2 (resp. 5) dans E_{10}, on divise par 10 (i.e. on translate l'écriture d'un cran vers la droite), puis l'on multiplie par 5 (resp. 2). Par suite on peut diviser par $2^m 5^r$ $(m, r \in \mathbb{N}^*)$.
Si la base de numération est *un nombre premier* $b = p$, on peut donner un algorithme simple pour le calcul de l'inverse de tout élément de Z_p *dont la première décimale est non nulle* (*cf.* l'exercice III.3.7 de la page 522). Cet inverse est dans $\mathbb{Z}_p$. On donnera plus loin une autre preuve de l'existence de l'inverse dans $\mathbb{Z}_p$.

Remarques.

1. Avec les écritures décimales à droite, on a $1/3 = 0,33333\ldots$. On peut ainsi deviner qu'il n'est pas raisonnable d'utiliser des écritures infinies à la fois à droite *et* à gauche.
2. Nous verrons plus loin qu'un élément de $\mathbb{Z}_{10}$ est inversible dans $\mathbb{Z}_{10}$ si, et seulement si, son premier chiffre est 1, 3, 7, ou 9. On peut diviser dans $\mathbb{Z}_{10}$ par un nombre premier différent de 2 et 5. Par suite on peut diviser par un nombre entier $n \in \mathbb{N}^*$ qui n'est divisible ni par 2 ni par 5. On en déduit que l'on peut diviser dans E_{10} par tout $n \in \mathbb{N}^*$. On peut alors identifier $\mathbb{Q}$ à un sous-ensemble de E_{10}. L'addition et la multiplication de E_{10} induisent sur $\mathbb{Q}$ les opérations usuelles.
3. Nous verrons plus loin qu'il y a dans $\mathbb{Z}_{10}$ des diviseurs de zéro. Par contre ce problème disparaît si l'on utilise comme base de numération un nombre premier p (ou une puissance d'un tel nombre).

2 Introduction aux nombres p-adiques

On peut vérifier « artisanalement » les propriétés de l'addition et de la multiplication pour les nombres décadiques, définies dans la partie précédente, comme nous l'avons fait pour la construction de $\mathbb{R}$ par les écriture décimales, mais c'est assez fastidieux. Nous allons utiliser un argument plus abstrait, mais plus élégant, en reformulant les définitions. Cet argument est indépendant du choix de la base. Nous nous placerons donc dans le cadre des nombres b-adiques ($b \in \mathbb{N}$, $b \geqslant 2$). Certaines propriétés ne sont vraies que pour b premier. Dans ce cas nous noterons le plus souvent $b = p$, conformément à la tradition. La présentation ci-dessous est traditionnellement appelée *définition algébrique des nombres p-adiques*.

2.1 Anneaux b-adiques — Corps p-adiques Présentation algébrique

Dans toute cette partie $b \in \mathbb{N}$, $b \geqslant 2$, et p est un nombre premier. Pour tout $n \in \mathbb{N}^*$, on pose $A_n = \mathbb{Z}/b^n\mathbb{Z}$. Pour $n \geqslant 2$ l'on a $b^n\mathbb{Z} \subset b^{n-1}\mathbb{Z}$. Un élément de *l'anneau* A_n définit donc un élément de A_{n-1} et l'on obtient un homomorphisme surjectif d'anneaux $\varphi_n : A_n \to A_{n-1}$ dont le noyau est l'idéal $b^{n-1}A_n$ de A_n.
La suite :

$$\cdots A_n \xrightarrow{\varphi_n} A_{n-1} \xrightarrow{\varphi_{n-1}} \ldots \xrightarrow{\varphi_3} A_2 \xrightarrow{\varphi_2} A_1$$

forme ce que l'on appelle un *système projectif d'anneaux* indexé par l'ensemble ordonné $\mathbb{N}^*$.
On va définir la *limite projective* A de ce système. On la notera $A = \varprojlim A_n$. Par définition, un élément de *l'ensemble* A est une suite $x = (\ldots, x_n, \ldots, x_1)$, avec :

$$\forall n \in \mathbb{N},\ \ x_n \in A_n \quad \text{et} \quad \varphi_n(x_n) = x_{n-1} \ \text{ si } \ n \geqslant 2. \tag{2}$$

Une telle suite, est dite *cohérente*. C'est un élément du produit dénombrable d'ensembles $\prod_{n\in\mathbb{N}^*} A_n$. Ainsi $A \subset \prod_{n\in\mathbb{N}^*} A_n$.

On définit une addition et une multiplication sur A « coordonnées par coordonnées ». En utilisant le fait que les φ_n sont des homomorphismes d'anneaux, on vérifie que si les suites x et y sont cohérentes, il en est de même des suites :

$$\ldots(x_n+y_n)\ldots(x_1+y_1) \quad \text{et} \quad \ldots(x_ny_n)\ldots(x_1y_1).$$

Ainsi A muni de l'addition et de la multiplication est un sous-anneau du produit dénombrable d'anneaux $\prod_{n\in\mathbb{N}^*} A_n$. En particulier c'est un anneau. Il est commutatif et on vérifie qu'il est de caractéristique nulle. On a, pour tout $n \in \mathbb{N}$, une application $\varepsilon_n : A \to A_n$, définie par $\varepsilon_n(x) = x_n$. On l'appelle *réduction modulo* b^n. On vérifie que c'est un homomorphisme surjectif d'anneaux. Si $x, y \in A$ et $\varepsilon_n(x) = \varepsilon_n(y)$, on note $x \equiv y \mod b^n$.
Nous allons maintenant définir un isomorphisme naturel d'ensembles $i : \mathbb{Z}_b \to A$ et montrer que c'est un homomorphisme pour les lois d'addition et de multiplication sur $\mathbb{Z}_b$ et A. La définition de i est basée sur la remarque de la page 441. Pour un certain nombre de vérifications faciles, que nous ne détaillerons pas, on utilise le lemme 7. Pour tout $n \in \mathbb{N}^*$, on note $B_n = \{\alpha_{n-1}\ldots\alpha_0 \mid \alpha_i \in [\![1, b-1]\!]\}$ l'ensemble des écritures à n-chiffres en base b. Il s'identifie à $[\![1, b-1]\!]^n$. On le munit d'une addition et d'une multiplication en utilisant les règles élémentaires et en « oubliant » la dernière retenue si elle existe. On a une application $B_n \to \mathbb{N}$ définie par $\alpha_{n-1}\ldots\alpha_0 \mapsto \sum_{i=0}^{n-1} \alpha_i b^i$. On en déduit par passage au quotient une application $i_n : B_n \to A_n$. On vérifie qu'elle est bijective et que c'est un homomorphisme pour l'addition et la multiplication. Par suite B_n est un anneau commutatif, puisque A_n en est un. On définit, pour tout $n \geqslant 2$, une application $\psi_n : B_n \to B_{n-1}$ par troncation de la dernière décimale à gauche : $\psi_n : \alpha_{n-1}\ldots\alpha_0 \mapsto \alpha_{n-2}\ldots\alpha_0$. On vérifie : $\varphi_n \circ i_n = i_{n-1} \circ \psi_n$. On en déduit que les applications i_n ($n \in \mathbb{N}^*$) induisent un isomorphisme d'anneaux $i : B = \varprojlim B_n \to A = \varprojlim A_n$. L'ensemble $\mathbb{Z}_b$ s'identifie à B (on associe à une écriture à gauche entière la suite de ses troncatures à gauche) et les lois d'addition et de multiplication que nous avons définies plus haut sur $\mathbb{Z}_b$ s'identifient à celles de B. Ces lois munissent donc $\mathbb{Z}_b$ d'une structure d'anneau et cet anneau est isomorphe à $A = \varprojlim \mathbb{Z}/b^n\mathbb{Z}$. Dans la suite nous identifierons le plus souvent $\mathbb{Z}_b$ et A. L'application $\varepsilon_n : A \to A_n$ est alors identifiée à la réduction modulo b^n : $\varepsilon_n : \mathbb{Z}_p \to \mathbb{Z}/b^n\mathbb{Z}$. On a en particulier une application de réduction modulo b : $\varepsilon_1 : \mathbb{Z}_b \to \mathbb{Z}/b\mathbb{Z}$.
Le détour abstrait par les limites projectives nous a ainsi permis d'établir le résultat suivant.

Proposition 8. Les lois d'addition et de multiplication définies sur l'ensemble $\mathbb{Z}_b$ des écritures en base b à gauche entières, en utilisant les algorithmes élémentaires (avec retenues), munissent $\mathbb{Z}_b$ d'une structure d'anneau commutatif unitaire. L'anneau $\mathbb{Z}$ s'identifie à un sous-anneau de $\mathbb{Z}_b$.

On peut écrire un élément x de $\mathbb{Z}_b$, de façon unique, comme une série, pour l'instant *formelle*, $x = \sum_{n\in\mathbb{N}^*} \alpha_n b^n$. Cette écriture est dite *série de Hensel*. On a (avec un abus de langage) $E_b = \mathbb{Z}_b[b^{-1}]$. On a donc, plus généralement, pour $x \in E_b$ une *série de Hensel* : $x = \sum_{n\geqslant n_0} \alpha_n b^n$, avec $n_0 \in \mathbb{Z}_-$ convenable. En utilisant des multiplications par des puissances de b, on prouve facilement le résultat suivant.

Corollaire 9. Les lois d'addition et de multiplication définies sur l'ensemble E_b des écritures en base b à gauche, en utilisant les algorithmes élémentaires (avec retenues), munissent E_b d'une structure d'anneau commutatif unitaire ; $\mathbb{Z}_b$ est un sous-anneau de E_b.

Supposons b *premier*. En utilisant la bijection ensembliste entre $[\![0, b-1]\!]$ et F_b et en remplaçant p par X dans l'écriture de Hensel, on obtient un isomorphisme *ensembliste* entre $\mathbb{Z}_b$ et l'ensemble des séries formelles $F_b[[X]]$. *Ce n'est pas un isomorphisme d'anneaux* : $\mathbb{Z}_b$ est de caractéristique nulle et $F_b[[X]]$ est de caractéristique p (le problème vient des retenues dans l'addition pour $\mathbb{Z}_b$). Les deux anneaux *se ressemblent*, mais il faut utiliser cette ressemblance avec précautions.

Exercice 5.
Soit $x = \sum_{n\in\mathbb{N}^*} \alpha_n 10^n \in \mathbb{Z}_{10}$, on lui associe la série $\sum_{n\in\mathbb{N}^*} \alpha_n 10^{-n}$, que l'on peut considérer comme une série de nombres décimaux. Montrer que celle-ci *converge dans* $\mathbb{R}$ vers un nombre réel appartenant au segment $[0,1]$. On définit ainsi une application $f : \mathbb{Z}_{10} \to [0,1]$. Montrer que cette application est surjective. Est-ce qu'elle est injective ?
Solution.
On majore la série dans $\mathbb{R}$ par la série numérique $\sum_{n\in\mathbb{N}^*} 9 \times 10^{-n}$ qui converge vers 1. Tout réel $t \in [0,1[$ a une (unique) écriture propre $0,\alpha_1\alpha_2\ldots$ On vérifie que $t = f(x)$ avec $x = \sum_{n\in\mathbb{N}^*} \alpha_n 10^n \in \mathbb{Z}_{10}$. Si $t = 1$, alors $t = f(x)$, avec $x = \ldots 999999$
Ainsi f est surjective. Elle *n'est pas injective*, car si $0,\alpha_1\alpha_2\ldots$ et $0,\alpha'_1\alpha'_2\ldots$ sont les écritures propres et impropres d'un même *décimal* $t \in [0,1[$:

$$x := \sum_{n\in\mathbb{N}^*} \alpha_n 10^n \quad \text{et} \quad x' := \sum_{n\in\mathbb{N}^*} \alpha'_n 10^n$$

sont deux éléments distincts de $\mathbb{Z}_{10}$ et l'on a $f(x) = f(x') = t$.

Théorème 10. Soient $b \in \mathbb{N}$, $b \geqslant 2$ et p un nombre premier. On note $A_n := \mathbb{Z}/b^n\mathbb{Z}$ et $\mathbb{Z}_b \xrightarrow{\varepsilon_n} A_n$ la surjection canonique (nn identifie $\mathbb{Z}_b$ et la limite projective A des A_n).

(i) Pour tout $m \in \mathbb{N}^*$, la suite $0 \to \mathbb{Z}_b \xrightarrow{b^m} \mathbb{Z}_b \xrightarrow{\varepsilon_m} A_m \to 0$ est *exacte*.

(ii) Un élément x de $\mathbb{Z}_b$ est inversible si, et seulement si, son image $\varepsilon_1(x) \in A_1 = \mathbb{Z}/b\mathbb{Z}$ est inversible dans $\mathbb{Z}/b\mathbb{Z}$. Les éléments inversibles de $\mathbb{Z}_p$ forment un groupe multiplicatif

(iii) Un élément de $\mathbb{Z}_p$ (resp. $A_n\mathbb{Z}/p^n\mathbb{Z}$) est inversible si, et seulement s'il n'est pas divisible par p. Les éléments inversibles de $\mathbb{Z}_p$ forment un groupe multiplicatif $\mathbb{U}$. Tout élément non nul de $\mathbb{Z}_p$ s'écrit de façon unique $p^m u$, avec $m \in \mathbb{N}$ et $u \in \mathbb{U}$.

(iv) L'anneau $\mathbb{Z}_p$ est *principal*. Ses idéaux non triviaux sont de la forme $p^m\mathbb{Z}_p$, $m \in \mathbb{N}^*$. Il existe un *unique idéal maximal* $p\mathbb{Z}_p$. Le quotient $\mathbb{Z}_p/pZ_p$ est isomorphe au corps fini $F_p = \mathbb{Z}/p\mathbb{Z}$.

On dit que $\mathbb{U}$ est le *groupe des unités* de $\mathbb{Z}_p$ et que F_p est *le corps résiduel* de $\mathbb{Z}_p$. On peut identifier $Z_p/p^n\mathbb{Z}_p$ à $A_n = \mathbb{Z}/p^n\mathbb{Z}$. On a $\varepsilon_1(\mathbb{U}) = F_p^*$.

Démonstration.

(i) Soit $x = (\dots, x_n, \dots, x_1) \in \mathbb{Z}_b$. Par définition $(x_n \in A_n)$ est une suite cohérente. Si $bx = 0$, on a, pour tout $n \in \mathbb{N}$, $bx_{n+1} = 0$ dans A_{n+1}. Il existe donc donc $y_n \in A_{n+1}$ tel que $x_{n+1} = b^n y_{n+1}$. En appliquant φ_{n+1}, on en déduit :

$$x_n = \varphi_{n+1}(x_{n+1}) = \varphi_{n+1}(b^n)\varphi_{n+1}(y_{n+1}) = 0.$$

Par suite $\mathbb{Z}_b \xrightarrow{b} \mathbb{Z}_b$ est *injective*. Par itération il en est de même pour la multiplication par b^n.

On peut aussi utiliser les écritures b-adiques. Sur ces écritures la multiplication par b^m se traduit par un décalage de m crans vers la gauche (en remplaçant par 0 les termes manquants à droite). L'injectivité de celui-ci est évidente.

Le noyau de ε_m contient $b^m\mathbb{Z}_p$. Inversement si $x \in \mathbb{Z}_b$ appartient à ce noyau, on a, pour tout $k \geqslant m$, $x_k \equiv 0 \mod b^m$. Il existe donc $y_{k-m} \in A_{k-m}$ unique tel que $x_k = b^m y_{k-m}$. On vérifie que les y_k définissent un élément $y = \mathbb{Z}_p$ et l'on a $x = b^m y$.

L'application ε_n est surjective (les φ_n le sont).

(ii) Si $x \in \mathbb{Z}_b$ est inversible, il existe $y \in \mathbb{Z}_b$ tel que $xy = 1$.

Alors $1 = \varepsilon_1(xy) = \varepsilon_1(x)\varepsilon_1(y)$ et $\varepsilon_1(x)$ est inversible dans $\mathbb{Z}/b\mathbb{Z}$. Inversement, soit $x \in \mathbb{Z}_b$ tel que $x_1 := \varepsilon_1(x)$ soit inversible dans $\mathbb{Z}/b\mathbb{Z}$. Montrons d'abord que, pour tout $n \in \mathbb{N}^*$, $x_n := \varepsilon_n(x)$ est inversible dans A_n. Par hypothèse, il existe un entier $m \in \mathbb{N}$ tel que $mx_n \equiv 1 \mod b$. Notons y_n l'image de m dans A_n. On a $x_n y_n \equiv 1 \mod b$ dans A_n, il existe donc $z_n \in A_n$ tel que $x_n y_n = 1 - bz_n$. Donc $x_n x'_n = 1$, avec $x'_n = y(1 + pz + \cdots + p^{n-1}z^{n-1})$. L'inverse x'_n de x_n dans A_n est *unique*. La suite $(x'_n \in A_n)_{n\in\mathbb{N}^*}$ est donc *cohérente* et elle définit un élément $x' \in \mathbb{Z}_b$. On a $xx' = 1$. Il est évident que les inversibles forment un groupe multiplicatif.

(ii) Les inversibles de $\mathbb{Z}/p\mathbb{Z} = F_p$ sont les éléments non nuls. D'après (ii), $x \in \mathbb{Z}_p$ est inversible si, et seulement si, $\varepsilon_1(x) \neq 0$, c'est-à-dire si, et seulement si, x n'est pas divisible par p. Soit $x = (x_n) \in \mathbb{Z}_p^*$. Il existe $r \in \mathbb{N}$ tel que $x_r \neq 0$. Soit m le plus grand entier tel que $x_m := \varepsilon_m(x)$ soit nul (ou $m = 0$). Alors $x = p^m u$, avec $u \in \mathbb{Z}_p$ non divisible par p. D'après ce qui précéde, on a $u \in \mathbb{U}$. L'unicité est évidente.

(iv) Pour tout $r \in \mathbb{N}^*$, $p^r\mathbb{Z}_p$ est un idéal de $\mathbb{Z}_p$. Inversement, soit $\mathcal{I} \subset \mathbb{Z}_p$ un idéal non trivial. Si $x \in \mathcal{I}$, $x = p^m u$ ($m \in \mathbb{N}$, $u \in \mathbb{U}$), donc $p^m \in \mathcal{I}$ et $p^m\mathbb{Z}_p \subset \mathcal{I}$. On en déduit que, si l'on désigne par r le plus petit entier tel que $p^r\mathbb{Z}_p \subset \mathcal{I}$, l'on a $\mathcal{I} = p^r\mathbb{Z}_p$. Ainsi $\mathbb{Z}_p$ est un anneau principal. L'idéal $p\mathbb{Z}_p$ est maximal et c'est l'unique idéal maximal. D'après (i) (avec $n = 1$), on a $\mathbb{Z}_p/p\mathbb{Z}_p = A_1 = F_p$. ■

En utilisant les écritures à gauche, on peut donner un algorithme très simple pour le calcul de l'inverse de $x \in \mathbb{Z}_p$ non divisible par p (et par suite de tout x non nul de $\mathbb{Q}_p$), *cf.* l'exercice III.3.7 de la page 522.

Exercice 6.
Soit $b \in \mathbb{N}$, $b \geqslant 2$. Montrer que $m \in \mathbb{Z}$ est inversible dans $\mathbb{Z}_b$ si, et seulement s'il est premier avec b.
Solution. L'entier m est inversible dans $\mathbb{Z}/b\mathbb{Z}$ si, et seulement s'il est premier avec b. On utilise ensuite l'assertion (ii) du théorème 10 ci-dessus.

Soit p premier. Si $x \in \mathbb{Z}_p$ n'est pas nul, on a $x = p^m u$, avec $m \in \mathbb{Z}$ et $u \in \mathbb{U}$. On dit que l'entier m est la *valuation p-adique* de x et l'on note $v_p(x) := m$. On convient de $v_p(0) := +\infty$ On a :

$$\forall x, y \in \mathbb{Z}_p, \quad v_p(xy) = v_p(x)v_p(y) \quad \text{et} \quad v_p(x+y) \geqslant \min\,(v_p(x), v_p(y)). \tag{3}$$

On vérifie que v_p coïncide avec l'ordre ω_p défini plus haut. On prendra garde au fait que, si b n'est pas un nombre premier (ou une puissance d'un tel nombre), on a seulement une *inégalité* $\omega_p(xy) \geqslant \omega_p(x)\omega_p(y)$. Par exemple :

$$1 = \omega_{10}(10 = 2 \times 5) > 0 = 0 \times 0 = \omega_{10}(2)\omega_{10}(5).$$

Dans ce cas, l'ordre n'est pas une valuation.
Soient $x, y \in \mathbb{Z}_p$. Si x et y ne sont pas nuls, on a $v_p(x) \in \mathbb{N}$ et $v_p(y) \in \mathbb{N}$, donc $v_p(xy) \in \mathbb{N}$ et xy n'est pas nul. L'anneau $\mathbb{Z}_p$ est *intègre*.

On prouve alors facilement l'énoncé suivant.

Corollaire 11. Soit p un nombre premier, l'anneau Z_p est *intègre* et l'anneau des nombres p-adiques E_p est l'anneau sous-jacent à son corps des fractions. On note $\mathbb{Q}_p$ ce corps. On a (avec un abus de notation) $\mathbb{Q}_p = Z_[p^{-1}]$. Tout élément x non nul de $\mathbb{Q}_p$ s'écrit de façon unique $x = p^m u$, avec $m \in \mathbb{Z}$ et $u \in \mathbb{U}$.

L'anneau Z_p contient un sous-anneau isomorphe à $\mathbb{Z}$. Par suite, $\mathbb{Q}_p$ contient un sous-corps isomorphe à $\mathbb{Q}$.

Supposons p premier et $m \in \mathbb{N}^*$. L'exercice suivant permet de ramener l'étude des nombres p^m-adiques à celle des nombres p-adiques.

Exercice 7.
On suppose $b = c^m$, où $c \in \mathbb{N}$, $c \geqslant 2$, et $m \in \mathbb{N}^*$. Montrer que les anneaux $\mathbb{Z}_b$ et $\mathbb{Z}_c$ sont isomorphes. Montrer que si p est un nombre premier, alors l'anneau Z_{p^m} est intègre. Montrer que l'anneau E_{p^m} est isomorphe à l'anneau sous-jacent au corps des fractions $\mathbb{Q}_p$ de $\mathbb{Z}_p$
Solution.

1. Un nombre $\beta \in [\![1, 99]\!]$ s'écrit de manière unique $\beta = 10\alpha + \alpha' = (\alpha, \alpha')$, avec $\alpha, \alpha' \in [\![0, 9]\!]$. On peut remplacer dans l'écriture d'un nombre de $\mathbb{Z}_{100}$ chaque chiffre $\beta \in [\![1, 99]\!]$ par un couple ordonné de chiffres de $[\![0, 9]\!]$. On obtient ainsi une bijection de $\mathbb{Z}_{100}$ sur $\mathbb{Z}_{10}$. On peut faire de même pour $b = c^m$. On remplace dans l'écriture d'un nombre de $\mathbb{Z}_b$ chaque chiffre $\beta \in [\![1, b-1]\!]$ par un m-uple ordonné de chiffres de $[\![0, c-1]\!]$. On obtient ainsi une *bijection* de $\mathbb{Z}_b$ sur $\mathbb{Z}_c$. On peut vérifier en utilisant

les écritures que c'est un isomorphisme d'anneaux. Il est plus simple d'utiliser les limites projectives.

On a : $\mathbb{Z}_b = A = \varprojlim A_n$ et $\mathbb{Z}_c = A' = \varprojlim A'_{n'}$, où :

$$A'_{n'} = \mathbb{Z}/c^{n'}\mathbb{Z} \quad \text{et} \quad A_n = \mathbb{Z}/b^n\mathbb{Z} = \mathbb{Z}/c^{mn}\mathbb{Z} = A'_{mn}.$$

Considérons $\varphi_n : A_n \to A_{n-1}$. On peut l'écrire $\psi_n : A'_{mn} \to A'_{m(n-1)}$ et l'on a

$$\psi_n = \varphi'_{mn} \circ \varphi'_{mn-1} \ldots \circ \varphi'_{mn-m+1}. \tag{4}$$

Les homomorphismes $\varepsilon_n : A \to A_n$ peuvent s'écrire $\theta_n : A \to A'_{mn}$. On en déduit, en utilisant (4), un homomorphisme d'anneaux

$$\mathbb{Z}_b = A = \varprojlim A_n \to \mathbb{Z}_c = A' = \varprojlim A'_{n'}.$$

On vérifie, en utilisant les écritures, que l'application $\mathbb{Z}_b \to \mathbb{Z}_c$ obtenue s'identifie à la bijection définie plus haut. On a ainsi un isomorphisme d'anneaux $\mathbb{Z}_b \to \mathbb{Z}_c$.

2. Supposons $b = p^m$, avec p premier. Puisque l'anneau $\mathbb{Z}_p$ est intègre, l'anneau isomorphe $\mathbb{Z}_{p^m}$ l'est aussi. On peut aussi vérifier que v_{p^m} est une *valuation* sur $\mathbb{Z}_{p^m}$: pour tous $x, y \in \mathbb{Z}_{p^m}$, $v_{p^m}(xy) = v_{p^m}(x)v_{p^m}(y)$, puis en déduire que $\mathbb{Z}_{p^m}$ est intégre.

3. Les corps des fractions de $\mathbb{Z}_{p^m}$ et $\mathbb{Z}_p$ sont isomorphes. On vérifie facilement :

$$E_{p^m} = \bigcup_{n\in\mathbb{N}} p^{-mn} Z_{p^m}. = \bigcup_{n'\in\mathbb{N}} p^{-n'} Z_{p^m}.$$

On en déduit que l'anneau E_{p^m} est l'anneau sous-jacent au corps des fractions de $\mathbb{Z}_{p^m}$.

Nous verrons plus loin que si $\mathbb{Z}_b$ est intègre, alors b est la puissance d'un nombre premier (*cf.* le corollaire 48 de la page 477).

2.2 Distances et topologie. Suites et séries

Dans cette partie b est un entier supérieur ou égal à 2, la base de la numération. On ne suppose pas nécessairement que b est premier.

2.2.1 Distances et topologie

L'idée implicite sous-jacente à la manipulation des écritures décimales infinies à gauche est que $\ldots\alpha_{n+1}\alpha_n 00\ldots 00$ est « petit » si n est « grand ». On formalise ceci en utilisant l'analogie des séries de Hensel avec les séries formelles : l'ordre $\omega(f)$ d'une série formelle $f \in K[[X]]$ est l'unique entier m tel que $f = X^m g$, avec $g(0) \neq 0$. On a défini plus haut *l'ordre* $\omega(x)$ d'un élément non nul $x = \ldots\alpha_n\alpha_{n-1}\ldots\alpha_0$ de $\mathbb{Z}_{10}$ comme l'entier k tel que $\alpha_k \neq 0$ et $\alpha_h = 0$ pour tout $0 \leqslant h < k$. Par convention, l'ordre de 0 est $+\infty$.

On a une notion similaire pour tout $\mathbb{Z}_b$. S'il est nécessaire de rappeler la base b de l'écriture, on écrit ω_b. On étend de façon évidente la notion d'ordre à E_b.

Si $x \in \mathbb{Z}_b$, l'ordre $\omega_b(x)$ est le plus grand entier $m \in \mathbb{N}$ tel que p^m divise x : $x \in p^{\omega_b(x)}\mathbb{Z}_b$ et $x \notin p^{\omega_b(x)+1}\mathbb{Z}_b$. Si $b = p$ est *un nombre premier*, l'ordre coïncide avec la *valuation p-adique* introduite plus haut (*cf.* la page 449) : $\omega_b = v_b$ et l'on a $x = p^{\omega_b(x)}u$, avec $u \in \mathbb{U}$.

On vérifie immédiatement que l'ordre possède les propriétés suivantes :

1. $x = 0 \Leftrightarrow \omega_b(x) = +\infty$;
2. pour tous $x, y \in E_b$, $\omega_b(xy) \geqslant \omega_b(x)\omega_b(y)$;
3. pour tous $x, y \in E_b$, $\omega_b(x+y) \geqslant \min(\omega_b(x), \omega_b(y))$.

Notons que l'on peut avoir dans 2 une inégalité stricte. Par exemple $\omega_{10=2\times 5}(10) = 1$ et $\omega_{10}(2) = \omega_{10}(5) = 1$. On a toujours une égalité si b est un nombre premier (*cf.* (3)). Pour $x \in E_b$, on *définit* $|x|_b := b^{-\omega_b(x)} \in \mathbb{Q}_+$. On convient de $b^{-\infty} = 0$.

On vérifie facilement les résultats suivants.

Proposition 12. On a :

1. $|x|_b = 0 \Leftrightarrow x = 0$;
2. pour tous $x, y \in E_b$, $|xy|_b \leqslant |x|_b|y|_b$;
3. pour tous $x, y \in E_b$, $|(|x+y)_b \leqslant \max(|x|_b, |y|_b)$.
4. L'image $|E_b^*|_b \subset \mathbb{Q}_+$ de E_b^* par l'application $|\ |_b$ est $b^{\mathbb{Z}}$.

Si $b = p$ est premier, on a :

2'. pour tous $x, y \in E_b$, $|xy|_p = |x|_p|y|_p$.

On dit alors que $|\ |_p$ est *une valeur absolue*. C'est la valeur absolue p-adique sur $\mathbb{Q}_p$. Notons que, si $\mathbb{Q}$ est muni de sa métrique ordinaire, le sous-ensemble $b^{\mathbb{Z}} \subset \mathbb{Q}$ est *discret*.

Exercice 8.

Soient $x, y \in E_b$. On suppose $|x|_b \neq |y|_b$. Alors $|x+y|_b = \max(|x|_b, |y|_b)$.

Solution. Quitte à permuter x et y, on peut supposer $|x|_b > |y|_b$. On a alors :

$$|x+y|_b \leqslant |x|_b = |(x+y) - y|_b \leqslant \max(|x+y|_b, |y|_b).$$

Puisque $|y|_b < |x|_b$, on a $\max(|x+y|_b, |y|_b) = |x+y|_b$ et $|x|_b = |x+y|_b$.

L'application $|\ |_b$ est une valeur absolue seulement si b est un nombre premier (ou la puissance d'un nombre premier).

Par contre, l'on a, par exemple, $|10 = 2 \times 5|_{10} = 1/10$ et $|2|_{10}|5|_{10} = 1 \times 1 = 1$. Il faut donc manipuler la notation $|\ |_b$ avec précaution.

Nous allons utiliser la notion de distance et les propriétés des espaces métriques (*cf.* [L2, II.1, 1]). Nous rappelons que la boule ouverte (resp. fermée) de centre a et de rayon $r > 0$ d'un espace métrique (X, d) est :

$$B(a,r) := \{x \in X \mid d(a,x) < r\} \quad (\text{resp. } B'(a,r) : \{x \in X \mid d(a,x) \leqslant r\}).$$

Les boules ouvertes (resp. fermées) sont des ouverts (resp. des fermés), mais nous allons voir qu'une boule ouverte peut être un fermé et qu'une boule fermée peut être un ouvert.

On peut avoir $B(a,r) = B(a',r')$ avec $a \neq a'$ et $r \neq r'$.

En utilisant la proposition 12, on vérifie que l'application $d_b : E_b \times E_b \to \mathbb{Q}_+$ définie par $(x, y) \mapsto d_b(x, y) := b^{-\omega_b(x-y)}$ est une distance sur E_b (en convenant de $b^{-\infty} = 0$) :

(i) $\forall x, y \in E_b,\ \ d_b(x, y) = 0 \Leftrightarrow x = y$;

(ii) $\forall x, y \in E_b,\ \ d_b(x, y) = d(y, x)$;

(iii) $\forall x, y, z \in E_b,\ \ d_b(x, y) \leqslant d_b(x, z) + d_b(z, y)$

On a même une version plus précise de (iii) :

(iii') $\forall x, y, z \in E_b,\ \ d_b(x, y) \leqslant \max\{d_b(x, z), d_b(z, y)\}$. Une distance vérifiant cette condition est appelée *ultramétrique*. Un espace muni d'une telle distance est dit *espace ultramétrique*.

La distance $d_b(x, y)$ est « petite » si l'ordre $\omega_b(x - y)$ est « grand ». Le choix de $b^{-\omega_b(x-y)}$ est contingent. On aurait aussi bien pu choisir $a^{-\omega_b(x-y)}$, avec $a \in \mathbb{N}^*$ arbitraire. Nous justifierons ce choix plus loin (*cf.* le théorème 93 de la page 506). Pour tout $x \in E_b$, on a $d_b(0, x) = |x|_b \in b^{\mathbb{Z}}$. Puisque $b^{\mathbb{Z}} \subset \mathbb{Q}_+$ est discret, deux boules de même centre et de rayons différents peuvent être égales.

On a, avec la distance $d := d_p$ sur $X := \mathbb{Q}_p$: $B'(0, 1) = \mathbb{Z}_p$ et :

$$B(0, 1) = B'(0, p^{-1}) = pB'(0, 1) = p\mathbb{Z}_p.$$

Proposition 13.

(i) L'anneau E_b muni de la distance d_b est un *anneau topologique* (i. e. l'addition, la multiplication et le passage à l'opposé sont continus) .

(ii) Les sous-ensembles $\mathbb{N}$ et $\mathbb{Z}$ de $\mathbb{Z}_b$ sont denses dans $(\mathbb{Z}_b, d_b)$; $\mathbb{Q}$ est dense dans (E_b, d_b).

Démonstration.

(i) Pour tout $x \in E_b$, on a $|-x|_b = |x|_b$: si α est le premier chiffre non nul de x, le premier chiffre non nul de $-x$ est $p - \alpha$, il est donc non nul. On montre la continuité des opérations comme dans $\mathbb{R}$ (en replaçant $|\ |_\infty$ par $|\ |_b$).

(ii) Pour tout $x \in \mathbb{Z}_b$ et pour tout $k \in \mathbb{N}^*$, on a $x_k := \mathrm{trg}_k(x) \in \mathbb{N}$ et $\left|x - \mathrm{trg}_k(x)\right|_b \geqslant k$, donc $d\left(x, \mathrm{trg}_k(x)\right) \leqslant p^{-k}$. On en déduit que $\mathbb{N}$ est dense dans $\mathbb{Z}_b$ (la suite (x_k) des troncatures converge vers x). A fortiori $\mathbb{Z}$ est dense dans $\mathbb{Z}_b$, puisque $\mathbb{N} \subset \mathbb{Z}$. On montre, en utilisant la densité de $\mathbb{N}$ dans $\mathbb{Z}_b$, que $\mathcal{D}_+$ est dense dans E_b. La densité de $\mathbb{Q}$ dans E_b en résulte, puisque $\mathcal{D}_+ \subset \mathbb{Q}$. ■

La topologie de (E_b, d_b) est invariante par translation. Pour la comprendre, il suffit de décrire un système fondamental de voisinages de 0, ce que nous ferons dans la suite.

Exemple. Soit K un corps. Dans [L2, I.1, 4.2], on a défini une notion de distance sur un corps de séries formelles $K((X))$ à l'aide de l'ordre ω $\ d(f, g) := a^{-\omega(f-g)}$ (où $a > 1$ est un réel fixé). Cette distance est ultramétrique. On a défini une valeur absolue $|f| := a^{-}\omega(f)$. Pour cette distance, $K[X]$ est dense dans $K[[X]]$, $K(X)$ est dense dans $K((X))$. On a montré que l'espace métrique $(K((X)))$ est *complet* (*cf.* [L2, I.1, 4.2.1, théorème 60]).

Exercice 9.

Soit p un nombre premier. On munit $\mathbb{N}$ de la distance d_p. Montrer que de toute suite de $\mathbb{N}$ on peut extraire une suite de Cauchy.

Solution. On procède par récurrence par extractions de suites et on utilise ensuite un procédé diagonal. Soit $x := (x_n)_{n\in\mathbb{N}}$ une suite de $\mathbb{N}$. On écrit les x_n en base p. Le nombre de chiffres $0, 1, \ldots, p-1$ étant fini, il existe un chiffre α_0 et une suite $x^{(1)} = (x_n^{(1)} = x_{\varphi(n)})_{n\in\mathbb{N}}$ extraite x de telle que, pour tout $n \in \mathbb{N}$, le premier chiffre de $x_n^{(1)}$ soit α_0. Il existe un chiffre α_1 et une suite $x^{(2)} = (x_n^{(2)})_{n\in\mathbb{N}}$ extraite de $x^{(1)}$ telle que, pour tout $n \in \mathbb{N}$, les deux premiers chiffres de $x_n^{(2)}$ soient $\alpha_1\alpha_0$. On définit ainsi par récurrence, pour tout $k \in \mathbb{N}^*$, un chiffre α_{k-1} et une suite $x^{(k)} = (x_n^{(k)})_{n\in\mathbb{N}}$ telle que, pour tout $n \in \mathbb{N}$, les k premiers chiffres de $x_n^{(k)}$ soient $\alpha_{k-1}\ldots\alpha_1\alpha_0$. On définit ensuite une suite $y = (y_n)$ par $y_0 = x_0$ et, pour tout $n \in \mathbb{N}^*$, $y_n := x_n^{(n)}$. C'est une suite extraite de x et, pour tous $r, s \geqslant n$, les n premiers chiffres de y_r et y_s coïncident. Par suite y est une suite de Cauchy.

Exercice 10.

Soit p un nombre premier.

1. Soit $r \in \mathbb{R}_+^*$. Montrer que l'ensemble S_r des nombres rationnels appartenant à $p^{\mathbb{Z}}$ et strictement inférieurs à r a un maximum. On le note r'.
   Montrer que $B(0, r) = B'(0, r')$.
2. Soit $x \in \mathbb{Z}_p$. Montrer que le singleton $\{x\}$ est l'intersection des boules de centre x et de rayon p^{-m}, pour $m \in \mathbb{N}$.

Solution.

1. L'ensemble S_r n'est pas vide : quand $m \in \mathbb{Z}$ tend vers $-\infty$, p^m tend vers 0. Soit $r_0 \in S_r$. Le sous-ensemble $p^{\mathbb{Z}} \subset \mathbb{R}$ est discret, donc il n'existe qu'un nombre *fini* d'éléments de S_r dans le segment $[r_0, r]$. Le plus grand est le maximum r' de S_r. Si $r' < r_1 < r$, alors il n'existe pas de point $x \in B(0, r)$ tel que $d_p(0, x) = r_1$, donc $B(0, r) = B'(0, r')$.
2. C'est immédiat en utilisant les définitions. Ces boules forment un système fondamental de voisinages de x.

Définition 6. Un espace métrique (X, d) est dit *totalement discontinu* si la composante connexe de tout point $x \in X$ est le singleton $\{x\}$.

Proposition 14. Soit (X, d) un espace métrique. On suppose que tout point x admet un système fondamental de voisinages à la fois ouverts et fermés. Alors (X, d) est totalement discontinu.

Démonstration. Un espace métrique est séparé. Soit C_x la composante connexe de x.

Si $C_x \neq \{x\}$, il existe $y \in C_x$, $y \neq x$. Puisque l'espace est séparé, il existe un voisinage V de x qui ne contient pas y. Par hypothèse on peut supposer que V est ouvert et fermé. Notons $A := C_x \cap V$ et $B := C_x \setminus V$. On a $C_x = A \cup B$, la réunion étant disjointe. On a $x \in A$ et $y \in B$, donc $A \neq \varnothing$ et $B \neq \varnothing$. Par ailleurs A est un ouvert de C_x et B est aussi un ouvert de C_x (A est aussi fermé), donc C_x n'est pas connexe. On aboutit à une contradiction. ∎

Proposition 15. Soit (X, d) un espace ultramétrique.

(i) *Tout triangle est isocèle*. La longueur commune de deux côtés égaux majore celle du troisième.

(ii) Tout point d'une boule (ouverte ou fermée) en est "le" centre.

(iii) Les boules sont *à la fois ouvertes et fermées*.

(iv) Étant données deux boules ouvertes (resp. fermées), elles sont disjointes ou l'une est contenue dans l'autre.

(v) L'espace (X, d) est *totalement discontinu*.

Démonstration.

(i) C'est une généralisation de l'exercice 8. La longueur d'un côté d'un triangle est majorée par la plus grande des longueurs des deux autres côtés. Soient a, b, c les longueurs. On suppose que a et la plus grande. On a $a \leqslant \max\,(b, c)$, donc $a = b$ ou $a = c$. Si $a = b$, alors $c \leqslant a$.

(ii) On traite le cas d'une boule ouverte (celui d'une boule fermée est similaire). On a $b \in B(a, r)$ si, et seulement si, $d(a, b) < r$. Soit $x \in B(a, r)$. On a $d(a, x) < r$, donc $d(b, x) \leqslant \max\,(d(a, b), d(a, x)) < r$ et $x \in B(b, r)$. Ainsi $B(a, r) \subset B(b, r)$. Échangeant a et b, on a l'inclusion inverse, donc $B(a, r) = B(b, r)$.

(iii) On traite le cas d'une boule ouverte $B(a, r)$. Celui d'une boule fermée est très similaire et laissé au lecteur. Montrons que $B(a, r)$ est fermée. ceci revient à montrer que son complémentaire $C := \{x \in X \mid d(a, x) \geqslant r\}$ est ouvert.

Soient $y \in C$ et $0 < r' < r$. Nous allons montrer que la boule ouverte $B(y, r')$ est contenue dans C, ce qui prouvera que C et ouvert. Soit $z \in B(y, r')$. On a $d(z, y) < r' < r \leqslant d(a, y)$. d'après (i), appliqué au triangle (a, y, z) on a $d(a, z) = d(a, y) \geqslant r$, c'est-à-dire $y \in C$ et $B(y, r') \subset C$.

(iv) On traite le cas des boules ouvertes. Celui des boules fermées est similaire et laissé au lecteur. Soient $B(a, r)$ et $B(a', r')$. On peut supposer $r \leqslant r'$.

On suppose $B(a, r) \cap B(a', r') \neq \varnothing$. Soit $c \in B(a, r) \cap B(a', r')$.

D'après (ii), on a $B(a, r) = B(c, r) \subset B(c, r') = B(b, r')$.

(v) Tout point possède un système fondamental de voisinages formé de boules ouvertes. D'après (iii), elles sont aussi fermées. La proposition 14 permet alors de montrer que X est totalement discontinu. ∎

Ces propriétés des espaces ultramétriques peuvent sembler au premier abord assez étranges. Nous conseillons au lecteur d'essayer de les comprendre en étudiant des exemples de boules dans le cas des espaces $(\mathbb{Q}_p, d_p)$, en s'aidant de l'exercice 10. Il pourra aussi considérer l'exemple des anneaux (ou des corps) de séries formelles.

Exercice 11.
Soit $(X, d) := (\mathbb{Q}_p, d_p)$. Montrer que la boule unité fermée $B'(0,1) = \mathbb{Z}_p$ de $\mathbb{Q}_p$ est une réunion finie disjointe de boules unité ouvertes. Montrer qu'elle est une réunion disjointe de p boules fermées de rayon $1/p$. Montrer que, pour tout $m \in \mathbb{N}^*$, $B'(0,1)$ est une réunion disjointe de p^m boules fermées de rayon $1/p^m$. Comparer la frontière de $B'(0,1)$ et $B'(0,1) \setminus B(0,1)$.
Solution.
On a $B(0,1) = pZ_p = B'(0,1/p)$. On a $\mathbb{Z}_p = pZ_p \cup (1 + p\mathbb{Z}_p) \cup \ldots \cup (p-1+p\mathbb{Z}_p)$, donc $B'(0,1) = \mathbb{Z}_p = B(0,1) \cup B(1,1) \cup B(2,1) \cup \ldots \cup B(p-1,1)$.
Si, pour $i, j \in \{0, 1, \ldots, p-1\}$, l'on a $B(i,1) \cap B(j,1) \neq \varnothing$, alors $B(i,1) = B(j,1)$, donc $d_p(i,j) < 1$ et $i = j$: si $i \neq j$, on a $d_p(i,j) = |i-j|_p = 1$.
On a, pour $i \in \{0, 1, \ldots, p-1\}$, $B(i,1) = B'(i,1/p)$ donc :

$$\mathbb{Z}_p = B'(0,1) = B'(0,1/p) \cup B'(1,1/p) \cup \ldots \cup B'(p-1,1/p). \tag{5}$$

On a $B'(i,1/p) = i + \frac{1}{p}B'(0,1)$. En reportant dans l'égalité (5) ci-dessus, puis en remplaçant chaque $B'(0,1)$ par une réunion de p boules fermées en utilisant à nouveau (5), on écrit $B'(0,1)$ comme une réunion de p^2 boules fermées de rayon $1/p^2$. On poursuit par récurrence.
La frontière de $B'(0,1)$ est *vide*, puisque $B'(0,1)$ est ouvert et fermé.
Par contre, $S := B'(0,1) \setminus B(0,1)$ n'est pas vide. Puisque $B(0,1) = B'(0,1/p)$, on a $S = B'(1,1/p) \cup \ldots \cup B'(p-1,1/p)$. Ainsi S est la réunion de $p-1$ boules fermées de rayon $1/p$.

Il est possible de transformer par un homéomorphisme le compact $\mathbb{Z}_p$ en un compact K de $\mathbb{R}$. Les propriétés de $\mathbb{Z}_p$ mises en évidence dans l'exercice ci-dessus se traduisent alors en une *structure fractale* sur K. Le prototype est le cas de $\mathbb{Z}_2$ en prenant pour K l'ensemble de Cantor K_3 (*cf.* ci-dessous l'exemple de la page 457). Tout ceci sera détaillé dans la correction de l'exercice type III.3.0 de la page 518.

On rappelle la définition suivante (*cf.* [L2, II.1, 1.2]).

Définition 7. Soit $((X_n, d_n))_{n\in\mathbb{N}}$ une famille dénombrable d'espaces métriques. On note $X := \prod_{n\in\mathbb{N}} X_n$. La formule suivante définit une distance δ sur X :

$$\delta(x,y) := \sum_{n\in\mathbb{N}} 2^{-n} \min\left(1, d_n(x_n, y_n)\right),$$

pour tous $x = (x_n)$ et $y = (y_n)$ appartenant à X.

Nous utiliserons des variantes. Soit $m \in \mathbb{N}$, $m \geqslant 2$. L'ensemble $\{\min(1, d_n(x_n, y_n))\}_{n\in\mathbb{N}}$ est majoré par 1, et $m^{-n} \leqslant 1$, donc la borne supérieure $\sup\limits_{n\in\mathbb{N}} (m^{-n} \min(1, d_n(x_n, y_n)))$ existe. On vérifie facilement que :

$$d^{(m)}(x, y) := \sup_{n\in\mathbb{N}} \left(m^{-n} \min(1, d_n(x_n, y_n))\right) \tag{6}$$

définit une distance sur X.

Théorème 16. Soit $((X_n, d_n))_{n\in\mathbb{N}}$ une famille dénombrable d'espaces métriques *compacts*. Le produit $X := \prod\limits_{n\in\mathbb{N}} X_n$ muni de la distance $d^{(m)}$ définie ci-dessus par (6) est *compact*.

Démonstration. En remplaçant d_n par $\min(1, d_n)$, on se ramène au cas où, pour tout $n \in \mathbb{N}$, $d_n \leqslant 1$. On note $x = (x_n)_{n\in\mathbb{N}}$ et, pour tout $n \in \mathbb{N}$, $\pi_n : X \to X_n$ la projection canonique ($\pi_n(x) = x_n$).

On vérifie qu'elle est continue en utilisant : $d_n(x_n, y_n) \leqslant m^n d^{(m)}(x, y)$.

1. On montre d'abord qu'une suite $x = (x_{(k)})_{k\in\mathbb{N}}$ de X est convergente si, et seulement si, toutes les suites $\pi_n(x) = (x_{(k),n})_{k\in\mathbb{N}}$ sont convergentes. La nécessité est évidente d'après ce qui précède. Montrons la suffisance. On suppose que toutes les suites $(x_{(k),n})$ convergent respectivement vers a_n. On note $a := (a_n) \in X$.

   Soit $\varepsilon > 0$. Il existe N^* tel que, pour $n \geqslant N$, on a $m^{-n} \leqslant \varepsilon$. On a donc, pour tout $k \in \mathbb{N}$, $\sup\limits_{n\geqslant N} \left(m^{-n} d_n(x_{(k),n}, a_n)\right) \leqslant \varepsilon$. On considère ensuite les $(N-1)$ suites convergentes $(x_{(k),n})_{k\in\mathbb{N}}$, $n \in [\![1, N-1]\!]$. Pour un tel n, il existe N'_n tel que, pour tout $k \geqslant N'_n$, on a $\sup\left(m^{-n} d_n(x_{(k),n}, a_n)\right) \leqslant \varepsilon$. On pose $N' := \max\limits_{n\in[\![1,N-1]\!]} N'_n$.

   Alors, pour tout $k \geqslant N'$, on a : $d^{(m)}(x_{(k)}, a) = \sup\limits_{n\in N} \left(m^{-n} d_n(x_{(k),n}, a_n)\right) \leqslant \varepsilon$.

2. D'après la définition séquentielle de la compacité d'un espace métrique (*cf.* [L2, II.1,3.1, définition 25]), il suffit, pour établir la compacité de X, de montrer que de toute suite $(x_{(k)})_{k\in\mathbb{N}}$ de X on peut extraire une suite convergente. On va montrer que l'on peut extraire une suite $(y_{(k)})$ de $(x_{(k)})$ telle que toutes les suites $\left(y_{(k),n} := \pi_n(y_{(k)})\right)$ convergent. La convergence de $(y_{(k)})$ dans X en résultera d'après 1. On va utiliser un procédé diagonal.

   L'espace métrique X_1 est compact, on peut donc extraire de $(x_{(k)})_{k\in\mathbb{N}}$ une suite $(x^{(1)}_{(k)} = x_{(\varphi(k))})_{k\in\mathbb{N}}$ telle que la suite $\left(x^{(1)}_{(k),1} := \pi_1\left(x^{(1)}_{(k)}\right)\right)_{k\in\mathbb{N}}$ converge vers a_1 dans X_1. Puisque X_2 est compact, on peut ensuite extraire de $(x^{(1)}_{(k)})_{k\in\mathbb{N}}$ une suite $(x^{(2)}_{(k)})_{k\in\mathbb{N}}$ telle que la suite $\left(x^{(1)}_{(k),2} := \pi_2\left(x^{(2)}_{(k)}\right)\right)_{k\in\mathbb{N}}$ converge vers a_2 dans X_2. On construit par récurrence, par extractions successives, des suites $(x^{(n)}_{(k)})_{k\in\mathbb{N}}$ telles que les suites $\left(x^{(n)}_{(k),n} := \pi_n\left(x^{(n)}_{(k)}\right)\right)_{k\in\mathbb{N}}$ convergent respectivement vers a_n dans X_n.

On peut alors définir une *suite diagonale* $(y_{(k)})_{k\in\mathbb{N}}$ de E, extraite de $(x_{(k)})_{k\in\mathbb{N}}$, par $y_{(k)} := x_{(k)}^{(k)}$. Pour tout $n \in \mathbb{N}$, la suite $(y_{(k),n})_{k\in\mathbb{N}}$ est une suite extraite de $(x_{(k),n})_{k\in\mathbb{N}}$; elle converge donc vers a_n dans X_n. D'après 1, la suite $(y_{(k)})$ converge vers $a := (a_n)_{n\in\mathbb{N}}$. ∎

Le théorème reste vrai avec la distance d définie plus haut (cf la définition 7) (nous l'avions admis dans ce cas dans le livre [L2] (*cf.* II.1, 3.1, proposition 48)). En effet les suites convergentes pour $(X, d^{(m)})$ et (X, δ) sont les mêmes.

On rappelle qu'une distance d sur un ensemble X est dite *discrète* si, pour tous $x, y \in X$, $d(x, x) := 0$ et $d(x, y) := 1$, si $x \neq y$. Alors, les singletons sont ouverts et fermés. Un ensemble *fini* muni de la distance discrète est *compact*. On en déduit le résultat suivant.

Corollaire 17. Soit $((X_n, d_n))_{n\in\mathbb{N}}$ une famille dénombrable d'ensembles *finis*, chacun étant muni de la distance *discrète* d_n . Alors l'espace métrique $\left(X, d^{(m)}\right)$ est *compact*

On déduit du corollaire ci-dessus le résultat suivant.

Proposition 18. L'espace métrique $(\mathbb{Z}_b, d_b)$ est *compact*.

Démonstration. On a $\mathbb{Z}_b = [\![0, b-1]\!]^{\mathbb{N}}$. Pour tout $n \in \mathbb{N}$, on met sur le $n^{ème}$ facteur $[\![0, b-1]\!]$ la distance discrète : $\mathbf{d}_n(\alpha, \beta) := 0$ si $\alpha = \beta$ et $\mathbf{d}_n(\alpha, \beta) := 1$ si $\alpha \neq \beta$. On vérifie que la distance b-adique $d_b(x, y)$ est égale à la distance :

$$\mathbf{d}^{(p)}(x, y) := \sup_{n\in\mathbb{N}} \left(p^{-n}\mathbf{d}(x_n, y_n)\right).$$ ∎

Une autre preuve de la proposition ci-dessus est proposée dans l'exercice III.3.11 de la page 523 .

Exemple. L'ensemble triadique de Cantor est homéomorphe au produit dénombrable d'ensembles finis $\{0, 1\}^{\mathbb{N}}$ muni de la distance définie ci-dessus (*cf.* [L2, II.1, 3.2, proposition 53]). Par suite, l'ensemble de Cantor est homéomorphe à $\mathbb{Z}_2$.

En rassemblant des résultats précédents, on obtient la proposition suivante (on rappelle qu'un espace métrique localement compact est complet).

Proposition 19. Soit $b \in \mathbb{N}$, $b \geqslant 2$.

(i) Pour tout $n \in \mathbb{N}$, l'idéal $p^n\mathbb{Z}_b$ (de $\mathbb{Z}_b$) est la boule fermée de rayon p^{-n} dans (E_b, d_b). Ces boules sont compactes. En particulier $\mathbb{Z}_b$ est la boule unité fermée.

(ii) Les boules (ouvertes et fermées) de (E_b, d_b) sont compactes.

(iii) Les boules $(p^n\mathbb{Z}_b)_{n\in\mathbb{N}}$ forment un système fondamental de voisinages compacts de 0 dans (E_b, d_b).

(iii) L'espace métrique (E_b, d_b) est localement compact. Il est une réunion dénombrable de compacts. Il est totalement discontinu.

(iv) L'espace métrique (E_b, d_b) est *complet*.

On a vu que $K[[X]]$ muni de la distance définie par l'ordre $d(f,g) := 2^{\omega(f-g)}$ est *complet*. Il est *compact* si, et seulement si, K est *fini* (*cf.* l'exercice III.3.13).

Exercice 12.

Montrer que $\mathbb{Z}_{10}$ a des diviseurs de zéro. Montrer, plus généralement, que, si b n'est pas une puissance d'un nombre premier, alors $\mathbb{Z}_b$ a des diviseurs de zéro.

Solution. On suppose que b n'est pas la puissance d'un nombre premier. Soit b_1 un diviseur de b tel que b_1 et $b_2 := b/b_1$ soient premiers entre eux et $b_1, b_2 \geqslant 2$. Par exemple $b = 10$, $b_1 = 5$, $b_2 = 2$.

On a, dans $\mathbb{Z}/b\mathbb{Z}$, $\bar{b}_1^m \neq \bar{0}$ et $\bar{b}_2^m \neq \bar{0}$, pour tout $m \in \mathbb{N}$. En effet, si l'on avait $\bar{b}_1^m = \bar{0}$, un diviseur premier de b_2 diviserait b_1, ce qui est impossible.

On considère la suite $(b_1^n)_{n\in\mathbb{N}^*}$ de $\mathbb{Z}_b$. Puisque $\mathbb{Z}_b$ et *compact*, on peut en extraire une sous-suite convergente. Il existe donc une suite strictement croissante d'entiers $(\mu(n))$ telle que la suite $\left(b_1^{\mu(n)}\right)_{n\in\mathbb{N}^*}$ converge. On note x_1 sa limite. Sa première décimale n'est pas nulle, donc $x_1 \neq 0$. On considère la suite $\left(b_2^{\mu(n)}\right)_{n\in\mathbb{N}^*}$ du compact $\mathbb{Z}_b$. On peut en extraire une sous-suite convergente $\left(b_2^{\nu(n)}\right)_{n\in\mathbb{N}^*}$. On note x_2 sa limite. Elle n'est pas nulle. Par ailleurs, la suite $\left(b_1^{\nu(n)}\right)_{n\in\mathbb{N}^*}$ converge vers x_1. Donc la suite produit $\left(b^{\nu(n)}\right)_{n\in\mathbb{N}^*}$ converge vers x_1x_2. Mais cette suite produit converge clairement vers 0 dans $\mathbb{Z}_b$, donc $x_1x_2 = 0$, avec x_1 et x_2 non nuls.

Nous admettons le résultat suivant (dû à Brouwer, 1910).

Théorème 20 (Brouwer). Soit X un espace métrique. Tout compact, non vide, totalement discontinu, sans points isolés de X est homéomorphe à l'ensemble de Cantor.

Corollaire 21.

(i) Tout produit dénombrable d'ensembles finis, munis d'une distance discrète, est homéomorphe à l'ensemble de Cantor.

(ii) Si $b, b' \in \mathbb{N}$, $b, b' \geqslant 2$, alors $\mathbb{Z}_b$ et $\mathbb{Z}_{b'}$ sont homéomorphes. En particulier, Si p et p' sont des nombres premiers $\mathbb{Z}_p$ et $\mathbb{Z}_{p'}$ sont homéomorphes ; $\mathbb{Q}_p$ et $\mathbb{Q}_{p'}$ sont homéomorphes.

Remarque. Il y a des similitudes entre les corps $\mathbb{R}$ et $\mathbb{Q}_p$. Ils sont métrisables, dans les deux cas $\mathbb{Q}$ est dense, ils sont localement compacts et complets, aucun des deux n'est algébriquement clos, ... Mais il y a aussi de grandes différences : il y a sur $\mathbb{R}$ une relation d'ordre total, tandis que $\mathbb{Q}_p$ n'est pas totalement ordonnable ; $\mathbb{R}$ est connexe, tandis que $\mathbb{Q}_p$ est totalement discontinu ; la boule unité fermée de $\mathbb{Q}_p$ est $\mathbb{Z}_p$, elle est ouverte et c'est un sous-anneau ; la boule unité fermée de $\mathbb{R}$ est $[-1,1]$, elle n'est pas ouverte, elle est stable par la multiplication, mais elle ne l'est pas par l'addition, ...

2.2.2 Convergence des suites et des séries

La plupart des propriétés des suites et des séries que nous connaissons dans $\mathbb{R}$ restent valables dans $\mathbb{Q}_p$ avec, *mutatis mutandis*, les même preuves. Mais, du fait que la norme de $\mathbb{Q}_p$ est ultramétrique, il y a d'importantes simplifications dans ce cas. Bien sûr une suite ou une série peuvent converger dans $\mathbb{Q}_p$ et pas dans $\mathbb{R}$ (et inversement). Si une suite de $\mathbb{Q}$ converge vers un rationnel respectivement dans $\mathbb{R}$ et $\mathbb{Q}_p$, il est possible que ces deux rationnels soient *différents* (*cf.* l'exercice III.3.23 de la page 524).
Nous traitons le cas, plus général, des suites et séries dans E_b, pour $b \in \mathbb{N}$, $b \geqslant 2$. Rappelons que si $b := p$ est un nombre premier $E_b = \mathbb{Q}_p$. Dans cette partie nous écrirons souvent $|\ |$ au lieu de $|\ |_b$ quand il n'y aura pas d'ambiguïté.

Suites et séries

La série géométrique $\sum\limits_{n\in\mathbb{N}} p^n$ converge vers $\frac{1}{1-p}$ dans $\mathbb{Q}_p$, : $\left|\sum\limits_{n\geqslant m} p^n\right|_p = p^{-m}$. Elle est évidemment divergente dans $\mathbb{R}$. Pour $p = 2$, on obtient l'égalité[13] : $1 + 2 + 4 + \cdots = -1$, dans $\mathbb{Q}_2$.
Plus généralement, si $x = \ldots\alpha_h\ldots\alpha_0, \alpha_{-1}\ldots\alpha_{n_0}$, la série de Hensel $\sum\limits_{n\geqslant n_0} \alpha_n p^n$ converge vers x dans $\mathbb{Q}_p$: $\left|\sum\limits_{n\geqslant m} \alpha_n p^n\right|_p \leqslant p^{-m}$. De façon équivalente la suite (x_n) des tronquées à gauche de x converge vers x.

Exercice 13.
Soit p un nombre premier. On définit une suite récurrente $(x_n)_{n\in\mathbb{N}^*}$ de $\mathbb{Q}_p$ par $x_1 := 1+p$ et $x_n : (x_{n-1})^p$, pour $n > 1$. Montrer que, dans $\mathbb{Q}_p$ cette suite converge vers 1 quand n tend vers $+\infty$. Qu'en est-il dans $\mathbb{R}$?

Solution. On a $(1+p)^p = 1 + \binom{p}{1}p + \binom{p}{2}p^2 + \cdots + p^p$ et $\binom{p}{1} = \frac{p(p-1)}{2}$ est divisible par p. Donc $x_2 \equiv 1 \mod p^2$: $x_2 = 1 + ap^2$. En répétant l'argument, en remplaçant p par ap^2, on obtient $x_3 \equiv 1 \mod p^3$. Puis, par récurrence, $x_n \equiv 1 \mod p^n$ ($n \in \mathbb{N}^*$), c'est-à-dire $|x_n - 1|_p \leqslant p^{-n}$. Ainsi, dans $\mathbb{Q}_p$, la suite x_n converge vers 1 quand n tend vers $+\infty$.
Dans $\mathbb{R}$, on a $p \geqslant 2$ et $x_1 \geqslant 3$. On en déduit, par récurrence, $x_n \geqslant 3^{2^{n-1}}$, donc la suite tend vers $+\infty$ quand n tend vers $+\infty$.

Proposition 22. Une suite de $\mathbb{Z}_b$ converge si, et seulement si, pour tout $k \in \mathbb{N}^*$, la suite des nombres formés par ses k-premiers chiffres (c'est-à-dire la suite des k-tronquées) est constante à partir d'un certain rang (qui dépend de k).

Démonstration. Soit (x_n)convergeant vers a. En remplaçant (x_n) par $(x_n - a)$, on se ramène au cas d'une suite tendant vers 0. Alors $\lim\limits_{n\to+\infty} |x_n|_b = b^{-\omega_b(x_n)} = 0$, donc $\lim\limits_{n\to+\infty} \omega_b(x_n) = +\infty$. Pour tout $k \in \mathbb{N}$, il existe $N_k \in \mathbb{N}$ tel que, pour tout $n > N_k$, $\omega_b(x_n) > k$. Alors les k premiers chiffres de x_n sont nuls.

[13] Cette égalité apparaît chez L. Euler dans le *De seriebus divergentibus* en 1760, avec évidemment une argumentation différente.

Inversement, supposons que pour tout $k \in \mathbb{N}^*$, la suite finie des k-premiers chiffres de x_n est constante pour $n > N_k$. Notons $\alpha_{k-1} \dots \alpha_0$ ces chiffres et $a_k := \alpha_{k-1} \dots \alpha_0$. Si l'on supprime le $k^{ème}$ chiffre α_{k-1} de a_k, on obtient a_{k-1}. La suite (a_k) est cohérente et définit $a \in \mathbb{Z}_b$. On vérifie que la suite (x_n) converge vers a. ■

Proposition 23.
Une suite (x_n) de E_b converge si, et seulement si, $\lim\limits_{n \to +\infty} |x_{n+1} - x_n|_b = 0$.

Démonstration. La condition est nécessaire.
Inversement, supposons $\lim\limits_{n \to +\infty} |x_{n+1} - x_n| = 0$. On va d'abord montrer que (x_n) est *une suite de Cauchy*. On suppose $m = n + r > n$. On a :

$$\begin{aligned} |x_m - x_n| &= |x_{n+r} - x_{n+r-1} + x_{n+r-1} - x_{n+r-2} + \cdots + x_{n+1} - x_n| \\ &\leqslant \max(|x_{n+r} - x_{n+r-1}|, |x_{n+r-1} - x_{n+r-2}|, \dots, |x_{n+1} - x_n|). \end{aligned}$$

Le résultat s'en déduit. La suite étant de Cauchy elle converge, puisque (E_b, d_b) est *complet* (*cf.* la proposition 19 de la page 457). ■
Bien sûr ce résultat est faux dans $\mathbb{R}$.

Proposition 24. Une série $\sum\limits_{n \in \mathbb{N}} a_n$ de E_b converge si, et seulement si, son terme général a_n tend vers 0. De plus, dans ce cas on a $\Big| \sum\limits_{n \in \mathbb{N}} a_n \Big|_b \leqslant \max\limits_{n \in \mathbb{N}} |a_n|_b$.

Démonstration. La nécessité est suffisante. Inversement, supposons $\lim\limits_{n \to +\infty} a_n = 0$. On note (S_n) la suite des sommes partielles. On a $S_{n+1} - S_n = a_n$. En appliquant à (S_n) la proposition 23 , on obtient la convergence de la suite (S_n), donc, par définition, celle de la série.
Puisque la série converge, on a :

$$\Big| \sum_{n \in \mathbb{N}} a_n \Big| = \lim_{m \to +\infty} \Big| \sum_{n \in [\![1,m]\!]} a_n \Big| \leqslant \lim_{m \to +\infty} \max_{n \in [\![1,m]\!]} |a_n| = \sup_{n \in \mathbb{N}} |a_n|.$$

Puisque $\lim\limits_{n \to +\infty} a_n = 0$, il existe $n_0 \in \mathbb{N}$, tel que $\sup\limits_{n \in \mathbb{N}} |a_n| = \max\limits_{n \in [\![1,n_0]\!]} |a_n| = \max\limits_{n \in \mathbb{N}} |a_n|$. ■

Exemple. Soit $x \in \mathbb{U}_1 = 1 + p\mathbb{Z}_p$. On écrit $x = 1 - py$, avec $y \in \mathbb{Z}_p$. On a $v_p(py) \geqslant 1$, donc $v_p\left((py)^n\right) \geqslant n$. Par suite la série $1 + py + (py)^2 + \cdots$ converge. Notons $x' \in \mathbb{Z}_p$ sa somme. On a $xx' = (1 - py)(1 + py + (py)^2 + \cdots) = 1$. Donc $x' \in \mathbb{Z}_p$ est l'inverse de x.
Pus généralement, soit $z \in Z_p$, non divisible par p. Il existe $m \in [\![1, p-1]\!]$ tel que $x := mz \equiv 1 \mod p$. On a $x = mz = 1 - py$, avec $y \in \mathbb{Z}_p$. On pose :

$$z' := m(1 + py + (py)^2 + \cdots) \in \mathbb{Z}_p.$$

On a $zz' = mz(1 + py + (py)^2 + \cdots) = 1$, donc z' est l'inverse de z.

La convergence des séries est souvent plus facile à manipuler dans le cas b-adique que dans le cas réel : *cf.* les deux propositions qui suivent (et l'exercice III.3.17).
Nous admettons le résultat suivant (une preuve fait l'objet de l'exercice III.3.16).

Proposition 25. Soient $\sum_{n\in\mathbb{N}} a_n$ et $\sum_{n\in\mathbb{N}} b_n$ deux séries convergentes de $\mathbb{Q}_p$, de sommes respectives a et b. La série produit de terme général $c_n = \sum_{i=0}^{n} a_i b_{n-i}$ est convergente et sa somme est ab.

Nous admettons le résultat suivant (une preuve fait l'objet de l'exercice III.3.18)

Proposition 26. Soit $(a_{m,n})_{m,n\in\mathbb{N}}$ une suite double de $\mathbb{Q}_p$. On suppose que :

1. pour tout $m \in \mathbb{N}$, $\lim_{n\to+\infty} a_{m,n} = 0$;
2. $\lim_{m\to+\infty} a_{m,n} = 0$ *uniformément* en n.

Alors les deux séries $\sum_{m\in\mathbb{N}} \left(\sum_{n\in\mathbb{N}} a_{m,n}\right)$ et $\sum_{n\in\mathbb{N}} \left(\sum_{m\in\mathbb{N}} a_{m,n}\right)$ convergent et leurs sommes sont égales.

La factorielle $n!$ apparaît dans de nombreuses suites et séries « intéressantes ». Il est utile d'en connaître une estimation p-adique. L'outil essentiel est une formule découverte par le mathématicien français Adrien-Marie Legendre[14] en 1808. Nous admettons les résultats suivants. Des preuves font l'objet de l'exercice III.3.15 de la page 523. L'idée est d'introduire, pour tout $v \in \mathbb{N}^*$, le nombre m_v d'éléments de $[\![1, n]\!]$ divisibles par p^v et de montrer que m_v est la partie entière de n/p^v en utilisant la division euclidienne. On rappelle que $v_p(m)$ est la valuation p-adique d'un entier m.

Proposition 27 (Formule de Legendre). Soit p un nombre premier.
On note $E(x)$ la *partie entière* d'un rationnel x.

(i) On a[a], pour tout $n \in \mathbb{N}^*$: $v_p(n!) = \sum_{v\in\mathbb{N}^*} E(n/p^v)$. C'est la *formule de Legendre*.

(i) On a, pour tout $n \in \mathbb{N}^*$: $v_p(n!) \leqslant \frac{n}{p-1}$ et $v_p(n!) \sim \frac{n}{p-1}$, quand $n \to +\infty$.

[a] En convenant de ne pas écrire les termes nuls.

On notera que le comportement asymptotique de $|n!|_p = p^{-v_p(n)}$ est très différent de celui de $n!$ dans le cadre classique : du point de vue p-adique $n!$ tend vers 0 quand n tend vers $+\infty$. Une conséquence notable est que la série exponentielle p-adique à un rayon de convergence *fini* $p^{1/(p-1)}$ (*cf.* l'exercice III.3.15 de la page 523).

Continuité et dérivabilité

Nous donnons les définitions et propriétés « minimales », en vue de l'application aux sommes de séries entières convergentes.
Soient $X \subset \mathbb{Q}_p$ et $a \in X$. On définit la continuité en a d'une fonction $f : X \to \mathbb{Q}_p$ en utilisant la métrique de $\mathbb{Q}_p$. On définit la dérivabilité comme dans le cas réel.

[14] Voir : A. M. Legendre, *Théorie des Nombres*, Paris : Firmin Didot, 1830.

Définition 8. Soient $V \subset \mathbb{Q}_p$ un ouvert et $f : V \to \mathbb{Q}_p$ une fonction. On dit que f est *dérivable* en $x \in V$ si la limite $\lim\limits_{h \to 0, h \neq 0} \frac{f(x+h)-f(x)}{h}$ existe.

On note alors $f'(x)$ la dérivée.

Une fonction localement constante sur une boule est dérivable, de dérivée identiquement nulle, mais n'est pas nécessaire constante. Une fonction peut être dérivable, de dérivée identiquement nulle sans être localement constante.

Exercice 14.

1. Trouver une fonction $f : B'(0,1) \to \mathbb{Q}_p$ non constante et de dérivée identiquement nulle sur $B'(0,1)$.
2. On définit une fonction $f : \mathbb{Z}_p \to \mathbb{Z}_p$ de la façon suivante. Si $x = \sum \alpha_n p^n$, $f(x) = \sum \alpha_n p^{2n}$. Montrer qu'elle est injective et dérivable sur $\mathbb{Z}_p$, de dérivée identiquement nulle

Solution.

1. On a l'union disjointe $B'(0,1) = B(0,1) \cup \ldots \cup B(p-1,1)$. La fonction f égale à i sur $B(i,1)$ est localement constante, mais n'est pas constante.
2. L'injectivité est évidente. La fonction f n'est donc pas localement constante. Nous allons estimer $|f(x) - f(y)|$ en fonction de $|x-y|_p$. On a $|x-y|_p = p^{-m}$, pour $m \in \mathbb{N}$ convenable. On suppose $m \geqslant 2$. Alors les $m-1$ premiers chiffres de x et y coïncident, donc les $(2m-1)$ chiffres de $f(x)$ et $f(y)$ coïncident. Par suite $|f(x) - f(y)|_p \leqslant p^{-2m} = |x-y|_p^2$. On en déduit que, pour tous $x, y \in \mathbb{Q}_p$, on a $\left|\frac{f(x)-f(y)}{x-y}\right|_p \leqslant |x-y|_p$. Par suite f est dérivable de dérivée nulle.

On vérifie facilement que la règle classique de dérivation des fonctions composées (*cf.* [L1, IV.2, 2.2, théorème 36]) reste valable.

Séries entières convergentes

Dans ce paragraphe, nous allons étudier *la convergence* des séries entières formelles[15] de $\mathbb{Q}_p[[X]]$ pour des valeurs p-adiques de la variable et donner quelques propriétés des fonctions définies par les sommes. On peut, *mutatis mutandis*, également étudier, pour $a \in \mathbb{Q}_p$, les séries entières en $(X-a)$. Nous nous limiterons au minimum. Notre but est essentiellement de permettre au lecteur de comprendre quelques exemples « intéressants ». Les preuves ne sont pas très difficiles, mais parfois assez longues. Certaines seront traitées sous forme d'exercices. Pour quelques autres nous donnerons seulement des références. On pourra, en première lecture, admettre les résultats énoncés dans ce paragraphe. Ils sont, pour l'essentiel, analogues à ceux des cas réels et complexes (*cf.* [L2, II.2 et II.6]), avec des « simplifications ». Il y a une exception, la composition de fonctions sommes de séries convergentes est plus délicate à étudier dans le cas p-adique.

[15] Pour ce qui concerne les séries entières formelles, le lecteur se reportera à [L2, I.1, 4]

Définition 9. Soit $\hat{f} := \sum_{n\in\mathbb{N}} a_n X^n \in \mathbb{Q}_p[[X]]$ une série entière formelle.

On définit $\rho(\hat{f}) := \sup\{\rho > 0 \mid \lim_{n\to+\infty} |a_n|_p \rho^n = 0\} \in \mathbb{R}_+ \cup \{+\infty\}$. On l'appelle le rayon de convergence de $\hat{f}$.

Si $\rho(\hat{f}) > 0$ on dit que la série entière $\hat{f}$ est *convergente*. Le domaine de convergence d'une série entière convergente $\hat{f}$ est l'ensemble des $x \in \mathbb{Q}_p$ tels que la série $\hat{f}$ converge.

La série définissant $\hat{f}(x)$ converge si, et seulement si, $|a_n x^n|_p$ tend vers 0. On peut, comme pour $\mathbb{R}$ et $\mathbb{C}$, calculer le rayon de convergence par la *formule de Hadamard* :

$$\rho(\hat{f}) = \frac{1}{\limsup_{n\to+\infty} |a_n|_p^{1/n}}. \tag{7}$$

(Nous l'admettons. La preuve est similaire à la preuve classique, elle est l'objet de l'exercice III.3.19 de la page 524).

On vérifie facilement le résultat suivant.

Proposition 28. Soit $\hat{f} := \sum_{n\in\mathbb{N}} a_n X^n \in \mathbb{Q}_p[[X]]$ une série entière formelle.

Soit $x \in \mathbb{Q}_p^*$. Si la série $\hat{f}(x)$ converge, alors, pour tout $y \in \mathbb{Q}_p$ tel que $|y|_p \leqslant |x|_p$, la série $\hat{f}(y)$ converge. On a $\rho(\hat{f}) \geqslant |x|_p$. Le domaine de convergence de $\hat{f}$ est une boule de centre 0 ou $\mathbb{Q}_p$ tout entier. Il contient la boule ouverte $B\left(0, \rho(\hat{f})\right)$.

Le domaine de convergence ne contient pas nécessairement la boule fermée $B'(0, \rho(\hat{f}))$.

Exemple. Toute série de $\mathbb{Z}_p[[X]]$ converge sur la boule ouverte $B(0,1)$. Son rayon de convergence est $\geqslant 1$.

Exercice 15.

Soit $\hat{f} := \sum_{n\in\mathbb{N}} a_n X^n \in \mathbb{Q}_p[[X]]$ une série entière formelle convergente. Soit $r > 0$ tel que $B'(0,r)$ soit contenu dans le domaine de convergence. Alors la fonction f définie par $\hat{f}$ est bornée et uniformément continue sur $B'(0,r)$.

Solution.

On a $|f(x)| \leqslant \max |a_n| r^n$, donc f est bornée sur $B'(0,r)$. On note $M_r := \max |a_n| r^n$.

Montrons que f est continue. Pour $x, y \in B'(0,r)$, on a :

$$f(x) - f(y) = \sum_{n\geqslant 1} a_n (x^n - y^n) = (x-y) \sum_{n\geqslant 1} a_n (x^{n-1} + x^{n-2}y + \cdots + y^{n-1}).$$

On en déduit : $|f(x) - f(y)| \leqslant \frac{1}{r} M_r |x-y|$. La continuité uniforme en résulte. Notons que la continuité uniforme et la propriété d'être bornée résultent aussi de la continuité, puisque $B'(0,r)$ est compacte.

En utilisant la proposition 25 (pour (ii)), on prouve facilement le résultat suivant.

Proposition 29. Soient $\hat{f}, \hat{g} \in \mathbb{Q}_p[[X]]$ deux séries entières formelles. on suppose qu'elles convergent en $x \in \mathbb{Q}_p$. On note $\hat{f}(x)$ et $\hat{g}(x)$ les sommes. Alors :

(i) La série $\hat{f} + \hat{g}$ converge en x et $(\hat{f} + \hat{g})(x) = \hat{f}(x) + \hat{g}(x)$.

(ii) La série $\hat{f}\hat{g}$ converge en x et $(\hat{f}\hat{g})(x) = \hat{f}(x)\hat{g}(x)$.

Dans le cas p-adique, la composition des séries entières pose un problème. Il y a par comparaison avec les cas réels et complexes, une condition supplémentaire. Si elle n'est pas réalisée le résultat attendu peut être faux (*cf.* par exemple le cas des exponentielles et logarithmes 2-adiques dans la question 6 de l'exercice III.3.25). Nous admettons le résultat suivant[16].

Proposition 30. Soient $\hat{f} = \sum a_n X^n, \hat{g} = \sum b_n X^n \in \mathbb{Q}_p[[X]]$ deux séries entières formelles, avec $\hat{g}(0) = 0$. On note $\hat{h} = \hat{f} \circ \hat{g}$ la composée. On suppose que :

(i) $\hat{g}$ converge en x ;

(ii) si $r := \max\limits_{n \in \mathbb{N}^*} (|b_n x^n|_p)$, on a $\lim\limits_{n \to +\infty} |a_n|_p r^n = 0$.

Alors la série $\hat{h}$ converge en x, la série $\hat{f}$ converge en $\hat{g}(x)$ et l'on a $\hat{h}(x) = \hat{f}(\hat{g}(x))$

Il y a des variantes de cet énoncé.
On peut par exemple utiliser le résultat de l'exercice suivant.

Exercice 16.
Les notations sont celles de la proposition 30 ci-dessus. On suppose que $\hat{f}(\hat{g}(x))$ est défini et que, pour tout $n \in \mathbb{N}$, $|b_n x^n|_p \leqslant |g(x)|_p$. Montrer que la condition (ii) de la proposition 30 ci-dessus est satisfaite.
Solution. On a toujours $|g(x)|_p \leqslant \max(|b_n x^n|_p))$. Donc l'hypothèse entraîne :

$$|g(x)|_p = \max(|b_n x^n|_p) = r.$$

Si la série $\hat{(f)}$ converge en $\hat{g}(x)$, on a $\lim\limits_{n \to +\infty} |a_n|_p r^n = 0$.

Remarque. Soit $\hat{f} = \sum a_n X^n \in \mathbb{Q}_p[[X]]$ de rayon de convergence $\rho > 0$. La proposition 30 ci-dessus montre l'intérêt de la fonction réelle de variable réelle suivante (appelée *module de croissance* de $\hat{f}$) :

$$M_{\hat{f}} : [0, \rho[\mapsto M_{\hat{f}}(r) := \max_{n \in \mathbb{N}^*} (|a_n x^n|_p).$$

On vérifie facilement (*cf.* l'exercice III.3.20 de la page 524) que $M_{\hat{f}}$ est continue, *convexe* et qu'il existe une suite strictement croissante (σ_k) de $[0, \rho[$ (finie ou infinie, éventuellement vide) telle que, entre deux valeurs consécutives de cette suite $M_{\hat{f}}$ coïncide avec une fonction monomiale. La suite est finie ou elle converge vers ρ.

[16] Voir, par exemple, le livre de F. Gouvêa, *p-adic numbers, An Introduction*

On montre, comme dans le cas réel (ou complexe), le résultat suivant.

Proposition 31. Soient $\hat{f}, \hat{g} \in \mathbb{Q}_p[[X]]$ deux séries formelles convergeant sur une boule ouverte $B(0,r)$. Soit (x_n) une suite de $B(0,r)$ tendant vers 0 quand n tend vers $+\infty$, telle que $x_n \neq 0$ pour tout $n \in \mathbb{N}$.
On suppose que, pour tout $n \in \mathbb{N}$, $f(x_n) = g(x_n)$. Alors $\hat{f} = \hat{g}$.

Proposition 32.
Une série entière et sa série dérivée $\hat{f}'$ ont le même rayon de convergence.

Démonstration. On emploie la règle de Hadamard. On note que la série $f'(x)$ converge si, et seulement si, la série $xf'(x)$ converge. On et alors ramené à prouver :

$$\limsup_{n\to+\infty} \sqrt[n]{|a_n|_p} = \limsup_{n\to+\infty} \sqrt[n]{|na_n|_p}$$

Il suffit de montrer que $\lim\limits_{n\to+\infty} \sqrt[n]{|n|_p} = 1$. On a $1/n \leqslant |n|_p \leqslant 1$, donc $n^{-1/n} \leqslant |n|_p^{1/n} \leqslant 1$. Le résultat s'en déduit en utilisant $\lim\limits_{n\to+\infty} n^{-1/n} = 1$. ∎

Exemple. La dérivée de la série entière formelle $\log(1+X) := \sum\limits_{n\in\mathbb{N}^*} (-1)^{n-1}\dfrac{X^n}{n}$ est $\sum\limits_{n\in\mathbb{N}} (-1)^n X^n$. Le rayon de convergence de la seconde est 1. Le rayon de convergence de la série du logarithme est donc aussi 1. La série du logarithme ne converge pas pour $x = 1$. En effet le terme général $(-1)^{n-1}/n$ ne tend pas vers 0 : si n et divisible par p^m, $\left|(-1)^{n-1}/n\right|_p \geqslant p^m$. Par suite le domaine de convergence de la série du logarithme est $B(0,1)$. Pour $y \in B(1,1)$, on note $\log_p(y)$ la somme de la série $\log(1+(y-1))$. La fonction $\log_p$ est appelée *logarithme p-adique*.

On admet le résultat suivant (une preuve fait l'objet de l'exercice III.3.21 de la page 524).

Lemme 33. . Soit (f_n) une suite de fonctions définie sur la boule fermée $B'(0,r)$ de $\mathbb{Q}_p$, à valeurs dans $\mathbb{Q}_p$. On suppose que $\lim\limits_{n\to+\infty} f_n(h) = 0$ *uniformément* en h sur $B'(0,r)$. On note $g(h)$ la somme de la série $\sum\limits_{n\in\mathbb{N}} f_n(h)$. Alors, si toutes les limites et la somme écrites existent, on a l'égalité : $\lim\limits_{h\to 0} g(h) = \sum\limits_{n\in\mathbb{N}} \left(\lim\limits_{h\to 0} f_n(h)\right)$.

Proposition 34. Soit $\hat{f} \in \mathbb{Q}_p[[X]]$ une série entière convergente.
On note f sa somme sur son domaine de convergence $\mathcal{D}$. Soit $x \in \mathcal{D}$.
Alors $\hat{f}'(x) = \lim\limits_{h\to 0, h\neq 0} \frac{f(x+h)-f(x)}{h}$.

Cela s'écrit $\hat{f}'(x) = f'(x)$.
Démonstration. On note $\hat{f} = \sum a_n X^n$. Alors, pour $x, h \in \mathcal{D}$, $h \neq 0$:

$$\frac{f(x+h)-f(x)}{h} = \sum_{n\in\mathbb{N}^*}\sum_{k=1}^{n}\binom{n}{k}a_n x^{n-k}h^{k-1}.$$

Quand h tend vers 0, à x fixé, le premier membre tend vers $f'(x)$. Nous allons calculer la limite du second d'une autre façon, en utilisant le lemme 33 ci-dessus.
On suppose x *fixé*. Si $x \neq 0$, on suppose $|h| < |x| \leqslant r_1$.

Si $x = 0$, on suppose $|h| \leqslant r_1$ $(0 < r_1)$. On pose, $f_n(h) := \displaystyle\sum_{k=1}^{n}\binom{n}{k}a_n x^{n-k}h^{k-1}$. Le second membre s'écrit alors $\sum\limits_{n\in\mathbb{N}^*} f_n(h)$. Nous allons voir que $\lim\limits_{h\to 0}\sum\limits_{n\in\mathbb{N}^*} f_n(h) = \sum\limits_{n\in\mathbb{N}^*}\lim\limits_{h\to 0} f_n(h)$.
On a, pour tout $n \in \mathbb{N}^*$, $\lim\limits_{h\to 0} f_n(h) = na_n x^{n-1} = c_n$. La série $\sum\limits_{n\in\mathbb{N}^*} c_n = \sum\limits_{n\in\mathbb{N}^*} na_n x^{n-1}$ converge et sa somme est, par définition, $\hat{f}'(x)$.
On a, par ailleurs : $\left|\binom{n}{k}a_n x^{n-k}h^{k-1}\right| \leqslant |a_n|\rho_1^{n-1}$ (les $\binom{n}{k}$ sont des entiers). Cette majoration est indépendante de k et h, et $\lim\limits_{n\to+\infty}|a_n|\rho_1^{n-1} = 0$. On en déduit que la suite de polynômes (f_n) converge vers 0, uniformément en h, quand n tend vers $+\infty$. On peut donc appliquer le lemme 33 et l'on obtient : $\lim\limits_{h\to 0}\sum\limits_{n\in\mathbb{N}^*} f_n(h) = \sum\limits_{n\in\mathbb{N}^*} c_n$.
Par suite $f'(x) = \sum\limits_{n\in\mathbb{N}^*} c_n = \hat{f}'(x)$. ∎

Corollaire 35. Soient $\hat{f}, \hat{g} \in Q_p[[X]]$ deux séries convergentes sur une boule ouverte $B(0,r)$. On note f et g les sommes respectives. On suppose que $f' = g'$ sur $B(0,r)$. Alors il existe une constante $C \in \mathbb{Q}_p$ telle que $\hat{f} = \hat{g} + C$.
On a alors $f = g + C$ sur le domaine de convergence de f.

Démonstration. D'après la proposition 32, les deux séries ont le même rayon de convergence $\rho > 0$. D'après la proposition 34 ci-dessus, on a $f' = g'$ sur $B(0,r)$. On a alors, d'après la proposition 31, $\hat{f}' = \hat{g}'$ dans $\mathbb{Q}_p[[X]]$ et ceci implique que, sauf éventuellement le premier, les coefficients de $\hat{f}$ et $\hat{g}$ sont égaux. ∎

Soit $\hat{f} \in \mathbb{Q}_p[[X]]$ une série entière convergente. On note f sa somme sur son domaine de convergence D. On peut montrer que f est développable en série entière au voisinage de tout point de son domaine de convergence.

Proposition 36. Soit $\hat{f} = \sum a_n X^n \in \mathbb{Q}_p[[X]]$ une série entière convergente. On note f sa somme sur son domaine de convergence $\mathcal{D}$. Soit $x_0 \in \mathcal{D}$, $x_0 \neq 0$.

(i) Pour tout $m \in \mathbb{N}$, la série $\sum\limits_{n\geqslant m}\binom{n}{m}a_n x_0^{n-m}$ converge. On note b_m sa somme.

(ii) On considère la série formelle $\hat{g} := \sum\limits_{m\in\mathbb{N}} b_m(X-x_0)^m \in \mathbb{Q}_p[[X-x_0]]$. Alors, les séries $\hat{f}$ et $\hat{g}$ ont le même domaine de convergence (« en X ») et, pour tout x appartenant au domaine de convergence, l'on a $g(x-x_0) = f(x)$.

Démonstration.

(i) Les $\binom{n}{m}$ sont des entiers et $x_0 \in \mathcal{D}$, donc :

$$\left|\binom{n}{m} a_n x_0^{n-m}\right|_p \leqslant |a_n x_0^{n-m}|_p = |x_0^{-m}|_p |a_n x_0^n|_p$$

On a $\lim\limits_{n\to+\infty} |x_0^{-m}|_p |a_n x_0^n|_p = 0$, donc la série définissant b_m converge.

(ii) Soit $x \in \mathcal{D}$. On a :

$$f(x) = \sum_{n\in\mathbb{N}} a_n (x - x_0 + x_0)^n = \sum_{n\in\mathbb{N}} \sum_{m\leqslant n} \binom{n}{m} a_n x_0^{n-m} (x - x_0)^m.$$

Posons $c_{m,n} := \binom{n}{m} a_n x_0^{n-m} (x - x_0)^m$, si $m \leqslant n$ et $c_{m,n} := 0$ si $n < m$.
Alors $f(x) = \sum\limits_{n\in\mathbb{N}} \sum\limits_{m\in\mathbb{N}} c_{m,n}$. Nous admettons que l'on peut appliquer la proposition 26 de la page 461 sur les suites doubles et *intervertir les sommations* (une preuve fait l'objet de l'exercice III.3.22 de la page 524).
Alors :

$$\begin{aligned} f(x) &= \sum_{n\in\mathbb{N}} \sum_{m\in\mathbb{N}} c_{m,n} = \sum_{m\in\mathbb{N}} \sum_{n\in\mathbb{N}} c_{m,n} = \sum_{m\in\mathbb{N}} \sum_{n\geqslant m} \binom{n}{m} a_n x_0^{n-m} (x - x_0)^m \\ &= \sum_{m\in\mathbb{N}} b_m (x - x_0)^m = g(x - x_0). \end{aligned}$$

On vérifie facilement que l'on peut intervertir les rôles de $\hat{f}$ et $\hat{g}$. Par suite les domaines de convergence sont les mêmes. ■

Ainsi la somme d'une série convergente sur son disque de convergence est développable en série entière en tout point du disque de convergence. Mais, contrairement ce qui peut se passer dans le cas complexe, ceci ne permet jamais de prolonger la fonction en dehors du « disque de convergence » car tout point du disque est centre du disque. Pour certaines fonctions il y a d'autres méthodes. Par exemple pour la fonction définie, pour $x \in B(0,1)$ par $x \mapsto 1 + x + x^2 + \cdots + x^n + \cdots$, on peut utiliser le fait qu'elle est la restriction à $B(0,1)$ d'une fraction rationnelle $\frac{1}{1-x}$ pour la prolonger à $\mathbb{Q}_p \setminus \{1\}$.

Exercice 17.

Montrer que la fonction $f : \mathbb{Q}_p \setminus \{1\}$ définie par $x \mapsto f(x) := \frac{1}{1-x}$ est développable en série entière au voisinage de tout point de son domaine de définition.

Solution. Soit $x_0 \neq 1$. On a, pour tout $x \in B\left(x_0, |1 - x_0|_p\right)$:

$$\begin{aligned} f(x) &= \frac{1}{1-x} = \frac{1}{1-x_0} \frac{1}{1 - \frac{x-x_0}{1-x_0}} \\ &= \frac{1}{1-x_0} \left(1 + (1-x_0)^{-1}(x-x_0) + \cdots + (1-x_0)^{-n}(x-x_0)^n + \cdots\right). \end{aligned}$$

Remarque : prolongement analytique.

Dans le cas complexe, un outil essentiel de l'étude des *fonctions analytiques*, c'est-à-dire des fonctions localement définies par la somme d'une série entière convergente, est *la théorie du prolongement analytique* (*cf.* [L2, II.6]). Elle et basée sur des propriétés de *connexité*. Dans le cas p-adique, les seuls sous-ensembles connexes sont les singletons et les seuls chemins continus sont les chemins constants. il n'est donc pas possible d'étendre la méthode utilisée pour $\mathbb{C}$. Il existe toutefois une autre méthode pour faire des prolongements analytiques. Elle a été initiée par le mathématicien français Marc Krasner. L'idée est d'utiliser l'approximation par des *fractions rationnelles*[17]. Cette approche a été développée en particulier par John Tate et Phlippe Robba.

2.2.3 Caractérisation des rationnels par leurs écritures p-adiques

Dans le cas des écritures décimales à droite, nous connaissons une caractérisation des rationnels : un nombre réel est rationnel si, et seulement si, son écriture décimale est *périodique à partir d'un certain rang* (*cf.* [L1, IV.1, 3.1.2]). Il y a une caractérisation semblable pour les écritures à droite en base b. Nous allons voir que, si $b = p$ est un nombre premier, il y a aussi un résultat similaire pour les écritures p-adiques à gauche. On peut admettre le résultat en première lecture.

On définit les écritures à gauche périodiques (resp. périodiques à partir d'un certain rang) comme pour les écritures à droite, *mutatis mutandis* (*cf.* [L1, IV.1, 3.1.2]).

Exercice 18.

Soit $b = 3$. On considère le nombre $x \in \mathbb{Z}_3$ défini par $x := \ldots 01210121 \ldots 0121\underline{0121}1$. Montrer qu'il est rationnel.

Solution. Le bloc périodique est 0121. C'est l'écriture triadique de $16 = 1+6+9 = 1+2\times 3+3^2$. On a, en écritures triadiques : $x = \ldots 0 \ldots 01 + (0121)(\ldots 00010001 \ldots 0001)$, donc

$$x = 1 + 3 \times 16(1 + 3^4 + 3^8 + \cdots) = 1 + \frac{48}{1-3^4} = 1 - \frac{48}{80} = \frac{32}{80} = \frac{2}{5}.$$

Proposition 37. Soit $b \in \mathbb{N}$, $b \geqslant 2$. Si l'écriture de $x \in \mathbb{Q}_b$ est périodique à partir d'un certain rang, alors x est rationnel.

Démonstration. On procède comme dans l'exercice ci-dessus. On se ramène facilement au cas d'une écriture entière périodique.

Soit $x := \ldots \beta_{k-1} \ldots \beta_0 \beta_{k-1} \ldots \beta_0 \ldots \beta_{k-1} \ldots \beta_0 \underline{\beta_{k-1} \ldots \beta_0}$. Le bloc périodique est formé de k chiffres ($\beta_i \in [\![0, b-1]\!]$). Ce bloc $\beta_{k-1} \ldots \beta_0$ est l'écriture b-adique d'un entier $m \in \mathbb{N}$. On a $x = m + mp^k + mp^{2k} + \cdots$, donc $x = \frac{m}{1-p^k}$ qui est un rationnel. On note que x est négatif et vérifie $-1 < x \leqslant 0$. ∎

Exercice 19.

Soit $b = 3$.

1. Trouver un entier $k \in \mathbb{N}^*$ tel que $3^k \equiv 1 \mod 11$.
2. Montrer que l'écriture 3-adique de $-5/11$ est périodique.

[17] Dans le cas complexe on a dans ce cadre un résultat important, le théorème de Runge. Pour le cas p-adique, on pourra consulter par exemple le livre d'Yvette Amice, *Les nombres p-adiques*.

Solution. On vérifie $3^5 = 243 = 242+1 = 1+22\times 11$, donc $3^5 \equiv 1 \mod 11$ ($k=5$ est la plus petite valeur). Par suite :

$$-\frac{5}{11} = -\frac{5\times 22}{11\times 22} = \frac{110}{1-3^5} = \frac{2+3^2+3^4}{1-3^5} = (2+3^2+3^4)(1+3^5+3^{10}+\cdots).$$

On a donc en écritures triadiques à gauche :

$$-\frac{5}{11} = 11002\times \ldots 000010000100\underline{00001} = \ldots 110021100211\underline{002}.$$

Le bloc périodique est 11002.

On note que $-\frac{5}{11} \in [-1,0[$ et $\left|-\frac{5}{11}\right|_3 = 1$. Le résultat de l'exercice ci-dessus se généralise alors de la façon suivante (la preuve est similaire).

Lemme 38. Soit p un nombre premier. Soit $r \in \mathbb{Q}$ tel que $|r|_p = 1$ et $r \in [-1,0[$. Alors l'écriture p-adique à gauche de r est *périodique*.

Démonstration. Puisque $|r|_p = 1$ et $r<0$, on peut écrire $r = a/b$, avec $a,b \in \mathbb{Z}$, $a<0$ et $b \geqslant 1$, a et b n'étant pas divisibles par p. Les entiers p et b sont premiers entre eux, donc, d'après un théorème d'Euler (*cf.* [L1, II.2, 3.4, Théorème 50]), il existe un entier $k \in \mathbb{N}^*$ tel que $p^k \equiv 1 \mod b$. Il existe donc $c \in \mathbb{N}^*$ tel que $bc = p^k - 1$. Par suite : $r = \dfrac{a}{b} = \dfrac{ac}{bc} = \dfrac{ac}{p^k-1} = \dfrac{-ac}{1-p^k}$. On pose $d := -ac$. On a $d \in \mathbb{N}$ et $0 < d < p^k - 1$, donc d écrit en base p a au plus k chiffres : $d := \beta_{k-1}\ldots\beta_0$ (l'entier d n'est pas divisible par p, donc $\beta_0 \neq 0$). On a $\frac{1}{1-p^k} = 1+p^k+p^{2k}+\cdots$. L'écriture de r est donc périodique : $r = \ldots\beta_{k-1}\ldots\beta_0\beta_{k-1}\ldots\beta_0\underline{\beta_{k-1}\ldots\beta_0}$. ∎

Contrairement à une première intuition le résultat suivant ne nous parait pas si évident[18] .

Lemme 39. Soit p un nombre premier. Soient x et y deux écritures p-adiques à gauche entières telles que $x+m=y$, avec $m \in \mathbb{N}$. Si l'écriture de y est périodique à partir d'un certain rang, alors celle de x l'est aussi.

Démonstration. Si $y \in \mathbb{N}$, alors $x \in \mathbb{Z}$. À partir d'un certain rang les chiffres de x sont tous 0 ou tous $p-1$, donc le résultat est vrai. On suppose $y \notin \mathbb{N}$. Notons $y = \ldots\beta_h\ldots\beta_0$. On considère la suite des tronquées $y_n := \beta_{n-1}\ldots\beta_0$. C'est une suite d'entiers. La suite (β_n) est, par hypothèse, périodique à partir d'un certain rang et $y \notin \mathbb{N}$, donc le bloc périodique contient au moins un chiffe non nul. Par suite $\lim\limits_{n\to+\infty} y_n = +\infty$ (la limite étant prise au sens usuel) et il existe donc $k \in \mathbb{N}^*$ tel que $y_k > m$. Alors $x+m=y$ s'écrit $x = y_k - m + z$, avec $z := \sum\limits_{h\geqslant k} \beta_h p^h = \ldots\beta_h\ldots\beta_k 0\ldots 0$.

On a $y_k - m \in \mathbb{N}$ et $y_k - m \leqslant y_k < p^k$, donc $y_k - m = \gamma_{k-1}\ldots\gamma_0$. Par suite :

$$x = \ldots\beta_h\ldots\beta_k\gamma_{k-1}\ldots\gamma_0$$

et l'écriture de x est périodique à partir d'un certain rang, puisque celle de y l'est. ∎

[18] Il semble l'être pour plusieurs auteurs et il l'était apparemment pour Hensel en 1904.

Théorème 40. Soit p un nombre premier. Un nombre p-adique est *rationnel* si, et seulement si, son écriture p-adique à gauche est *périodique à partir d'un certain rang*.

Démonstration. La preuve est facile en utilisant les deux lemmes 38 et 39 ci-dessus. Il faut analyser un certain nombre de cas. Nous laissons les détails en exercice au lecteur. ∎

2.3 Zéros d'un polynôme dans un corps p-adique

2.3.1 Lemme de Hensel

Soit K un corps. On s'intéresse aux solutions d'une équation algébrique $f = 0$, avec $f \in K[X]$. Dans le cas réel $K := \mathbb{R}$, on obtient une solution de $f = 0$ à partir d'une « solution approchée x_1 », comme limite de la suite définie par récurrence par $x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}$, sous réserve que $f'(x_m)$ ne s'annule pas. C'est la *méthode de Newton*. Géométriquement x_{m+1} est l'abscisse à laquelle la tangente au graphe de f en $(x_m, f(x_m))$ coupe l'axe des abscisses.

Soit maintenant $f \in \mathbb{Z}[X]$. Supposons que l'on connaisse une « solution approchée » de $f = 0$, au sens suivant : on a $x_m \in \mathbb{Z}$ tel que $f(x_m) \equiv 0 \mod p^m$, avec $m \in \mathbb{N}^*$. Il est naturel de chercher à « améliorer l'approximation » en cherchant $x_{m'} \in \mathbb{Z}$ tel que $f(x_{m'}) \equiv 0 \mod p^{m'}$, avec $m' > m$ et $x_{m'} \equiv x_m \mod p^m$. L'idée de Hensel pour ce faire est d'imiter la méthode de Newton. Si l'on peut poursuivre indéfiniment le procédé, il est naturel de *définir* $\mathbb{Z}_p$ pour « donner un sens au résultat » et de dire qu'il existe une solution $\tilde{x}$ de $f = 0$ dans $\mathbb{Z}_p$ égale à x_m modulo p^m. Comme dans le cas réel, il y aura des conditions sur f'. Il y a de trés nombreuses versions et généralisations du lemme de Hensel. Nous commençons par « la plus simple ».

Proposition 41 (Lemme de Hensel). Soient p un nombre premier et $f \in \mathbb{Z}[X]$.

(i) Soit $m \in \mathbb{N}^*$. On suppose qu'il existe un entier $x \in \mathbb{Z}$ tel que :
$$f(x) \equiv 0 \mod p^m \quad \text{et} \quad f'(x) \not\equiv 0 \mod p.$$
Alors il existe un entier $y \in \mathbb{Z}$, *unique* module p^{2m}, tel que :
$$y \equiv x \mod p^m \quad \text{et} \quad f(y) \equiv 0 \mod p^{2m}.$$
De plus, on peut prendre $y = x - f(x)u$, où u est un entier tel que $uf'x) \equiv 1 \mod p^m$.

(ii) Soient $m \in \mathbb{N}^*$ et une solution $x_m \in A_m := \mathbb{Z}/p^m\mathbb{Z}$ de $f = 0$ dans A_m, telle que l'image de $f'(x_m)$ dans F_p ne soit pas nulle.

Alors il existe une solution *unique* $x_{m+1} \in A_{m+1} := \mathbb{Z}/p^{m+1}\mathbb{Z}$ de $f = 0$ dans A_{m+1} telle que $\varphi_{m+1}(x_{m+1}) = x_m$ (c'est-à-dire $x_{m+1} \equiv x_m \mod p^m$).

(iii) Si $x_1 \in \mathbb{F}_p$ est une solution de $f(X) = 0$ dans F_p, telle que $f'(x_1) \neq 0$ dans F_p, alors il existe une *unique* solution $x \in \mathbb{Z}_p$ de $f(X) = 0$ dans $\mathbb{Z}_p$ telle que l'image de x dans F_p soit x_1 (c'est-à-dire $x_1 = \varepsilon_1(x)$).

Démonstration.

(i) Il est naturel de chercher y sous la forme $x + tp^m$, avec $t \in \mathbb{Z}$.

D'après la formule de Taylor pour les polynômes de $\mathbb{Z}[X]$ (à l'ordre 1), on a, dans l'anneau $\mathbb{Z}[X]$: $f(X + tp^m) \equiv f(X) + f'(X)tp^m \mod (tp^m)^2\,\mathbb{Z}[X]$.

Par suite, si $t \in \mathbb{Z}$, $f(x+tp^m) \equiv f(x) + f'(x)tp^m \mod p^{2m}$.
Par hypothèse, il existe $a \in \mathbb{Z}$ tel que $f(x) = ap^m$. Alors :

$$f(x+tp^m) \equiv \big(a + f'(x)t\big)\, p^m \mod p^{2m}$$

et $f(x+tp^m) \equiv 0 \mod p^{2m}$ si, et seulement si, $a + f'(x)t \equiv 0 \mod p^m$. Cette derniére condition équivaut à $f(x) + f'(x)tp^m \equiv 0 \mod p^{2m}$. Si u est un entier tel que $uf'(x) \equiv 1 \mod p^m$ (un tel entier existe puisque l'image de $f'(x)$ dans F_p n'est pas nulle), elle est équivalente à $tp^m = -uf(x) \mod p^{2m}$. Ainsi $y := x - uf(x)$ répond à la question et si y' est une autre solution, $y \equiv y' \mod p^{2m}$.

(ii) C'est une conséquence immédiate de (i).

(iii) En utilisant (ii), on construit par récurrence une suite unique $(x_m)_{m\in\mathbb{N}^*}$, de premier terme la valeur $x_1 \in F_p$ donnée, telle que, pour tout $m \in \mathbb{N}^*$:

* $x_m \in \mathbb{Z}/p^m\mathbb{Z}$;
* $\varphi_{m+1}(x_{m+1}) = x_m$;
* x_m est une solution de $f(X) = 0$ dans $\mathbb{Z}/p^m\mathbb{Z}$.

La suite (x_m) est *cohérente* et définit un élément $x \in \mathbb{Z}_p$. Cet élément x répond à la question et c'est le seul. ∎

Exemples.

1. Soient $p = 5$ et $f(X) := X^2 + 1$. On a $f(\bar{2}) = \bar{0}$ et $f'(\bar{2}) = \bar{4} \neq \bar{0}$. On peut donc appliquer le lemme de Hensel et l'on obtient $x \in \mathbb{Z}_5$ tel que $x^2 + 1 = 0$.
2. Soient $p = 3$ et $f(X) := X^2 - 7$. On a $f(\bar{1}) = \bar{0}$ et $f'(\bar{1}) = \bar{2} \neq \bar{0}$.
   On obtient $x \in \mathbb{Z}_3$ tel que $x^2 - 7 = 0$.
3. Soient $p = 2$ et $f(X) := X^3 - X - 2$. On a $f(\bar{0}) = \bar{0}$ et $f'(\bar{0}) = \bar{1}$.
   On obtient $x \in \mathbb{Z}_2$ tel que $x^3 - x + 1 = 0$.
   En appliquant l'algorithme, on trouve $x = 2 + 2^2 + 2^4 + 2^7 + \cdots$.
   Par contre $f(\bar{1}) = \bar{0}$ et $f'(\bar{1}) = 0$. Il n'existe pas de $y \in \mathbb{Z}_2$ relevant $\bar{1} \in F_2$ tel que $f(y) = 0$. Il n'existe même pas de relèvement z dans $\mathbb{Z}/4\mathbb{Z}$ tel que $f(z) = 0$. On devrait avoir $z \equiv 1$ ou $z \equiv 3$, mais $f(1) \equiv 2 \mod 4$ et $f(3) \equiv 2 \mod 4$.

Exercice 20.

Soient p un nombre premier, $a, b \in \mathbb{Z}_p$, avec b non divisible par p. Montrer que le polynôme $f(X) := aX^2 + bX + p$ admet une racine dans Z_p. En déduire que $b^2 - 4ap$ est un carré dans $\mathbb{Z}_p$.

Solution. Le discriminant de f dans F_p est l'image de $b^2 - 4ap$ par $\varepsilon_1 : \mathbb{Z}_p \to F_p$. C'est donc un carré et il est non nul car b n'est pas divisible par p. Ainsi f admet deux racines simples dans F_p. D'après le Lemme de Hensel (proposition 41), $f(X)$ a une racine c dans $\mathbb{Z}_p$.

On a $4af(X) = (2aX + b)^2 - (b^2 - 4ap)$, donc $b^2 - 4ap = (2ac + b)^2$. Ainsi $b^2 - 4ap$ est *un carré* dans $\mathbb{Z}_p$.

Proposition 42. Le corps p-adique $\mathbb{Q}_p$ n'est pas totalement ordonnable.

Démonstration. Supposons qu'il existe sur $\mathbb{Q}_p$ un ordre en faisant un corps totalement ordonné. Alors cet ordre induit l'ordre ordinaire sur $\mathbb{Z}$ (*cf.* [L1, IV.1, 1.2.1, Corollaire 2]). On peut donc le noter $\leqslant$. Si x est un carré de $\mathbb{Q}_p$, on a $0 \leqslant x$ (*cf.* [L1, IV.1, 1.2.1, Proposition 1]).
On a $1 - 4p \in \mathbb{Z} \subset \mathbb{Z}_p$ et $1 - 4p < 0$ dans $\mathbb{Z}$, donc $1 - 4p < 0$ dans $\mathbb{Z}_p$, mais $1 - 4p$ est un carré dans Z_p (*cf.* l'exercice ci-dessus, avec $a = b = 1$). On aboutit donc à une contradiction. ∎

Le résultat suivant est une généralisation du lemme de Hensel. La preuve est similaire. Elle fait l'objet de l'exercice III.3.30 de la page 527.

Proposition 43 (Lemme de Hensel. Cas singulier). Soient p un nombre premier et $f \in \mathbb{Z}_p[X]$. Soient $m, k \in \mathbb{N}$ tels que $0 \leqslant 2k < m$. Soit $x \in \mathbb{Z}_p$ tel que $f(x) \equiv 0 \mod p^m$ et $v_p(f'(x)) = k$. Alors il existe $y \in \mathbb{Z}_p$ tel que $f(y) \equiv 0 \mod p^{m+1}$, $y \equiv x \mod p^{m-k}$ et $v_p(f'(y)) = k$.

2.3.2 Racines de l'unité dans $\mathbb{Q}_p$

Proposition 44. Soit p un nombre premier. L'anneau $\mathbb{Z}_p$ contient l'ensemble $V := \{x \in \mathbb{Q}_p \mid x^{p-1} = 1\}$ des racines $(p-1)^{\text{èmes}}$ de l'unité dans le corps $\mathbb{Q}_p$. Ces racines forment un sous-groupe cyclique d'ordre $p - 1$ de $\mathbb{U}$. Ce sous-groupe est isomorphe à F_p^*.

Démonstration. Le sous-ensemble $F_p^* := \bar{1}, \ldots, \overline{p-1} \subset F_p$ forme le groupe des racines $(p-1)^{\text{èmes}}$ de l'unité dans $\mathbb{F}_p$ (petit théorème de Fermat). C'est un groupe cyclique d'ordre $p - 1$. Toutes les racines sont *simples*. Notons $f(X) = X^{p-1} - 1$. Si $x_1 \in F_p^*$, alors $f'(x_1) = (p-1)x_1^{p-2} \neq 0$. On peut donc appliquer le lemme de Hensel. Il existe $x \in \mathbb{Z}_p$ unique tel que $\varepsilon_1(x) = x_1$ et $f(x) = 0$. L'ensemble V des racines de $f(X) = X^{p-1} - 1 = 0$ dans $\mathbb{Q}_p$ a donc au moins $(p-1)$ éléments. Il n'y a pas d'autre racine car $\mathbb{Q}_p$ est un corps commutatif.
L'homomorphisme de réduction $\varepsilon_1 : \mathbb{Z}_p \to F_p$ induit un isomorphisme de groupes de V sur F_p^*. Par suite V est un sous-groupe cyclique d'ordre $p - 1$ de $\mathbb{U}$. ∎

Exercice 21.
Montrer que l'on a un isomorphisme de groupes $\mathbb{U} \approx V \times \mathbb{U}_1$, où $\mathbb{U}_1 := 1 + p\mathbb{Z}_p$ est un sous-groupe de $\mathbb{U}$.
Solution. L'homomorphisme de réduction $\varepsilon_1 : \mathbb{Z}_p \to F_p$ induit un isomorphisme de groupes de V sur F_p^*. Soit $x \in U$, on a $x_1 := \varepsilon_1(x) \in F_p^*$. D'après la proposition ci-dessus, Il existe $v \in V$ unique tel que $\varepsilon_1(v) = x_1$. Considérons $u := v^{-1}x$, on a $\varepsilon_1(u) = \bar{1}$, donc $u \in \mathbb{U}_1$. On a $x = vu$. On vérifie que $x \in \mathbb{U} \mapsto (v, u) \in V \times \mathbb{U}_1$ est un isomorphisme de groupes.

Proposition 45. Soit p un nombre premier.

(i) L'ensemble des racines de l'unité dans $\mathbb{Q}_p$ est un sous-groupe μ de $\mathbb{U}$.

(ii) Si $p \neq 2$, les racines de l'unité forment un *groupe cyclique d'ordre* $(p-1)$.

(iii) Le groupe des racines de l'unité de $\mathbb{Q}_2$ est $\{\pm 1\}$

Démonstration.

(i) Les racines de l'unité sont les éléments de $\mathbb{Z}_p^*$ tels que $x^m = 1$ pour un entier $m \in \mathbb{N}^*$ convenable. Si $x^m = 1$, $1 = |x^m|_p = |x|_p^m$ et $|x|_p = 1$, donc $x \in \mathbb{U}$.

On vérifie que, dans un groupe multiplicatif abélien quelconque, l'ensemble de tels éléments, pour $m \in \mathbb{N}^*$ arbitraire, forment un sous-groupe.

(ii) Supposons $p \neq 2$. On considère l'homomorphisme de réduction $\varepsilon_1 : \mathbb{Z}_p \to F_p$. Il induit un *homomorphisme* de groupes multiplicatifs : $\eta : \mu \to F_p^*$. D'après la proposition 44, cet homomorphisme est surjectif. Montrons qu'il est injectif. Soit x dans le noyau de η. Il s'écrit $x = 1 + py$, avec $y \in \mathbb{Z}_p$ et il existe $n \in \mathbb{N}^*$ tel que $x^n = 1$. En utilisant la formule du binôme, on obtient :

$$y\left(n + \binom{n}{2}py + \cdots + p^{n-1}y^{n-1}\right) = 0.$$

Donc $y = 0$ ou p divise n. Dans le second cas, on écrit $(x^p)^{n/p} = x^n = 1$ et on remplace x par x^p et n par n/p. En répétant l'opération, on se ramène à traiter le cas $n = p$. Alors, puisque $p \geqslant 3$, $p + \binom{p}{2}py + \cdots + p^{p-1}y^{p-1} = p + p^2(\ldots) \neq 0$ et $y = 0$. Ainsi, dans tous les cas $x = 1$ et η est un *isomorphisme*. Par suite μ est le groupe des racines $(p-1)^{\text{èmes}}$ de l'unité.

(iii) Supposons $p = 2$. On a $-1 = 1 + 2 + 2^2 + \cdots \in 1 + 2\mathbb{Z}_2$, donc $\{\pm 1\} \subset 1 + 2\mathbb{Z}_2$. Par ailleurs $F_2^* = \{1\}$ et les racines de l'unité dans $\mathbb{Z}_2$ ont nécessairement pour ordre une puissance de 2. Mais, dans $\mathbb{Z}_2$, -1 n'est pas un carré et 1 n'a pas de racine quatrième non triviale. Nous admettons cela (on peut le déduire des résultats de l'exercice III.3.28 de la page 526). ■

Exercice 22.

Soit p un nombre premier impair. Quels sont les corps $\mathbb{Q}_p$ qui contiennent $\sqrt{-1}$?

Solution. Le nombre $p-1$ est pair, donc -1 est racine de l'unité dans $\mathbb{Q}_p$. Le nombre -1 a une racine carrée dans $\mathbb{Q}_p$ si, et seulement si, $(p-1)/2$ est pair, c'est-à-dire si, et seulement si, $p \equiv 1 \mod 4$.

Exercice 23.

Soient p et ℓ deux nombres premiers. Montrer que les corps $\mathbb{Q}_p$ et $\mathbb{Q}_\ell$ sont isomorphes si, et seulement si, $p = \ell$.

Solution. La condition $p = \ell$ est évidemment suffisante. Montrons qu'elle est nécessaire. On suppose que $\mathbb{Q}_p$ et $\mathbb{Q}_\ell$ sont isomorphes. Alors, les groupes multiplicatifs des racines de l'unité de $\mathbb{Q}_p$ et $\mathbb{Q}_\ell$ sont isomorphes. On distingue trois cas.

- On suppose p et ℓ *impairs*. Alors, d'après la proposition 45 ci-dessus, les groupes cycliques d'ordres $p-1$ et $\ell-1$ sont isomorphes. Par suite $p = \ell$.

- On suppose $p = 2$ et ℓ impair, $\ell \neq 3$. Alors, d'après la proposition 45 les groupes cycliques d'ordres 2 et $\ell - 1$ sont isomorphes. Par suite $p = \ell$.
- Il reste le cas $p = 2$ et $\ell = 3$. On montre que les corps $\mathbb{Q}_2$ et $\mathbb{Q}_3$ ne sont pas isomorphes en considérant les carrés : $\mathbb{Q}_3^*/(\mathbb{Q}_3^*)^2$ est d'ordre 4 et $\mathbb{Q}_2^*/(\mathbb{Q}_2^*)^2$ est d'ordre 8 (*cf.* l'exercice III.3.28 et l'exercice III.3.29 de la page 527).

2.4 Nombres automorphes Nombres b-adiques et nombres p-adiques

L'étude des nombres automorphes en base 10 conduit naturellement à un isomorphisme entre l'anneau $\mathbb{Z}_{10}$ des nombres entiers décadiques et l'anneau produit $\mathbb{Z}_5 \times \mathbb{Z}_2$. Plus généralement on peut montrer que l'anneau $\mathbb{Z}_b$ est isomorphe à l'anneau produit $\prod_{p \in \mathcal{P}_b} \mathbb{Z}_p$ où $\mathcal{P}_b$ est l'ensemble des diviseurs premiers de b. On ramène ainsi l'étude des nombres b-adiques à celle des nombres p-adiques. Ces résultats seront établis dans la partie 2.4.2. Le lecteur peut, en première lecture, sauter la partie 2.4.1, et y revenir ultérieurement pour mieux comprendre l'isomorphisme : $\mathbb{Z}_{10} \to \mathbb{Z}_5 \times \mathbb{Z}_2$.

2.4.1 Nombres automorphes en base 10

Les nombres automorphes apparaissent dans des publications de mathématiques récréatives[19]. Leur introduction est attribuée aux mathématiciens français Gergonne et Lucas. Depuis les années 50, l'intérêt pour les nombres automorphes a été renouvelé par l'apparition des calculatrices, puis des ordinateurs. De très nombreux auteurs les ont étudiés en n'utilisant que des mathématiques très élémentaires et divers outils de calcul. Nous allons voir que les nombres automorphes conduisent naturellement à des exemples d'écritures décimales infinies à gauche, c'est à dire à des nombres décadiques, et à des sommes et multiplications de ces écritures au sens que nous avons défini. Nous montrerons ensuite comment l'étude de l'anneau Z_{10} permet d'élucider complètement les propriétés des nombres automorphes. Cette étude se ramène à celles de $\mathbb{Z}_5$ et $\mathbb{Z}_2$ et il y a un résultat similaire pour tous les anneaux $\mathbb{Z}_b$ en utilisant les diviseurs premiers de b.

Les nombres automorphes sont les entiers naturels n qui, dans le système décimal, ont la propriété d'écriture suivante : n^2 se termine par n (plus précisément, l'écriture décimale de n^2 se termine par celle de n). Les nombres 0 et 1 sont évidemment automorphes, mais il y en d'autres, par exemple : 5, 6, 25, 76, 376, 625 ($5^2 = 2\underline{5}$, $6^2 = 3\underline{6}$,$25^2 = 6\underline{25}$, $76^2 = 57\underline{76}$, $376^2 = 141\underline{376}$, $625^2 = 390\underline{625}$). Il y a bien sûr des variantes si l'on change de base de numération, en remplaçant 10 par b. Le site OEIS[20] consacré aux suites remarquable d'entiers donne de très grands nombres automorphes. Plus modestement :

$n := 1665095808638110005574234232308961090041066199773922562599182128906 25$, est un nombre automorphe dont le dernier chiffre est 5. En utilisant la définition des nombres automorphes, on obtient, par coupures à gauche, une suite finie de nombres automorphes se terminant tous par 5, avec $1, 2, 3, 4, 5, \ldots$ chiffres : 5, 25, 625, 0625, $90625, \ldots$.

On note $s_1 := 5$, $s_2 := 25$, $s_3 := 625$, $s_4 := 625$, $s_5 = 90625$. On pose $t_1 := 6$, $t_2 := 76$, $t_3 := 376$, $t_4 := 9376$, $t_5 := 09376$. On vérifie que les t_n sont des nombre automorphes se terminant tous par 6.

---

[19] *Recreational Mathematics Magazine*, *The Fibonacci Quarterly*, les articles *Mathematical Games* de Martin Gardner. . ..

[20] The On-Line Encyclopedia of Integer Sequences, de N.J.A. Sloane ; *cf.* “sequences" A018247, A018248.

Proposition 46. Il existe deux suites de nombres automorphes $(s_n)_{n\in\mathbb{N}^*}$ et $(t_n)_{n\in\mathbb{N}^*}$ telles que, pour tout $n \in \mathbb{N}^*$:

- s_n se termine par 5 et t_n se termine par 6 ;
- s_n (resp. t_n) est défini en écriture décimale par les n derniers chiffres de s_{n+1} (resp. t_{n+1}) ;
- $s_n + t_n = 1 + 10^n$.

Démonstration.

1. On construit d'abord la suite (s_n) par récurrence. On remarque que tous les s_n sont nécessairement divisibles par 5. On définit $s_1 := 5$. On cherche s_2 sous la forme $s_2 = 10k + s_1$, avec $k \in [\![1, 9]\!]$. On doit avoir $(10k + 5)^2 = 10k + 5 \mod 10^2$, c'est-à-dire $10^2k^2 + 10^2k + 5^2 = 10k + 5 \mod 10^2$, ou $10k = 25 - 5 = 20$ et $k = 2$. Ainsi il y a une solution unique pour s_2 : $s_2 = 25$. Supposons s_n connu, il est divisible par 5. On cherche s_{n+1} sous la forme $s_{n+1} = 10^nk + s_n$, avec $k \in [\![1, 9]\!]$. On doit avoir $(10^nk + s_n)^2 = 10^nk + s_n \mod 10^{n+1}$, c'est-à-dire $10^{2n}k^2 + (2s_n)10^n + s_n^2 = 10^nk + s^n \mod 10^{n+1}$. Puisque $2s_n$ est divisible par 10, c'est équivalent à $10^nk = s_n^2 - s_n$. Il y a une solution $k \in [\![1, 9]\!]$ unique. Ainsi il y a une solution unique pour s_{n+1}.
2. On construit une suite (t_n) unique en imposant, pour tout $n \in \mathbb{N}$ les conditions :
   - t_n a n chiffres (le ou les premiers chiffres à gauche pouvant être nuls) ;
   - $s_n + t_n = 1 + 10^n$. Cette suite répond aux conditions. Le seul point non trivial est qu'elle est automorphe. Montrons le. On a modulo 10^n :
   $$\begin{aligned} t_n^2 &= (1 + 10^n - s_n)^2 = 1 + 10^{2n} + 2.10^n - 2.10^n s_n + s_n^2 - 2s_n \\ &= 1 + s_n^2 - 2s_n = 1 - s_n, \end{aligned}$$
   donc $t_n^2 = t_n \mod 10^n$.

   Les suites d'entiers naturels (s_n) et (t_n) sont cohérentes, au sens de la remarque de la page 441 (*cf.* aussi la partie 2.1). Elles définissent donc des éléments s et t de Z_{10}. On a $s^2 = s$, $t^2 = t$ et $s + t = 1$. On a :
   $$1 = (s + t)^2 = s^2 + t^2 + 2st = s + t + 2st = 1 + 2st,$$
   donc $2st = 0$. On en déduit que $st = 0$. Donc l'anneau Z_{10} n'est pas intègre. Nous en avons donné plus haut une preuve « abstraite » (*cf.* l'exercice 12 de la page 458). Nous avons maintenant un exemple explicite de diviseurs de zéro. ■

Observations. On a $s \equiv 1 \mod 2$ et $s \equiv 0 \mod 5$. Plus généralement, on constate que, « pour les premières valeurs de n », $s \equiv 1 \mod 2^n$ et $s \equiv 0 \mod 5^n$. On fait des observations analogues pour t en échangeant les rôles de 2 et 5. On est ainsi conduit à penser que les écritures dyadiques de s et t sont $\ldots 00 \ldots 001$ et $\ldots 00 \ldots 000$ tandis que leurs écritures 5-adiques sont $\ldots 00 \ldots 000$ et $\ldots 00 \ldots 001$. Nous montrerons que c'est bien le cas dans le prochain paragraphe en étudiant l'application naturelle $\mathbb{Z}_{10} \mapsto \mathbb{Z}_2 \times \mathbb{Z}_5$.

Les décadiques s et t sont des solutions dans Z_{10} de l'équation algébrique $X^2 = X$. Nous montrerons qu'avec les solutions évidentes 0 et 1 ce sont les seules. Au préalable nous allons établir quelques résultats généraux.

2.4.2 Nombres b-adiques et nombres p-adiques

Nous allons montrer que l'étude des nombres b-adiques se ramène à celle des nombres p_i-adiques où les p_i sont les facteurs premiers de b. Si $b = 10$, cette réduction est étroitement liée aux nombres automorphes. On peut associer à un nombre décadique « ses développements en base 5 et en base 2 », c'est-à-dire un nombre pentadique et un nombre dyadique. On obtient ainsi une application $\varpi : \mathbb{Z}_{10} \to \mathbb{Z}_5 \times \mathbb{Z}_2$. Nous vérifierons que c'est un isomorphisme d'anneaux. Nous l'utiliserons pour élucider le mécanisme des suites automorphes.

Plaçons nous dans un cadre plus général. Soit b un entier, $b \geqslant 2$. Si $b' \geqslant 2$ divise b, on a, pour tout $n \in \mathbb{N}^*$, un homomorhisme naturel surjectf d'anneaux :

$$\tau_n : A_n := \mathbb{Z}/b^n\mathbb{Z} \to A'_n := \mathbb{Z}/(b')^n.$$

On a, pour tout $n \in \mathbb{N}$, des homomorphismes surjectifs d'anneaux :

$$\varepsilon_n : Z_b = A = \varprojlim A_n \to \mathbb{Z}/b^n = A_n \quad \text{et} \quad \varepsilon'_n : Z_{b'} = A' = \varprojlim A'_n \to \mathbb{Z}/(b')^n = A'_n.$$

Par composition, on obtient, pour tout $n \in \mathbb{N}$, des homomorphismes surjectifs d'anneaux, $\tau_n \circ \varepsilon_n : Z_b \to A'_n$. On vérifie que la famille des $\tau'_n := \tau_n \circ \varepsilon_n$ permet de définir un homomorphisme d'anneaux $\tau' : \mathbb{Z}_b \to \mathbb{Z}'_b$. C'est l'unique homomorphisme d'anneaux $\tau' : Z_b \to Z_{b'}$, tel que, pour tout $n \in \mathbb{N}$, l'on ait $\varepsilon'_n \circ \tau' = \tau'_n \circ \varepsilon_n$.
On identifie $\mathbb{Z} \subset Z_b$ et $\mathbb{Z} \subset Z_{b'}$. L'homomorphisme τ' induit l'identité sur $\mathbb{Z}$ (il envoie un entier naturel écrit en base b sur ce même entier écrit en base b') et il est *continu*. C'est (par densité) l'unique homomorphisme continu d'anneaux de Z_b dans $Z_{b'}$ induisant l'identité sur $\mathbb{Z}$.
Soient $b', b'' \geqslant 2$, premiers entre eux et tels que $b := b'b''$, d'après la formulation « moderne » du théorème chinois (*cf.* le théorème 52 de [L1, II.2, 3.4, page 149]), pour tout $n \in \mathbb{N}^*$, l'homomorphisme naturel d'anneaux :

$$\varpi_n : \mathbb{Z}/b^n\mathbb{Z} \to \mathbb{Z}/(b')^n\mathbb{Z} \times \mathbb{Z}/(b'')^n\mathbb{Z}. \tag{8}$$

est un *isomorphisme*.
En remplaçant b' par b'' on obtient un unique homomorphisme continu d'anneaux $\tau'' : Z_b \to Z_{b''}$ prolongeant l'identité. En utilisant le fait que les ϖ_n sont des isomorphismes, on vérifie que l'homomorphisme :

$$\varpi := (\tau', \tau'') : Z_b \to \mathbb{Z}/(b') \times \mathbb{Z}/(b'')$$

est *un isomorphisme*.
Rappelons que *l'application diagonale* de $\mathbb{Z}$ dans $\mathbb{Z} \times \mathbb{Z}$ est définie par $m \mapsto (m, m)$. Compte tenu de ce qui précède, on vérifie le résultat suivant.

Proposition 47. Soient $b', b'' \geqslant 2$ deux nombres entiers premiers entre eux et tels que $b := b'b''$. Il existe un unique homomorphisme continu d'anneaux $\varpi : Z_b \to \mathbb{Z}_{b'} \times \mathbb{Z}_{b''}$, prolongeant l'application diagonale $\mathbb{Z} \to \mathbb{Z} \times \mathbb{Z}$. De plus ϖ est un *isomorphisme* et un *homéomorphisme*.

En utilisant la proposition, une récurrence, puis le résultat de l'exercice 7 de la page 449, on obtient le résultat suivant.

Corollaire 48.

(i) Soient $b \in \mathbb{N}$ et $b_1, \ldots, b_r \geqslant 2$ premiers entre eux deux à deux, et tels que $b = b_1 \ldots b_r$. Il existe un unique homomorphisme continu d'anneaux $\varpi : Z_b \to \mathbb{Z}_{b_1} \times \cdots \times \mathbb{Z}_{b_r}$, prolongeant l'application « diagonale » $\mathbb{Z} \to \mathbb{Z} \times \cdots \times \mathbb{Z}$. De plus ϖ est un *isomorphisme* et un *homéomorphisme*.

(ii) Soient $b \geqslant 2$ et $\mathcal{P}_b$ l'ensemble des diviseurs premiers de b. Il existe un unique homomorphisme continu d'anneaux $\varpi_b : Z_b \to \prod_{p \in \mathcal{P}_b} \mathbb{Z}_p$ prolongeant l'application diagonale. De plus ϖ_b est un *isomorphisme* et un *homéomorphisme*.

L'étude des nombres b-adiques de E_b est ainsi ramenée à celle des nombres p-adiques de $\mathbb{Q}_p$, p étant premier. L'anneau $\mathbb{Z}_b$ est intègre si, et seulement si, b est la puissance d'un nombre premier.

Remarque. On est conduit ainsi à s'intéresser aux anneaux[21] $\hat{\mathbb{Z}} := \prod_{p \in \mathcal{P}} \mathbb{Z}_p$ (où $\mathcal{P}$ est l'ensemble des nombre premiers) et $\mathbb{A}_{\mathbb{Z}} := \mathbb{R} \times \hat{\mathbb{Z}}$. Le lecteur vérifiera à titre d'exercice que l'on peut identifier $\mathbb{Z}$ à un sous-anneau de $\mathbb{A}_{\mathbb{Z}}$.

2.4.3 Retour aux nombres automorphes

L'anneau des nombres décadiques $\mathbb{Z}_{10}$ est isomorphe, par l'application canonique $\varpi = \varpi_{10}$, au produit d'anneaux $\mathbb{Z}_2 \times \mathbb{Z}_5$. Les anneaux $\mathbb{Z}_2$ et $\mathbb{Z}_5$ sont *intégres*, mais $\mathbb{Z}_2 \times \mathbb{Z}_5$ ne l'est pas : $(1,0)$ et $(0,1)$ sont des diviseurs de 0 dans $\mathbb{Z}_2 \times \mathbb{Z}_5$, puisque $(1,0)(0,1) = (0,0)$, Notons que $\mathbb{Z}_2 \times \mathbb{Z}_5$ contient le sous-anneau : $\{(x,x) \mid x \in \mathbb{Z}\}$ qui est isomorphe à $\mathbb{Z}$, donc intègre, mais il contient aussi le sous-anneau plus grand : $\{x,y) \mid x,y \in \mathbb{Z}\}$ qui est isomorphe à $\mathbb{Z}^2$ et qui n'est pas intègre. Notons $\tilde{s} := \varpi^{-1}(1,0)$ et $\tilde{t} := \varpi^{-1}(0,1)$. On vérifie : $\tilde{s}^m = s$, $\tilde{t}^m = t$ $(m \in \mathbb{N}^*)$, $\tilde{s} + \tilde{t} = 1$, $\tilde{s}\tilde{t} = 0$.

Nous allons voir que nous avons en fait déjà rencontré $\tilde{s}$ et $\tilde{t}$ quand nous avons étudié les nombres décimaux *automorphes* : nous allons montrer que $\tilde{s} = s$ et $\tilde{t} = t$.

Dans $\mathbb{Z}_{10}$, l'on a : $\tilde{s}^2 = \tilde{s}$, $\tilde{t}^2 = \tilde{t}$, $1^2 = 1$ et $0^2 = 0$. On a ainsi quatre solutions de l'équation algébrique $X^2 = X$. Les solutions correspondantes dans $\mathbb{Z}_2 \times \mathbb{Z}_5$ sont : $(1,0)$, $(0,1)$, $(1,1)$ et $(0,0)$ et ce sont les seules. En effet 2 et 5 étant premiers, les anneaux $\mathbb{Z}_2$ et $\mathbb{Z}_5$ sont *intègres* et dans chaque corps commutatif $\mathbb{Q}_2$, $\mathbb{Q}_5$, l'équation $X^2 - X = 0$ a exactement deux solutions 0 et 1. Nous connaissons aussi, par ailleurs, deux solutions de $X^2 = X$, les nombres décadiques automorphes s et t. On a donc nécessairement $s = \tilde{s}$ et $t = \tilde{t}$ ou $s = \tilde{t}$ et $t = \tilde{s}$. On décide en considérant le premier chiffre.

Dans $\mathbb{Z}/10\mathbb{Z}$, l'équation $X^2 = X$ a exactement quatre racines $\bar{5}$, $\bar{6}$, $\bar{1}$ et $\bar{0}$. Ce sont respectivement les images de $\tilde{s}$, $\tilde{t}$, 1 et 0 par la projection $\varepsilon_1 : \mathbb{Z}_{10} \to \mathbb{Z}/10\mathbb{Z}$. On en déduit que $\tilde{s} = s$ et $\tilde{t} = t$.

Le nombre s engendre un idéal principal $(s) = s\mathbb{Z}_{10}$ de $\mathbb{Z}_{10}$. La multiplication par s dans $\mathbb{Z}_{10}$ est un *projecteur* de $\mathbb{Z}_{10}$ sur (s). On a des résultats similaires pour t, en notant $(t) = t\mathbb{Z}_{10}$. Puisque $s + t = 1$, l'anneau $\mathbb{Z}_{10}$ est *la somme directe* des idéaux (s) et (t). Puisque $st = 0$, si $x \in (s)$ et $y \in (t)$, alors $xy = 0$. On a une analogie avec des projecteurs orthogonaux.

[21] On peut interpréter $\hat{\mathbb{Z}}$ comme la limite projective des $\mathbb{Z}/n\mathbb{Z}$, $n \in \mathbb{N}^*$, et l'on appelle $\mathbb{A}_{\mathbb{Z}}$ un anneau adélique. Cet anneau et diverses généralisations sont en relation avec la *transformation de Fourier* et la *fonction zeta de Riemann*.

Un peu d'Histoire

Les nombres p-adiques A la suite des travaux de Dedekind et Weber, le mathématicien allemand Hensel a poursuivi l'analogie entre la surface de Riemann d'un corps de fonctions algébriques (i.e. d'une extension algébrique finie de $\mathbb{C}(X)$) et l'ensemble des idéaux premiers d'un corps de nombres (i.e. d'une extension algébrique finie de $\mathbb{Q}$). Comme pour les hiéroglyphes égyptiens, on a une « pierre de Rosette » : d'un côté le plan complexe, avec ses points a et les idéaux $(X-a)$ de $\mathbb{C}[X]$ (et, plus généralement, les extensions finies de $\mathbb{C}(X)$ et les courbes complexes associées) et de l'autre $\mathbb{Z}$ et ses idéaux $(p) := p\mathbb{Z}$ (et, plus généralement, les extensions finies de $\mathbb{Q}$). Hensel a introduit la notion de nombre p-adique en 1897, dans le cadre de l'étude de la factorisation de certains polynômes de $\mathbb{Z}[X]$ modulo un entier de la forme p^k (p premier). Il a ainsi défini $\mathbb{Z}_p$ et $\mathbb{Q}_p$. Plus tard (vers 1907), il a introduit un point de vue *topologique* avec la définition de la distance ultramétrique d_p sur $\mathbb{Q}_p$ (cette approche ne s'est vraiment clarifiée qu'avec les travaux de Fréchet et Riesz en topologie). En 1913, l'un des élèves de Hensel a découvert que, de même que $\mathbb{R}$ peut-être considéré comme le complété de $\mathbb{Q}$ pour la distance ordinaire (Cantor), $\mathbb{Q}_p$ est le complété de $\mathbb{Q}$ pour la distance d_p. L'étape suivante est, en 1918, le théorème d'Ostrowski donnant toutes les valeurs absolues sur $\mathbb{Q}$. En 1935 Ostrowski a généralisé son théorème aux corps de nombres algébriques. Un peu avant, Hasse et Schmidt avaient montré que si un corps est *localement compact* pour une topologie définie par une valeur absolue ultramétrique, alors c'est une extension de degré fini, ou bien d'un corps p-adique $\mathbb{Q}_p$, ou bien d'un corps de séries entières formelles $F_p((X)) = F_p[[X]][X^{-1}]$.
Le théorème d'Ostrowski suggère de « mettre sur le même plan » les complétés $\mathbb{R}$ et $\mathbb{Q}_p$ (p étant un nombre premier arbitraire). Ce point de vue apparaît naturellement dans la résolution de certaines *équations diophantiennes* à plusieurs variables. Considérons par exemple une *forme quadratique* $q : x = (x_1, \ldots, x_n) \mapsto q(x)$ à coefficients *entiers*, si l'équation $q(x) = 0$ a une solution non triviale dans $\mathbb{Z}^n$, elle a évidemment une solution non triviale dans $\mathbb{R}^n$ et une solution non triviale modulo p^k pour tout nombre premier p et tout entier k. Il revient au même de dire que $q = 0$ admet une solution dans $\mathbb{R}^n$ et dans tous les $\mathbb{Q}_p^n$. Le théorème dit de Hasse-Minkowski (Hasse 1923) est la réciproque de ce résultat. On en déduit un résultat similaire pour la classification des formes quadratiques (*cf.* le théorème 47 de la page 112, et pour plus de détails, le livre [Serre]). On ramène ainsi la preuve d'un résultat sur $\mathbb{Q}$ dit *global* à celle des résultats analogues sur $\mathbb{R}$ *et tous* les $\mathbb{Q}_p$ dits « locaux » $\big($la terminologie vient de l'analogie entre $\mathbb{Q}$ et $\mathbb{C}(z)\big)$, c'est le « principe du passage du local au global » ou principe de Hasse. On a l'impression d'avoir remplacé une question simple par une question compliquée, en fait on a « gagné » quelque chose d'essentiel : $\mathbb{R}$ et les $\mathbb{Q}_p$ sont *complets* et, de ce fait, le problème local est beaucoup plus simple à résoudre : par la *méthode de Newton* dans le cas réel et par son analogue p-adique, le *lemme de Hensel*, dans les autres cas (*cf.* la proposition 41).
On doit aussi à Hensel la définition des fonctions exponentielle p-adique et logarithme p-adique, à partir des développements en série usuels. Divers auteurs ont ensuite développé l'analyse p-adique. Dans ce cadre, on peut parler, nous l'avons vu, de dérivation et de série entière convergente. De la même façon qu'il est préférable d'étudier les séries entières classiques sur $\mathbb{C}$, il est préférable d'étudier les séries entières p-adiques sur $\mathbb{C}_p$ (la complétion de la clôture algébrique de $\mathbb{Q}_p$).

Un peu d'Histoire

La notion d'équation différentielle p-adique apparaît pour la première fois dans les travaux de la mathématicienne française Élisabeth Lutz[a] en 1937. Ce sujet a été ensuite développé par le mathématicien américain B. Dwork (en relation avec le célèbre problème des *conjectures de Weil*), puis par de nombreux auteurs.
Au début, les idées de Hensel semblent avoir été peu appréciées. Mais ultérieurement, son approche a eu tellement de succès qu'elle a changé en profondeur la théorie algébrique des nombres. Les nombres p-adiques ont pris au cours des années de plus en plus d'importance en arithmétique et ce mouvement se poursuit vigoureusement[b], avec en particulier l'étude des représentations p-adiques des groupes de Galois, Par exemple, la preuve du « grand théorème de Fermat » , par Andrew Wiles[c], utilise de façon essentielle (parmi bien d'autres choses...) des représentations de $\mathrm{Gal}\,\overline{\mathbb{Q}}/\mathbb{Q}$ (le groupe des automorphismes d'une clôture algébrique de $\mathbb{Q}$) dans $GL(2;\mathbb{Z}_\ell)$ (où ℓ est un nombre premier).
Les relations des nombre p-adiques avec le « monde réel » sont moins claires, malgré des tentatives vers une physique p-adique (à l'échelle de Planck).

[a] La motivation est la définition de fonctions elliptiques p-adiques.
[b] En 2018 le jeune mathématicien allemand Peter Scholze a obtenu la médaille Fields pour des travaux « fascinants » sur les nombres p-adiques.
[c] Cette preuve est très difficile. Le lecteur intéressé pourra lire la présentation assez abordable de Guy Henniart : *Fermat,Wiles et GL*(2), Journées x-ups 2009, 01 ; www.math.polytechnique.fr/xups/xups09-01

3 Coupures et sections commençantes

3.1 Construction de $\mathbb{R}$ par les sections commençantes

Après un long détour, revenons aux constructions de $\mathbb{R}$. La nécessité de donner un sens à la longueur de la diagonale d'un carré de côté de longueur 1 (un problème posé depuis l'antiquité) met en évidence la notion de « trou » dans le corps ordonné $\mathbb{Q}$: si l'on « visualise » les rationnels comme des points d'une « droite géométrique continue », alors il y a un « trou à la place de $\sqrt{2}$ ». L'idée de Dedekind est de formaliser la notion de trou dans $\mathbb{Q}$ (sous forme de *coupure*), puis de dire *qu'un trou est un nombre*. En ajoutant à l'ensemble $\mathbb{Q}$ l'ensemble de ses trous, on obtient l'ensemble $\mathbb{R}$. C'est une méthode basée sur la relation d'ordre total de $\mathbb{Q}$ et une idée heuristique de « continuité » du corps cherché, dont *on exige* qu'il soit totalement ordonné. Ceci permet de formaliser l'idée de « continuité » par la propriété de la borne supérieure.
Avec cette approche, il est facile de définir l'ensemble $\mathbb{R}$, d'étendre à cet ensemble la relation d'ordre de $\mathbb{Q}$ et de montrer la propriété de la borne supérieure. Les définitions de l'addition et de la multiplication sur $\mathbb{R}$ sont assez naturelles. Malheureusement, les justifications de ces définitions et l'étude des propriétés de ces opérations sont un peu techniques, avec de *très nombreuses vérifications*. Nous ne détaillerons que les points non immédiats en insistant sur les quelques preuves délicates. Un certain nombre de vérifications, aussi élémentaires que fastidieuses, seront laissées au lecteur (consciencieux...) pour ne pas alourdir sans réelle nécessité[22] la présentation. Le lecteur constatera que, dans la construction, les

[22] Peu d'ouvrages présentent de façon raisonnablement détaillée la construction de $\mathbb{R}$ par les coupures ou leurs variantes. On peut citer [Ru] et l'ancêtre de [R3] : Cagnac, Ramis, Commeau, *Traité de mathématiques spéciales.*

réels irrationnels sont comme l'arlésienne d'Alphonse Daudet[23], on en parle mais on ne les voit jamais vraiment. Toutes les preuves sont des manipulations de sous-ensembles de $\mathbb{Q}$ et elles reposent en général sur la théorie des ensembles.

Les deux autres constructions de $\mathbb{R}$ conduisent, par analogie, aux corps p-adiques $\mathbb{Q}_p$. Il n'en est pas de même de la construction à partir des coupures. Ce n'est pas surprenant car les corps $\mathbb{Q}_p$ ne sont pas ordonnables. Par contre, vers 1969, John Horton Conway, a proposé une vaste et élégante généralisation des nombres réels en utilisant *une version généralisée des coupures de Dedekind*[24]. Le point de départ de Conway est l'analyse des jeux, en particulier du jeu de Go. Les nombres de Conway ont été appelés *nombres surréels* par Donald Knuth[25].

3.1.1 Définitions

Définition 10. Une *coupure* de l'ensemble totalement ordonné $\mathbb{Q}$ est la donnée d'un couple (A, B) de sous-ensembles *non vides* de $\mathbb{Q}$ tels que :

(i) $A \cup B = \mathbb{Q}$;

(ii) tout élément de A est *strictement inférieur* à tout élément de B.

Une coupure est dite *de Dedekind* si elle vérifie la condition suivante :

(iii) si B a une borne inférieure dans $\mathbb{Q}$, alors celle-ci est dans B.

Si (A, B) est une coupure, alors les deux ensembles forment une *partition* de $\mathbb{Q}$. La notion intuitive de « trou dans $\mathbb{Q}$ » correspond au cas où B n'a pas de borne inférieure dans Q.
Si (A, B) est une coupure, alors $B = \mathbb{Q} \setminus A$. On peut traduire les conditions sur le couple (A, B) en des conditions sur A seul. On obtient ainsi la notion de *section commençante*. On pourrait aussi utiliser B et l'on obtiendrait la notion de *section finissante*.
Un rationnel $r \in \mathbb{Q}$ permet de définir *deux* coupures :

- $A_1 := \{x \in \mathbb{Q} \mid x \leqslant r\}$, $B_1 := \{y \in \mathbb{Q} \mid y > r\}$;
- $A_2 := \{x \in \mathbb{Q} \mid x < r\}$, $B_2 := \{y \in \mathbb{Q} \mid y \geqslant r\}$.

La seconde (A_2, B_2) est une coupure de Dedekind.

Définition 11. On appelle *section commençante* de $\mathbb{Q}$ tout sous-ensemble *propre* A de $\mathbb{Q}$ (c'est-à-dire $A \neq \varnothing$ et $A \neq \mathbb{Q}$) tel que, pour tout $u \in A$ et pour tout $v \leqslant u$, on a $v \in A$. On dit qu'une section commençante A de $\mathbb{Q}$ est *ouverte* si elle n'a pas d'élément maximum dans $\mathbb{Q}$ (c'est-à-dire s'il n'existe pas de rationnel x tel que $A = \{u \in \mathbb{Q} \mid u \leqslant x\}$).

Exemples. Les sous-ensembles $A := \{u \in \mathbb{Q} \mid u < 3\}$ et :

$$A' = \{u \in \mathbb{Q} \mid 0 < u \text{ et } u^2 < 2\} \cup \{v \in \mathbb{Q} \mid v \leqslant 0\}$$

sont des section commençantes ouvertes ; A est un intervalle ouvert de $\mathbb{Q}$ mais A' n'en est pas un.

Analyse. Tome 2. Tous les auteurs éludent un certain nombre de détails.

[23] *L'arlésienne* est une courte nouvelle des *Lettres de mon moulin*, 1869. Georges Bizet en a tiré un opéra du même nom. On y parle beaucoup de l'arlésienne mais elle n'apparaît jamais sur la scène.

[24] J.H. Conway, *On Numbers and Games, Academic Press*, 1976.

[25] En 1974, D. Knuth a écrit une introduction élémentaire au sujet sous forme de dialogue, *How Two Ex-Students Turned on to Pure Mathematics and Found Total Happiness*.

De façon analogue, un sous-ensemble propre $B \subset \mathbb{Q}$ tel que pour tout $u \in B$ et pour tout $v \geqslant u$ on a $v \in B$ est appelé *section finissante*. Une telle section est dite *ouverte* si elle n'a pas d'élément minimal.
Soient A une section commençante et $v \in \mathbb{Q}$ n'appartenant pas à A. Alors v est un majorant de A. En effet, si ce n'était pas le cas, il existerait $w \in A$ tel que $v < w$. Alors v appartiendrait à A et l'on aurait une contradiction.
On montre facilement le résultat suivant qui donne le « dictionnaire » entre la notion de section commençantes et celle de coupure. Dans la suite nous utiliserons uniquement les sections commençantes.

Proposition 49. Soit $A \subset \mathbb{Q}$.

(1) Les deux conditions suivantes sont équivalentes :
   (i) A est une section commençante ;
   (ii) $(A, \mathbb{Q} \setminus A)$ est une coupure.

(2) Les deux conditions suivantes sont équivalentes :
   (i) A est une section commençante *ouverte* ;
   (ii) $(A, \mathbb{Q} \setminus A)$ est une coupure *de Dedekind*.

Définition 12. On appelle ensemble des nombres réels, noté $\mathbb{R}$, le sous-ensemble de l'ensemble $\mathcal{P}(\mathbb{Q})$ (des parties de $\mathbb{Q}$) formé des sections commençantes *ouvertes* de $\mathbb{Q}$.

Compte tenu de la proposition ci-dessus, $\mathbb{R}$ pourrait aussi être défini comme l'ensemble des coupures de Dedekind (un sous-ensemble de $\mathcal{P}(\mathbb{Q}) \times \mathcal{P}(\mathbb{Q})$). C'est la définition utilisée par Dedekind.
La condition *ouverte* est essentielle. Un rationnel définit *deux* sections commençantes, mais une seule est *ouverte*. Nous avons rencontré un problème similaire avec les écritures décimales : un rationnel est représenté par *deux* écritures décimales, mais une seule est *propre*.
On notera, suivant les cas, $x \in \mathbb{R}$ ou $A \in \mathbb{R}$ les réels. On notera en général A pour manipuler le réel comme une partie de $\mathbb{Q}$.

La première étape est « d'interprêter $\mathbb{Q}$ comme un sous-ensemble de $\mathbb{R}$ ».

A tout rationnel $r \in \mathbb{Q}$ on associe la partie $A_r := \{x \in \mathbb{Q} \mid x < r\}$. On vérifie facilement que A_r est une section commençante ouverte. L'ensemble $B_r := \mathbb{Q} \setminus A_r$ est une section finissante et elle admet r comme plus petit élément.
Le résultat suivant est immédiat : si $r, r' \in \mathbb{Q}$, avec $r \neq r'$, alors $B_r \neq B_{r'}$, donc $A_r \neq A_{r'}$.

Proposition 50. L'application $\iota : r \in \mathbb{Q} \mapsto A_r \in \mathbb{R}$ est *injective*.

Nous appellerons cette application ι *l'injection naturelle* de $\mathbb{Q}$ dans $\mathbb{R}$. En l'utilisant, on peut « identifier » $\mathbb{Q}$ à un sous-ensemble de $\mathbb{R}$. Il faudra manipuler cette identification avec soin pour éviter les confusions. Il faudra en particulier vérifier que la relation d'ordre et les opérations sur $\mathbb{Q}$ induites par celles définies sur $\mathbb{R}$ coïncident avec celles connues initialement sur $\mathbb{Q}$. Nous avons déjà rencontré ce problème dans la construction de $\mathbb{R}$ par les

écritures décimales, mais ici nous ne renvoyons pas l'identification à la fin de la construction pour ne pas alourdir exagérément la rédaction. Nous noterons donc indifféremment r ou A_r un rationnel. c'est un abus de notation, mais le contexte permettra d'éviter les confusions. Les sections commençantes qui ne sont pas de la forme A_r, $r \in \mathbb{Q}$, seront dites *irrationnelles*. Si une section commençante (resp. finissante) ouverte n'est pas de la forme A_r, avec $r \in \mathbb{Q}$, alors son complémentaire dans $\mathbb{Q}$ est une section finissante (resp. commençante) ouverte. On notera par exemple souvent 0 pour $A_0 = \mathbb{Q}^*_-$ ou 1 pour A_1.

3.1.2 Relation d'ordre sur $\mathbb{R}$ et propriété de la borne supérieure

La relation d'inclusion sur $\mathcal{P}$ est une relation d'ordre (non total). Elle induit une relation d'ordre sur $\mathbb{R}$.

Définition 13. On définit une relation d'ordre $\leqslant$ sur $\mathbb{R}$ par : pour tous $A, A' \in \mathbb{R}$, $A \leqslant A'$ si $A \subset A'$.

Cette relation d'ordre est compatible avec la relation d'ordre ordinaire sur $\mathbb{Q}$.
La notation $\leqslant$ ne pose donc pas de problème. Soient $x, y \in \mathbb{R}$. Si $x \leqslant y$ et $x \neq y$, on note $x < y$. Si $A \in \mathbb{R}$ et $r \in A$, on a $r < A$ (puisque A est ouverte). Si $r \in \mathbb{Q}$ n'appartient pas à A, il majore A. On écrira indifféremment $A \leqslant A'$ ou $A \subset A'$.

Proposition 51. La relation $\leqslant$ est une relation d'ordre *total* sur $\mathbb{R}$.

Démonstration. Soient $A, A' \in \mathbb{R}$. Ou bien $A \subset A'$ on bien $A \not\subset A'$. Dans ce second cas, il existe $x \in A$ tel que $x \notin A'$. Soit $y \in A'$. On a $y < x$: sinon $x \leqslant y$ et $x \in A'$, ce qui n'est pas. Ainsi $y \in A$ et nous avons montré que $A' \subset A$. ∎

Corollaire 52. Entre deux réels distincts x et x' il existe toujours un rationnel r : $x' < r < x$, ou $x < r < x'$ (i. e. l'ensemble $\mathbb{Q}$ est dense dans $\mathbb{R}$).

Démonstration. Soient $A, A' \in \mathbb{R}$ distincts.
Puisque $\mathbb{R}$ est totalement ordonné, on a $A' \subset A$ ou $A \subset A'$. Supposons $A' \subset A$. Puisque $A \neq A'$ il existe un rationnel $r \in A$ n'appartenant pas à A'. On a $A' < r < A$. ∎

Proposition 53. Toute partie non vide majorée de $\mathbb{R}$ admet une borne supérieure.

Démonstration. La preuve est facile. Soit $C \subset \mathbb{R}$ un sous-ensemble non vide et majoré. On note $C = \{A_i\}_{i \in I}$, où les $A_i \subset \mathbb{Q}$ sont des sections commençantes.
On définit $A := \bigcup_{i \in I} A_i \subset \mathbb{Q}$. Nous allons voir que A est une section commençante ouverte et que c'est la borne supérieure de C.

1. Montrons que $A \subset \mathbb{Q}$ est un sous-ensemble *propre* :
   - L'ensembre C n'est pas vide, donc I n'est pas vide et A n'est pas vide.
   - Soit $A_0 \in \mathbb{R}$ un majorant de C. On a $A \subset A_0$ et, puisque $A_0 \neq \mathbb{Q}$, $A \neq \mathbb{Q}$.
2. Montrons que A est une section commençante. Soient $x \in A$ et $y \leqslant x$ ($y \in \mathbb{Q}$). Il existe $j \in I$ tel que $x \in A_j$. Alors $y \in A_j \subset A$. Ainsi, compte tenu de 1, A est une section commençante.

3. Montrons que A est une section commençante *ouverte*, c'est-à-dire $A \in \mathbb{R}$. Soit $x \in A$ ($x \in \mathbb{Q}$). Montrons qu'il existe $y \in A$ ($y \in \mathbb{Q}$) tel que $x < y$. Il existe $i \in I$ tel que $x \in A_j$. Puisque A_j est une section commençante *ouverte*, il existe $y \in A_j$ tel que $x < y$. On a $y \in A_j \subset A$.
4. Montrons que A majore C et que c'est le plus petit des majorants. Pour tout $i \in I$, $A_i \subset A$, donc A majore C. Si A' majore C, pour tout $i \in I$ on a $A_i \subset A'$.
   Par suite $A \subset A'$: A' majore A. ■

3.1.3 L'addition des réels

Soient $C, C' \subset \mathbb{Q}$ non vides, on définit une partie $C + C' \subset \mathbb{Q}$ par :

$$C + C' := \{x + x' \in \mathbb{Q} \mid x \in C,\ x' \in \mathbb{C}'\}.$$

On l'appelle la *somme* de C et C'. Elle n'est pas vide.

L'addition

Si $A, A' \in \mathbb{R}$, on considère la somme des deux sous-ensembles de $\mathbb{Q}$: $A + A'$. Montrons que c'est une section commençante ouverte, c'est-à-dire un élément de $\mathbb{R}$. Ainsi l'addition des parties induit une loi de composition interne sur $\mathbb{R}$.

- Montrons que le sous-ensemble $A + A' \subset \mathbb{Q}$ est *propre*. Il n'est évidemment pas vide. Il reste à montrer $A + A' \neq \mathbb{Q}$. Puisque $A \neq \mathbb{Q}$ et $A' \neq \mathbb{Q}$, il existe $x, x' \in \mathbb{Q}$ tels que $x \notin A$, $x' \notin A'$. Posons $z := x + x'$ et montrons que $z \notin A + A'$, ce qui prouvera que $A + A' \neq \mathbb{Q}$. On raisonne par l'absurde. Supposons $z \in A + A'$. Alors il existe $y \in A$ et $y' \in A'$ tels que $z = y + y'$. Puisque $x + x' = y + y'$, l'une des inégalités $x \leqslant y$, $x' \leqslant y'$ est vraie. Si, par exemple, $x \leqslant y$, alors $x \in A$ et l'on a une contradiction.
- Montrons que $A + A'$ est une section commençante. Soient $z \in A + A'$ et $u \leqslant z$. On pose $\varepsilon := z - u \in \mathbb{Q}_+$. Il existe $x \in A$ et $x' \in A'$ tels que $z = x + x'$. On pose $v := x - \varepsilon/2$ et $v' := x' - \varepsilon/2$. On a $v \in \mathbb{Q}$ et $v \leqslant x$, donc $v \in A$. De même $v' \in A'$. Par suite $u = z - \varepsilon = v + v' \in A + A'$. Ainsi, puisque $A + A'$ est propre, c'est une section commençante.
- Montrons que $A + A'$ est une section commençante. *ouverte*. Soit $z \in A + A'$. Il existe $x \in A$ et $x' \in A'$ tels que $z = x + x'$. Il existe $y \in A$ et $y' \in A'$ tels que $x < y$ et $x' < y'$. On pose $w = y + y' \in A + A'$. Alors $z = x + x' < y + y' = w$. Ainsi $A + A'$ est une section commençante ouverte, c'est-à-dire un élément de $\mathbb{R}$.

Proposition 54. La somme de deux sections commençantes ouvertes de $\mathbb{Q}$ est une section commençante ouverte de $\mathbb{Q}$. On définit ainsi l'addition de $\mathbb{R}$. Elle est associative et commutative.

L'associativité se déduit de celle de l'addition de $\mathbb{Q}$, nous laissons les détails au lecteur. La commutativité est évidente.
D'après le résultat suivant les additions sur $\mathbb{Q}$ et $\mathbb{R}$ sont compatibles avec l'injection canonique de $\mathbb{Q}$ dans $\mathbb{R}$: $r \mapsto A_r$.

Lemme 55. Si $r, r' \in \mathbb{Q}$, on a $A_r + A_{r'} = A_{r+r'}$.

Démonstration. On a évidemment $A_r + A_{r'} \subset A_{r+r'}$. Montrons l'inclusion opposée. Soit $x \in A_{r+r'}$. On a $r + r' - x > 0$, donc $\frac{r+r'-x}{2} > 0$. Par suite :

$$y := r - \frac{r + r' - x}{2} < r \quad \text{et} \quad y := r' - \frac{r + r' - x}{2} < r',$$

donc $y \in A$ et $y' \in A'$. On peut écrire : $x = r - \frac{r+r'-x}{2} + r' - \frac{r+r'-x}{2} = y + y'$. On a ainsi prouvé l'inclusion $A_{r+r'} \subset A_r + A_{r'}$. ∎

Élement neutre de l'addition. Opposé d'un nombre réel. Structure de groupe additif sur $\mathbb{R}$

Montrons que $A_0 = \mathbb{Q}^*_-$ est élément neutre pour l'addition. Soit $A \in \mathbb{R}$. On a évidemment l'inclusion $A + A_0 \subset A$. montrons l'inclusion opposée. Soit $x \in A$. Puisque A est ouverte, il existe $y \in A$ tel que $x < y$. Alors $x - y < 0$, c'est-à-dire $x - y \in A_0$. En écrivant $x = y + (x - y)$, on en déduit que $x \in A + A_0$. On a ainsi prouvé l'inclusion $A \subset A + A_0$. Donc $A = A + A_0 = A_0 + A$.

La définition de l'opposé d'un nombre réel est plus délicate. On distingue deux cas. Soit $A \in \mathbb{R}$.

1. On a $A = A_r$, avec $r \in \mathbb{Q}$. On *définit* alors l'opposé $-A$ de la section commençante ouverte A par $-A := A_{-r}$.
2. On suppose que A n'est pas de la forme A_r, avec $r \in \mathbb{Q}$.
   On *définit* alors $-A$ par $-A := \{x \in \mathbb{Q} \mid -x \notin A\}$.

Il reste à prouver que $-A \in \mathbb{R}$, puis que $A + (-A) = A_0$. Pour le second point, nous utiliserons le lemme suivant.

Lemme 56. Soit $A \in \mathbb{R}$. Pour tout $t \in \mathbb{Q}$, $t > 0$, il existe $x \in A$ tel que $x + t \notin A$.

Démonstration. On raisonne par l'absurde. On suppose qu'il existe $t \in \mathbb{Q}$, $t > 0$, tel que, pour tout $x \in A$, l'on ait $x + t \in A$. On en déduit, par récurrence sur n, que, pour tout $n \in \mathbb{N}$, on a $x + nt \in A$. Fixons $x_1 \in A$. Il existe $z \notin A$ ($z \in \mathbb{Q}$), donc, pour tout $n \in \mathbb{N}$, on a $x_1 + nt \leqslant z$ et $n \leqslant \frac{z-x_1}{t}$, ce qui contredit le fait que $n \in \mathbb{N}$ est arbitraire. ∎

1. Montrons $-A \in \mathbb{R}$. Si $A = A_r$, $-A = A_{-r} \in \mathbb{R}$. Il reste à traiter le cas où A n'est pas de la forme A_r. On fait donc cette hypothèse.
   - Montrons que $-A \subset \mathbb{Q}$ est un sous-ensemble propre. Puisque $A \neq \mathbb{Q}$, il existe $x \in \mathbb{Q} \setminus A$. Si $x \notin A$, alors $-x \in -A$ et $-A$ n'est pas vide. Puisque A n'est pas vide, il existe $y \in A$. Si $y \in A$, $-y \notin -A$ et $-A \neq \mathbb{Q}$.

- Montrons que $-A$ est une section commençante. Soient $x \in -A$ et $y \leqslant x$. Ceci équivaut à $-x \notin A$ et $-x \leqslant -y$. Alors $-x$ est un majorant strict de A et $-y \geqslant -x$, donc $-y$ est un majorant strict de A et $-y \notin A$, donc $y \in -A$. Ainsi, en utilisant 1, $-A$ est une section commençante.
- Montrons que la section commençante $-A$ est *ouverte*. On raisonne par l'absurde. On suppose que $-A$ admet un plus grand élément $z \in \mathbb{Q}$. Puisque $-A$ est une section commençante, on a $-A = \{x \in \mathbb{Q} \mid x \leqslant z\}$. Dire que $x \in -A$ équivaut à dire que $-x \notin A$, donc la condition $-x \notin A$ est équivalente à $x \leqslant z$. Par négation, la condition $-x \in A$ est équivalente à $x > z$, c'est-à-dire à $-x < -z$. Alors $A = A_{-z}$, ce qui et contraire à l'hypothèse.

2. Montrons $A + (-A) = A_0$.
   - On suppose $A = A_r$, avec $r \in \mathbb{Q}$. Ce cas est presque évident.
     * Montrons $A_r + A_{-r} \subset A_0$. On a $x \in A_r$ (resp. $y \in -A_r$) si, et seulement si, $x < r$ (resp. $y < -r$). Donc, si $x \in A_r$ et $y \in A_{-r}$, on a $x + y < 0$, c'est-à-dire $x + y \in A_0$.
     * Montrons $A_0 \subset A_r + A_{-r}$. Soit $x_0 \in A_0$. Alors $0 < 0$ et $x_0/2 < 0$. On a $r + x_0/2 < r$, donc $r + x_0/2 \in A_r$. De même $-r + x_0/2 \in A_{-r}$. Donc $x_0 = (r + x_0/2) + (r + x_0/2) \in A_r + A_{-r}$.
   - On suppose que A n'est pas de la forme A_r.
     * Prouvons d'abord $A + (-A) \subset A_0$. Soient $x \in A$ et $y \in -A$. On a $-y \notin A$, donc $-y$ majore strictement A et l'on a $x < -y$, c'est-à-dire $x + y < 0$ et $x + y \in A_0$. Ainsi $A + (-A) \subset A_0$.
     * Il reste à prouver l'inclusion inverse $A_0 \subset A + (-A)$. Si $x_0 \in A_0$, il s'écrit $x_0 = -t$, avec $t > 0$. D'après le lemme 56 ci-dessus, il existe $x \in A$ tel que $x - x_0 = x + t \notin A$. Alors $x_0 - x \in -A$. En écrivant $x_0 = x + x_0 - x$, on en déduit $x_0 \in A + (-A)$. On a ainsi montré l'inclusion $A_0 \subset A + (-A)$.

Proposition 57. L'ensemble $\mathbb{R}$ muni de l'addition est un groupe abélien. La relation d'ordre est compatible avec l'addition.

Démonstration. La structure de groupe résulte de ce qui précède. Montrons la compatibilité de l'addition et de la relation d'ordre. Il faut prouver que, pour tous $x, x', y \in \mathbb{R}$, $x \leqslant x' \Rightarrow x + y \leqslant x' + y$. C'est évident en raisonnant en termes de parties de $\mathbb{Q}$. Soient $A, A', B \in \mathcal{P}(\mathbb{Q})$. Montrons que : $A \subset A' \Rightarrow A + B \subset A' + B$. Soit $c \in A + B$. Il s'écrit $c = a + b$, avec $a \in A$ et $b \in B$. On a $a \in A'$, donc $c = a + b \in A' + B'$. ∎

3.1.4 La multiplication des réels

Multiplication des réels

On définit les réels positifs et négatifs : $A \in \mathbb{R}$ est positif si $A_0 \leqslant A$, négatif si $A \leqslant A_0$. On note, comme d'habitude, $\mathbb{R}_+$ et $\mathbb{R}_-$. Comme dans $\mathbb{Q}$, en utilisant la règle des signes, la multiplication est connue si on la connaît pour les réels positifs.

Définition 14.

- Soient A et A' deux réels *positifs*. On leur associe une partie AA' de $\mathbb{Q}$ définie par : $AA' := \{z \in \mathbb{Q} \mid z = xx',\ x \in A \cap \mathbb{Q}_+^*,\ x' \in A' \cap \mathbb{Q}_+^*\} \cup \mathbb{Q}_-^*$.
- Soient A et A' deux réels. On définit $AA' \subset \mathbb{Q}$ par :
  - Si $A \in \mathbb{R}_+$ et $A' \in \mathbb{R}_-$, $AA' := -(A(-A'))$;
  - Si $A \in \mathbb{R}_-$ et $A' \in \mathbb{R}_+$, $AA' := -((-A)A'))$;
  - Si $A \in \mathbb{R}_-$ et $A' \in \mathbb{R}_-$, $AA' := (-A)(-A')$.

Lemme 58. Si $A, A' \in \mathbb{R}$, AA' est un réel.

Démonstration. L'opposé d'un réel est un réel. Il suffit donc de prouver le lemme dans le cas où $A, A' \in \mathbb{R}_+$. Si $A = A_0$ ou si $A' = A_0$, on a $AA' = \mathbb{Q}_-^* = A_0$. On peut donc supposer que A et A' sont *strictement* positifs.

- Montrons que le sous-ensemble $AA' \subset \mathbb{Q}$ est *propre*. On a $\mathbb{Q}_-^* \subset AA'$, donc $AA' \neq \varnothing$. Par ailleurs, il existe $x \notin A$, $x > 0$ et $x' \notin A'$, $x' > 0$. On a $A < x$ et $A' < x'$. Montrons que $xx' \notin AA'$, ce qui prouvera $AA' \neq \mathbb{Q}$. On raisonne par l'absurde. On a $xx' > 0$. Si $xx' \in AA'$, il existe $y \in A \cap \mathbb{Q}_+^*$ et $y' \in A' \cap \mathbb{Q}_+^*$ tels que $yy' = xx'$. On a $0 < y < x$ et $0 < y' < x'$, donc $yy' < xx'$ et une contradiction.
- Montrons que AA' est une section commençante. Soient $x \in AA'$ et $y \leqslant x$. On considère les deux cas : $y \leqslant 0$ et $y > 0$. Si $y \leqslant 0$, $y \in \mathbb{Q}_- \subset AA'$. Supposons $y > 0$. Alors il existe $z \in A \cap \mathbb{Q}_+^*$ et $z' \in A' \cap \mathbb{Q}_+^*$ tels que $x = zz'$. On a $y < zz'$. On écrit $y = zz' - t$, avec $t \in \mathbb{Q}_+$, puis $y = z(z' - t/z)$. On a $0 < z' - t/z$ et $z' - t/z < z'$, donc $z' - t/z \in A' \cap \mathbb{Q}_+^*$. Ainsi $y \in AA'$.
- Montrons que AA' n'a pas de plus grand élément. On raisonne par l'absurde. On suppose que AA' admet un plus grand élément m. On a $m > 0$, donc il existe $s \in A$, $s > 0$ et $s' \in A$, $s' > 0$, tels que $m = ss'$. Mais A et A' n'ont pas de plus grand élément. Il existe donc $t \in A$, $t > s$ et $t' \in A'$, $t' > s'$. On a $tt' > m$ et une contradiction. ∎

Les propriétés d'associativité et de commutativité de la multiplication et la distributivité de la multiplication par rapport à l'addition se déduisent facilement des propriétés similaires sur $\mathbb{Q}$. Nous laissons les détails au lecteur. Il vérifiera aussi, un peu moins facilement (il faut utiliser la densité de $\mathbb{Q}$ dans $\mathbb{R}$ et s'inspirer du cas de l'addition), que les multiplications sur $\mathbb{Q}$ et $\mathbb{R}$ sont compatibles.

Proposition 59. L'ensemble $\mathbb{R}$ muni de la relation d'ordre, de l'addition et de la multiplication définies ci-dessus et un anneau commutatif totalement ordonné contenant un sous-anneau isomorphe à $\mathbb{Q}$.

Inverse d'un nombre réel.

Il reste à montrer que $\mathbb{R}$ est un *corps*, c'est-à-dire que tout élément non nul a un *inverse*. C'est assez facile (sans être tout à fait évident...) dans le cas d'un nombre rationnel, mais c'est plus délicat dans le cas d'un réel irrationnel.

Définition 15. Soit $A \in \mathbb{R}$, $A \neq 0$. On définit un sous-ensemble de $\mathbb{Q}$, noté $1/A$, par :

- si $A = A_r$ avec $r \in \mathbb{Q}^*$, alors $1/A_r := A_{1/r}$;
- si $A > 0$ et n'est pas de la forme $A = A_r$ avec $r \in \mathbb{Q}^*$, alors
$$1/A := \{x \in \mathbb{Q}_+^* \mid 1/x \notin A\} \cup \mathbb{Q}_- \ ;$$
- si $A < 0$, alors $1/A = -\left(1/(-A)\right)$.

Lemme 60. Soit $A \in \mathbb{R}$, $A \neq 0$. Alors :

(i) $1/A$ est un réel ;

(ii) si $B := 1/A$, on a $A = 1/B$.

Démonstration.

(i) Le résultat est évident si $A = A_r$, $r \in \mathbb{Q}^*$. Supposons A irrationnel. Il suffit clairement de prouver le résultat pour $A > 0$. Soit donc $A > 0$, irrationnel.

  * Montrons que le sous-ensemble $1/A \subset \mathbb{Q}$ est *propre*. On a $\mathbb{Q}_- \subset 1/A$, donc $1/A$ n'est pas vide. Il existe $x \in A$, $x > 0$. Alors $1/x \notin 1/A$ et $1/A \neq \mathbb{Q}$.
  * Montrons que $1/A$ est une section commençante. Soient $x \in 1/A$ et $y \leqslant x$. Il faut montrer que $y \in 1/A$. Le cas $y \leqslant 0$ est trivial.On peut donc supposer $0 < y \leqslant x$. Alors, $1/x \notin A$, donc $1/x$ majore strictement A. Pour tout $z \in A$, $z > 0$, on a $z < 1/x$. par suite $x < 1/z$ et $y < 1/z$, c'est-à-dire $z < 1/y$. Ainsi $1/y \notin A$ et $y \in 1/A$.
  * Montrons que $1/A$ n'a pas de plus grand élément. On raisonne par l'absurde. Supposons que m soit le plus grand élément de $1/A$. Puisque $A > 0$, $m > 0$. On a $1/m \notin A$. Donc $1/m$ majore A. On a $A \subset A_{1/m}$ et l'inclusion est stricte, puisque A est irrationnel. Il existe donc un rationnel $m' > m$ tel que $1/m'$ majore A . On a $1/m' \notin A$, donc $m' \in 1/A$ et une contradiction.

(ii) C'est évident. ∎

Il reste à montrer que le réel $1/A$ est bien l'inverse de A. On a un résultat préliminaire (presque trivial...).

Lemme 61. Soit $x \in \mathbb{Q}$, tel que $0 < x < 1$. La suite de rationnels $(y_n := x^{-n})_{n\in\mathbb{N}}$ est strictement croissante et n'est pas majorée.

Démonstration. On a $y_{n+1} - y_n = x^{-(n+1)}(1-x) > 0$. la suite est donc strictement croissante. On a $x^{-1} > 1$ et l'on peut écrire $x^{-1} = 1 + z$, avec $z > 0$. Alors $y_n = (1+z)^n > 1 + nz$ (d'après la formule du binôme, ou si l'on préfère par récurrence). La suite $(1 + nz)$n'étant pas bornée, (y_n) ne l'est pas non plus. ∎

Proposition 62. Tout réel non nul A admet un inverse. Cet inverse est le réel $1/A$ défini ci-dessus (*cf.* la définition 15 et le lemme 60).

Démonstration. Il suffit de prouver le résultat dans le cas d'un réel strictement positif. Le cas général s'en déduit immédiatement. On suppose donc $A > 0$.
On distingue deux cas. Le premier est très facile.

1. On suppose que $A = A_r$, $r \in \mathbb{Q}$. Il faut montrer $A_r A_{1/r} = 1$, c'est-à-dire les inclusions $A_r A_{1/r} \subset A_1$ et $A_1 \subset A_r A_{1/r}$.
   - Montrons $A_r A_{1/r} \subset A_1$. Soit $y \in A_r A_{1/r}$. Si $y \in \mathbb{Q}_-$, on a évidemment $y \in A_1$. Supposons $y > 0$. Alors, il existe $x \in A_r$, $x > 0$ et $x' \in A_{1/r}$, $x' > 0$ tels que $y = xx'$. On a $0 < x < r$ et $0 < x' < 1/r$, donc $y = xx' < 1$.
     Ainsi $A_r A_{1/r} \subset A_1$.
   - Montrons $A_1 \subset A_r A_{1/r}$. Soit $z \in A_1$. Si $z \in \mathbb{Q}_-$, on a évidemment $z \in A_r A_{1/r}$. Supposons $z > 0$. Puisque $z < 1$, il existe $t \in \mathbb{Q}$ tel que $z < t < 1$. On écrit $z = (tr)\left(\frac{z}{tr}\right)$. On a $0 < tr < r$, donc $tr \in A_r \cap \mathbb{Q}_+^*$, et $\frac{z}{tr} = \frac{z}{t}\frac{1}{r} < \frac{1}{r}$, donc $\frac{z}{tr} \in A_{1/r} \cap \mathbb{Q}_+^*$. Par suite $z \in A_r A_{1/r}$. Ainsi $A_1 \subset A_r A_{1/r}$.
2. On suppose que le réel A est irrationnel. Il faut montrer les inclusions $A(1/A) \subset A_1$ et $A_1 \subset A(1/A)$. Il faut travailler un peu plus pour prouver la seconde.
   - Montrons $A(1/A) \subset A_1$. On a $\mathbb{Q}_- \subset A_1$. Soient $x, x' > 0$, $x \in A$, $x' \in A'$. Pour prouver l'inclusion, il suffit de montrer que $xx' \in A_1$.
     Puisque $x' \in 1/A$, $1/x' \notin A$, donc $1/x'$ majore strictement A.
     En particulier $x < 1/x'$, donc $xx' < 1$, c'est-à-dire $xx' \in A_1$.
   - Montrons $A_1 \subset A(1/A)$. On distingue deux cas.
     * On suppose $1 < A$. Soit $x \in A_1 \cap \mathbb{Q}_+^*$, c'est-à-dire $0 < x < 1$. On applique le lemme à la suite $(y_n = x^{-n})_{n\in\mathbb{N}}$. On a $y_0 = 1 \in A$ et, puisque A est borné et que la suite est croissante non bornée, il existe un nombre fini, non nul, de termes de la suite appartenant à A. Soit $y_k = x^{-k} \in A$ le plus grand. On a $y_{k+1} = x^{-(k+1)} \notin A$, donc $x^{k+1} \in 1/A$. On a $x^{-k} \in A \cap \mathbb{Q}_+^*$ et $x^{k+1} \in 1/A \cap \mathbb{Q}_+^*$.De plus $x = x^{-k}x^{k+1}$, donc $x \in A(1/A)$. On a prouvé l'inclusion $A_1 \subset A(1/A)$.
     * On suppose $0 < A < 1$. Puisque $1 \notin A$, $1 \in 1/A$ et $1 < 1/A$. On peut donc échanger les rôles de A et $1/A$ dans la preuve précédente.
       Ainsi $A_1 \subset (1/A)A = A(1/A)$ (en utilisant la commutativité de la multiplication). ∎

Des propositions 59 et 62 on déduit le résultat suivant, qui achève la construction de $\mathbb{R}$ par les sections commençantes.

Théorème 63 (Théorème de Dedekind).
L'ensemble $\mathbb{R}$ des sections commençantes de $\mathbb{Q}$, muni de la relation d'ordre et des opérations d'addition et de multiplication définies ci-dessus, est un corps totalement ordonné vérifiant les axiomes des nombres réels.

3.2 Un théorème d'unicité

Le formalisme des sections commençantes permet de prouver l'unicité d'un corps totalement ordonné satisfaisant les axiomes des nombres réels à isomorphisme unique près, que nous avons admise dans le livre [L1].

Soit $\mathcal{R}$ un corps totalement ordonné satisfaisant les axiomes de $\mathbb{R}$. Pour tout $x \in \mathcal{R}$, on note $A_x := \{u \in \mathbb{Q} \mid u < x\}$. On remarque : $x \notin A_x$ et $A_x \neq \varnothing$, ainsi A_x est une section commençante ouverte de $\mathbb{Q}$.

Proposition 64. Soit $\mathcal{R}$ un corps totalement ordonné satisfaisant les axiomes de $\mathbb{R}$. On note $\sup$ la borne supérieure d'un sous-ensemble de $\mathcal{R}$, quand elle existe.

(i) Toute section commençante A de $\mathbb{Q}$ admet une borne supérieure dans $\mathcal{R}$.

(ii) Pour tout $x \in \mathcal{R}$, x est la borne supérieure de $A_x = \{u \in \mathbb{Q} \mid u < x\}$ dans $\mathcal{R}$ et A_x est une section commençante ouverte.

(iii) Si A est une section commençante ouverte de borne supérieure x dans $\mathcal{R}$, on a $A = A_x$.

(iv) Soient A et A' des sections commençantes ouvertes de $\mathbb{Q}$.
On note $A + A' := \{u + u' \mid u \in A,\ u' \in A'\} \subset \mathbb{Q}$.
On a $\sup A + \sup A' = \sup(A + A')$.
Pour tous $x, x' \in \mathcal{R}$, on a $A_x + A_{x'} = A_{x+x'}$.

(v) Soit A une section commençante ouverte de Q.
On lui associe $C := A \cap \{u \in \mathbb{Q} \mid u > 0\}$.
Si $A := A_x$, avec $x > 0$, on note $C := C_x$. Si C et C' ne sont pas vides, on note $CC' = \{uu' \mid u \in C,\ u' \in C'\} \subset \mathbb{Q}$.
Alors, pour tous $x,\ x' > 0$, on a $C_x C_{x'} = C_{xx'}$.

Démonstration. On utilise des variantes des arguments utilisés plus haut dans la construction de $\mathbb{R}$ par les sections commençantes. Cette démonstration peut être omise en première lecture.

(i) Soit A une section commençante de $\mathbb{Q}$. Puisque $A \neq \mathbb{Q}$, il existe $u \in \mathbb{Q}$ tel que $u \notin A$. Alors u majore A dans $\mathbb{Q}$ et dans $\mathcal{R}$. D'après la propriété de la borne supérieure dans $\mathcal{R}$, $A \neq \varnothing$ admet une borne supérieure dans $\mathcal{R}$.

(ii)
- On raisonne par l'absurde. Supposons que x ne soit pas la borne supérieure de A_x dans $\mathcal{R}$. Alors il existe $y \in \mathcal{R}$, $y < x$, majorant A_x. Il existe un rationnel r tel que $y < r < x$. On a alors, d'après la définition de A_x, $r \in A_x$, d'où une contradiction. Ainsi x est la borne supérieure de A_x.
- Le corps $\mathcal{R}$ est archimédien, il existe donc $n_1 \in -\mathbb{N}$ et $n_2 \in \mathbb{N}$ tels que $n_1 < x < n_2$. On a $n_1 \in A_x$ et $n_2 \notin A_x$, donc A_x est un sous-ensemble propre de $\mathbb{Q}$. C'est une section commençante. Puisque $x \notin A_x$, la section commençante A_x est ouverte.

(iii) Puisque A est ouverte, on ne peut pas avoir $x \in A$. Par suite, puisque x majore A, on a $A \subset A_x$. Supposons $A \neq A_x$, alors il existe $u \in A_x \setminus A$. On a $u < x$ et u majore A. Puisque x est la borne supérieure de A, il existe $r \in A$ tel que $u < r \leqslant x$, d'où une contradiction. Ainsi $A = A_x$.

(iv) Nous avons vu plus haut que $A + A'$ est une section commençante.

* Soient $x := \sup A$ et $x' := \sup A'$. On a $A = A_x$, $A' = A_{x'}$ et $A_x + A_{x'} \subset A_{x+x'}$. Notons z la borne supérieure de $A + A' = A_x + A_{x'}$. La borne supérieure de $A_{x+x'}$ dans $\mathcal{R}$ est $x + x'$ et elle majore z. Montrons que $z = x + x'$, c'est-à-dire $\sup(A + A') = \sup A + \sup A'$.

  On raisonne par l'absurde en supposant que $z < x + x'$. Alors, par densité de $\mathbb{Q}$ dans $\mathcal{R}$, il existe un rationnel s tel que $z < s < x + x'$. En utilisant $x + x' - s > 0$ et la caractérisation de la borne supérieure (*cf.* la proposition 4 de [L1, IV.1, 1.2.1]), on peut trouver $r \in A_x$ et $r' \in A_{x'}$ tels que :

$$x - \frac{x + x' - s}{2} < r < x \quad \text{et} \quad x' - \frac{x + x' - s}{2} < r' < x'. \tag{9}$$

  On a alors, en utilisant $x - \frac{x+x'-s}{2} + x' - \frac{x+x'-s}{2} = s$: $\quad r + r' \in A_x + A_{x'}$, et $z < s < r + r'$, d'où une contradiction puisque $r + r' \in A_x + A_{x'}$.

* Soient $x, x' \in \mathcal{R}$. On a $\sup(A_x + A_{x'}) = \sup A_x + \sup A_{x'} = x + x'$. On a aussi $A_x + A_{x'} = A_{\sup(A_x + A_{x'})}$, donc $A_x + A_{x'} = A_{x+x'}$.

(v) La preuve de l'égalité $C_{xx'} = C_x C_{x'}$ est une variante de celle de $A_x + A_{x'} = A_{x+x'}$. On remplace, en particulier, les inégalités (9) par des versions multiplicatives :

$$v = \frac{u}{uu'v^{-1}}u', \quad \frac{x}{\sqrt{xx's^{-1}}} < r < x \quad \text{et} \quad \frac{x'}{\sqrt{xx's^{-1}}} < r' < x'.$$

Nous laissons les détails au lecteur. ∎

Théorème 65. Soit $\mathcal{R}$ un corps totalement ordonné satisfaisant les axiomes de $\mathbb{R}$. Le corps $\mathbb{R}$ étant identifié à l'ensemble des sections commençantes ouvertes de $\mathbb{Q}$, on définit une application $f : \mathcal{R} \to \mathbb{R}$ par $f(x) := A_x = \{u \in \mathbb{Q} \mid u < x\}$. C'est un isomorphisme croissant d'anneaux totalement ordonnés.

Démonstration. Le théorème est essentiellement une traduction de la proposition 64 ci-dessus. Pour éviter les confusions, on notera dans cette démonstration $\preceq$ (resp. $\prec$) la relation d'ordre (resp. d'ordre strict) sur $\mathcal{R}$. On rappelle qu'elle induit sur $\mathbb{Q}$ la relation d'ordre usuelle.

1. Montrons que l'application f est injective. Si $f(x) = f(x')$, par définition $A_x = A_{x'}$. Alors $x = \sup A_x = \sup A_{x'} = x'$.
2. Montrons que l'application f est surjective. Soit A un élément de $\mathbb{R}$. C'est une section commençante. Soit $x = \sup A$. Alors $f(x) = A_x = A$.
3. Montrons que l'application f est strictement croissante. Si $x \prec x'$, $A_x \subset A_{x'}$. D'après la définition de la relation d'ordre sur $\mathbb{R}$, on a alors $f(x) \leqslant f(x')$.
4. Montrons que, pour tous $x, x' \in \mathcal{R}$, l'on a $f(x + x') = f(x) + f(x')$. Soient $x, x' \in \mathcal{R}$. On a $A_x + A'_x = A_{x+x'}$. D'après la définition de l'addition de $\mathbb{R}$, on en déduit $f(x) + f(x') = f(x + x')$.
5. Montrons que, pour tous $x, x' \in \mathcal{R}$, strictement positifs (i.e. , $0 \prec x, x'$), l'on a $f(xx') = f(x)f(x')$. On a $C(x)C(x') = C(xx')$. D'après la définition de la multiplication de $\mathbb{R}$, on en déduit $f(x)f(x') = f(xx')$.

6. Montrons que, pour tous $x, x' \in \mathcal{R}$, on a $f(xx') = f(x)f(x')$.
   On a $f(0_{\mathcal{R}}) = A_0 = 0_{\mathbb{R}}$, donc $0_{\mathbb{R}} = f(x - x) = f(x) + f(-x) = A_x + A_{-x}$ et $f(-x) = -f(x)$. Ceci permet de se ramener au cas précédent. ∎

Corollaire 66. Soient $\mathcal{R}_1$ et $\mathcal{R}_2$ deux corps totalement ordonnés vérifiant les axiomes des nombres réels.

(i) Il existe un unique isomorphisme croissant d'anneaux totalement ordonnés de $\mathcal{R}_1$ dans $\mathcal{R}_2$.

(ii) Soit $\mathcal{K}$ un corps totalement ordonné vérifiant la propriété de la borne supérieure. Il existe un unique isomorphisme croissant d'anneaux totalement ordonnés de $\mathcal{R}_1$ dans $\mathcal{K}$.

Démonstration.

(i) Cette assertion est une conséquence immédiate du théorème. On applique le théorème à $\mathcal{R}_1$ et $\mathcal{R}_2$. On obtient des isomorphismes $f_1 : \mathcal{R}_1 \to \mathbb{R}$ et $f_2 : \mathcal{R}_2 \to \mathbb{R}$. L'application $f_2^{-1} \circ f_1 : \mathcal{R}_1 \to \mathcal{R}_2$ est un isomorphisme croissant d'anneaux totalement ordonnés. On vérifie l'unicité en se ramenant au cas $\mathcal{R}_2 = \mathbb{R}$ (par composition par f_2).

(ii) Le corps totalement ordonné $\mathcal{K}$ est de caractéristique nulle. L'application $\mathbb{Z} \mapsto \mathcal{K}$ définie par $n \mapsto n.1$ est un homomorphisme injectif d'anneaux. Le plus petit sous-corps K de $\mathcal{K}$ contenant l'image de cette application (le sous-corps premier de $\mathcal{K}$) est isomorphe à $\mathbb{Q}$. On reprend *mutatis mutandis* les preuves ci-dessus en remplaçant $\mathcal{R}_2$ par $\mathcal{K}$ et l'identité de $\mathbb{Q}$ par l'isomorphisme $\mathbb{Q} \to K$. Nous laissons les détails au lecteur. ∎

Remarque.

Nous donnerons plus loin une autre preuve de l'unicité basée *sur la notion d'espace métrique complet*. Cette preuve est très différente.

4 Complétions de corps et construction de $\mathbb{R}$

Dans cette partie les corps sont, comme précédemment, toujours supposés *commutatifs*.

La valeur absolue ordinaire sur $\mathbb{Q}$ permet de le munir d'une distance. On peut alors parler des *suites de Cauchy de* $\mathbb{Q}$. Charles Meray, puis Georg Cantor (et Heinrich Eduard Heine en 1872) ont proposé une construction rigoureuse de de $\mathbb{R}$ basée sur les suites de Cauchy de rationnels. Nous allons la présenter (en version « modernisée »). Cette méthode est moins élémentaire que les deux précédentes : la *définition* est basée sur une notion plus « sophistiquée ». Par contre les propriétés des opérations sont plus simples à vérifier.

Cette troisième méthode a un autre avantage. On peut l'appliquer (avec quelques adaptations) en partant d'une valeur absolue sur $\mathbb{Q}$ différente de la valeur absolue ordinaire. En utilisant les valeurs absolues p-adiques (p étant un nombre premier), on obtient ainsi d'autres corps que $\mathbb{R}$, les *corps p-adiques* $\mathbb{Q}_p$ qui jouent en arithmétique un rôle presque aussi important[26] que $\mathbb{R}$. Nous les avons déjà rencontrés dans la partie 1, avec des défini-

[26] La classification des formes quadratiques sur $\mathbb{Q}$ sur laquelle nous donnons quelques indications dans le module I.2 est un exemple classique.

tions différentes. Ces autres complétions de $\mathbb{Q}$ sont aussi appelées *corps non-archimédiens*. Une fois que l'on dispose de $\mathbb{R}$, on peut définir la notion de corps normé. Le procédé utilisé pour compléter $\mathbb{Q}$ s'étend facilement, *mutatis mutandis*, à un corps normé arbitraire.
Pour la définition et les propriétés des suites de Cauchy d'un espace métrique, on se reportera à [L2, II.1, 2.1].

4.0.1 Suites d'un corps commutatif totalement ordonné

Pour les propriétés élémentaires des corps commutatifs totalement ordonnés le lecteur se reportera à [L1, IV.1, 1.2.1]. Un corps totalement ordonné K est de caractéristique nulle. Son sous-corps premier est isomorphe à $\mathbb{Q}$. Si ce sous-corps est $\mathbb{Q}$, l'ordre induit sur $\mathbb{Q}$ par l'ordre de K est l'ordre classique. On note $K_+ := \{x \in K \mid x \geqslant 0\}$.
La preuve du résultat suivant est facile et elle est laissée au lecteur.

Proposition 67.

(i) Soient G un groupe abélien et une partie $P \subset K$ vérifiant :
$$P + P \subset P, \;\; P \cap (-P) = \{0\}.$$
Alors il existe une unique relation d'ordre sur G qui en fasse un groupe ordonné dont l'ensemble des éléments positifs est P. C'est la relation $x \leqslant y$ si, et seulement si, $y - x \in P$.
L'ordre est total si, et seulement si, on a, de plus, $P \cup (-P) = G$.

(ii) Soient A un anneau commutatif et une partie $P \subset K$ vérifiant :
$$P + P \subset P, \;\; P \cap (-P) = \{0\}, \;\; P \cup (-P) = A, \;\; P.P \subset P.$$
Alors il existe une unique relation d'ordre sur A qui en fasse un anneau ordonné dont l'ensemble des éléments positifs est P. De plus la relation d'ordre est totale.

Signalons que l'on peut caractériser algébriquement les corps totalement ordonnables[27]. On montre qu'un corps K est totalement ordonnable (on dit aussi que c'est un *corps formellement réel*) si, et seulement si, -1 n'est pas une somme de carrés de K. La preuve utilise le lemme de Zorn (voir l'exercice III.3.38 de la page 528).
Soit K un corps totalement ordonné, on définit une valeur absolue sur K par :
$$|\;| : x \in K \mapsto \max(x, -x) \in K_+.$$
Elle vérifie :

(i) $|x| = 0 \Leftrightarrow x = 0$;

(ii) $\forall x, y \in K, \;\; |xy| = |x||y|$;

(iii) $\forall x, y \in K, \;\; |x + y| \leqslant |x| + |y|$.

On définit une *topologie* sur K en prenant comme base d'ouverts les intervalles ouverts.
Les suites de K sont les éléments de $K^{\mathbb{N}}$. On les munit de façon évidente d'une addition et d'une multiplication, qui font de $K^{\mathbb{N}}$ une K-algèbre commutative. L'élément neutre de l'addition est la suite nulle, celui de la multiplication est la suite constante définie par 1.
Dans un corps totalement ordonné, on a les notions de *suite majorée* (resp. *minorée*), *suite bornée*, *suite convergente* et *suite de Cauchy*. On reprend, *mutatis mutandis*, les définitions utilisées pour $\mathbb{R}$.

[27] Une étude systématique de ces corps a été faite entre 1924 et 1927 par E. Artin et O. Schreier. La motivation d'Artin était la solution du 17ème problème de Hilbert.

Définition 16. Soient K un corps totalement ordonné et $(x_n)_{n\in\mathbb{N}}$ une suite de K. On dit que :

(i) la suite (x_n) est *majorée* (resp. *minorée*) s'il existe $M \in K$ (resp. $m \in K$), tel que, pour tout $n \in \mathbb{N}$, $x_n \leqslant M$ (resp. $m \leqslant x_n$).

(ii) la suite (x_n) est *bornée* si elle est à la fois majorée et minorée ;

(iii) la suite (x_n) est *converge vers* $a \in K$ si, pour tout $\varepsilon \in K$, $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que, pour tout $n \geqslant N$, $|x - a| < \varepsilon$;

(iv) la suite (x_n) est *de Cauchy* si, pour tout $\varepsilon \in K$, $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que, pour tous $m, n \geqslant N$, $|x_m - x_n| < \varepsilon$.

Le lecteur vérifiera facilement les assertions suivantes. Une suite (x_n) est bornée si, et seulement si, la suite $(|x_n|)$ est majorée. Les suites bornées forment une sous-algèbre de $K^{\mathbb{N}}$. Toute suite convergente est de Cauchy. Les suites convergentes forment une sous-algèbre de $K^{\mathbb{N}}$. L'application qui à une suite convergente associe sa limite est un homomorphisme d'algèbres de l'algèbre des suites convergentes dans K. Toute suite de Cauchy est bornée. Les suites de Cauchy forment une sous-algèbre de $K^{\mathbb{N}}$.
On dit qu'un corps totalement ordonné K est *complet* si toute suite de Cauchy est convergente. On définit la notion de compacité d'une partie comme d'habitude en utilisant les recouvrements ouverts. On dit qu'un corps totalement ordonné K est *connexe* si l'on ne peut pas l'écrire comme réunion de deux ouverts non vides disjoints.
Le lecteur pourra se familiariser avec les propriétés des suites d'un corps totalement ordonné en faisant l'exercice III.3.36 de la page 528.

Malgré la similitude du vocabulaire et de certaines propriétés, un corps totalement ordonné n'est pas en général un espace métrique. Le corps $\mathbb{R}$ ne joue aucun rôle dans les définitions ci-dessus.

4.1 Complétion d'un corps normé

4.1.1 Complétions de $\mathbb{Q}$

Conventions de lecture

- Nous supposons dans ce paragraphe, que nous ne connaissons ni $\mathbb{R}$ ni les corps p-adiques $\mathbb{Q}_p$, ni les anneaux d'entiers Z_p. Nous allons les construire par une approche différente de celles utilisées auparavant. On peut montrer que l'on obtient les mêmes objets « à isomorphisme unique près ».

- Nous allons considérer des suites de $\mathbb{Q}$. Ne connaissant pas $\mathbb{R}$, on ne peut pas définir les notions de *distance*, de *suite convergente* et de *suite de Cauchy* de la façon usuelle. Les distances seront des applications $d : \mathbb{Q}^2 \to \mathbb{Q}_+$ vérifiant les conditions usuelles. Pour la convergence et la propriété de Cauchy, il faut remplacer les conditions du type "pour tout $\varepsilon > 0$" par "pour tout $\varepsilon > 0$ *rationnel*". Par la suite, il sera supposé implicitement, quand le contexte l'imposera, que $\varepsilon \in \mathbb{Q}$ quand nous écrirons "pour tout $\varepsilon > 0$".

Valuations

La notion de *valuation* a été introduite (sous une forme un peu différente et moins générale) en 1882 par les mathématiciens allemands Dedekind et Weber.

Définition 17. On appelle valuation une application v d'un anneau commutatif unitaire A vers vers $G \cup \{+\infty\}$, où G est un groupe commutatif totalement ordonné, vérifiant les trois conditions suivantes :

(i) $v(x) = +\infty$ si, et seulement si, $x = 0$;

(ii) pour tous $x, y \in A$, $v(xy) = v(x) + v(y)$;

(iiii) pour tous $x, y \in A$, $v(x+y) \geqslant \min\,(v(x), v(y))$.

Dans un premier temps, nous nous intéresserons surtout au cas où $A = \mathbb{Z}$ ou $\mathbb{Q}$ et où G est le groupe additif $(\mathbb{Z}, +)$. On considérera plus tard d'autres anneaux A, avec, si l'on connaît $\mathbb{R}$ (ce qui n'est pas ici dans nos conventions), $G = (\mathbb{R}, +)$. Si $G = \mathbb{Z}$, on dit que la valuation est *discrète*.

Proposition 68. Soit $v :\to G \cup \{+\infty\}$ une valuation. Alors :

(i) $v(1) = v(-1) = 0$ et, pour tout $x \in A$, $v(-x) = v(x)$;

(ii) pour tous $x, y \in A$, si $v(x) \neq v(y)$, on a $v(x+y) = \min\,(v(x), v(y))$;

(iii) l'anneau A est *intégre* ;

(iv) Soit K le corps des fractions de A.
Il existe une unique valuation $w : K \to G \cup \{+\infty\}$ qui prolonge v ; elle est définie par $w(x/y) := v(x) - v(y)$, pour $x \in A$, $y \in A^* := A \setminus \{0\}$.

Démonstration.

(i) On a $v(1) = v(1^2) = v(1) + v(1)$, donc $v(1) = 0$. Par suite :
$$0 = v(1) = v\left((-1)^2\right) = v(-1) + v(-1) \quad \text{et} \quad v(-1) = 0.$$
On a utilisé le fait que dans un groupe totalement ordonné $g + g = 0$ entraîne $g = 0$.
On a $v(-x) = v(-1) + v(x) = v(x)$.

(ii) On peut supposer $v(x) < v(y)$. Alors $v(x+y) \geqslant \min\,(v(x), v(y)) = v(x)$. Par ailleurs $v(x) = v\left((x+y) - y)\right) \geqslant \min\,(v(x+y), v(y))$ et nécessairement $v(x) \geqslant v(x+y)$. Alors $v(x) \geqslant v(x+y) \geqslant v(x)$, donc $v(x+y) = v(x)$.

(iii) Si $x, y \neq 0$, $v(xy) = v(x) + v(y) \in G$, donc $v(xy) \neq +\infty$ et $xy \neq 0$.

(iv) La vérification est facile. ∎

Exemples.

1. Soient $A := \mathbb{Z}$ et $G = \mathbb{Z}$. Soit p un nombre premier. Tout $a \in \mathbb{Z}^*$ s'écrit de façon unique $a = p^n b$, avec $n \in \mathbb{N}$ et $b \in \mathbb{Z}^*$ non divisible par p. On pose $v_p(a) := n$ et $v_p(0) = +\infty$. On vérifie que v_p est une valuation. Elle s'étend de manière unique en une valuation sur $\mathbb{Q}$, toujours notée v_p, à valeurs dans $\mathbb{Z}$. On l'appelle *la valuation p-adique* sur $\mathbb{Q}$. Nous l'avons déjà rencontrée dans la partie 2.

Tout rationnel $\in \mathbb{Q}$ s'écrit $r = p^v m/n$, avec $v, m \in \mathbb{Z}$, $n \in \mathbb{N}^*$, m et n étant non divisibles par p. Alors v est déterminé de manière unique et $v = v_p(r)$. On peut vérifier directement l'inégalité ultramétrique $v_p(x + x') \geqslant \min\,(v_p(x), v_p(x'))$. On écrit $y = p^{v'} m'/n'$. Supposons $v \leqslant v'$. On a $p^v \dfrac{m}{n} + p^{v'} \dfrac{m'}{n'} = p^v \dfrac{mn' + p^{v'-v} m'n}{nn'}$. L'inégalité s'en déduit. Si $v < v'$ il y a égalité. L'inégalité est stricte si, et seulement si, $v = v'$ *et* p divise $mn' + m'n$.

2. Soient K un corps commutatif, $A := K[X]$ et $G = \mathbb{Z}$. Soit $P_0 \in K[X]$ un polynôme unitaire *irréductible*. Tout polynôme $P \in K[X]$ non nul s'écrit de façon unique $P = P_0^n Q$, avec $n \in \mathbb{N}$ et $Q \in K[X]$, non divisible par P_0. On pose $v_{P_0}(P) := n$ et $v_{P_0}(0) = +\infty$. On vérifie que v_{P_0} est une valuation. Elle s'étend de manière unique en une valuation sur le corps de fractions rationnelles $K(X)$, toujours notée v_{P_0}, à valeurs dans $\mathbb{Z}$.

Normes et topologies sur $\mathbb{Q}$

Définition 18. Une *norme* (ou *valeur absolue*) sur $\mathbb{Q}$ est une application $x \mapsto |x|$ de $\mathbb{Q}$ dans $\mathbb{Q}_+$ vérifiant les trois propriétés suivantes :

(i) $\forall x \in \mathbb{Q},\ \ |x| = 0 \Leftrightarrow x = 0$;

(ii) $\forall x, y \in \mathbb{Q},\ \ |xy| = |x||y|$;

(iii) $\forall x, y \in \mathbb{Q},\ \ |x + y| \leqslant |x| + |y|$.

Une norme et dite *ultramétrique* si elle vérifie la condition

(iii') $\forall x, y \in \mathbb{Q},\ \ |x + y| \leqslant \sup\{|x|, |y|\}$.

qui est plus forte que la condition (iii).

Une norme $|\ |$ est dite *triviale* si $|x| = 1$ pour tout $x \in \mathbb{Q}^*$.
On a $1 = 1^2 = (-1)^2$, donc $|1| = |1|^2 = |-1|^2$. Puisque $|1| \neq 0$, $|1| = 1$ et, puisque la norme est positive, $|-1| = 1$.
La donnée d'une norme $||$ sur $\mathbb{Q}$ permet de définir une distance (*cf.* [L2, I.1]) : $d(x, y) := |x - y|$, à valeurs dans $\mathbb{Q}_+$.
Les exemples qui suivent sont, nous le verrons, « essentiellement les seuls » (*cf.* le théorème 92 de la page 506).
Exemples.

1. La valeur absolue ordinaire $|x| := \max\,(x, -x)$ est une norme sur $\mathbb{Q}$. On note aussi $|\ | = |\ |_\infty$
2. Soit $v = \mathbb{Q} \to \mathbb{Z}$ une valuation.

   Soit $s \in \mathbb{N}^*$. On vérifie facilement que $x \mapsto |x| := s^{-v(x)}$ est une norme *ultramétrique* sur $\mathbb{Q}$. La distance associée est notée d_p.

   En particulier, si p est un nombre premier, $x \mapsto |x|_p := p^{-v_p(x)}$ (*cf.* l'exemple 1 de la page 494) est une norme *ultramétrique* sur $\mathbb{Q}$. On l'appelle la norme p-adique (ou valeur absolue p-adique) sur $\mathbb{Q}$.

Complétions

Dans ce paragraphe, on choisit une norme $|\,|$ sur $\mathbb{Q}$. On suppose que $|\,|$ est la valeur absolue ordinaire ou une norme ultramétrique.

On note $\widetilde{\mathbb{Q}}$ l'ensemble des suites de Cauchy de $\mathbb{Q}$ pour la distance définie par la norme $|\,|$. On note $\mathcal{I}$ l'ensemble des suites tendant vers 0.

Lemme 69. Soit $a = (a_n) \in \widetilde{\mathbb{Q}} \setminus \mathcal{I}$. Alors, il existe $\delta \in \mathbb{Q}$, $\delta > 0$, et $N \in \mathbb{N}$ tels que, pour tout $n \geqslant N$, l'on ait $|a_n| \geqslant \delta$.

Démonstration. On raisonne par l'absurde. On suppose que la propriété $(\star)$ suivante est satisfaite : pour tout $\delta \in \mathbb{Q}$, $\delta > 0$, et tout $N \in \mathbb{N}$, il existe $n \geqslant N$ tel que $|a_n| < \delta$.

Soit $\varepsilon \in \mathbb{Q}$, $\varepsilon > 0$. Puisque a et une suite de Cauchy, il existe $N_0 \in N$ tel que $|a_n - a_m| < \varepsilon/3$, pour $n, m \geqslant N_0$.

On prend $\delta := \varepsilon/3$ et $N = N_0$ dans $(\star)$. Alors il existe $n_0 \geqslant N_0$ tel que $|a_{n_0}| < \varepsilon/3$. Par suite, pour tout $n \geqslant N_0$, on a $|a_n - a_{n_0}| < \varepsilon/3$, $|a_n| \leqslant |a_n - a_{n_0}| + |a_{n_0}| \leqslant 2\varepsilon/3 < \varepsilon$. On a ainsi montré que la suite $a = (a_n)$ converge vers 0, ce qui est contraire à l'hypothèse. Par suite la propriété $(\star)$ est fausse et son contraire est vrai. ∎

Lemme 70. On munit $\widetilde{\mathbb{Q}}$ de l'addition et de la multiplication des suites de $\mathbb{Q}$. Alors $\widetilde{\mathbb{Q}}$ est un anneau (commutatif unitaire) et $\mathcal{I}$ est un idéal *maximal* de $\widetilde{\mathbb{Q}}$. Le quotient $\widetilde{\mathbb{Q}}/\mathcal{I}$ est un *corps* noté $\widehat{\mathbb{Q}}$.

On note $\varphi : \widetilde{\mathbb{Q}} \to \widetilde{\mathbb{Q}}/\mathcal{I}$ l'application canonique.

Démonstration. On vérifie facilement que $\widetilde{\mathbb{Q}}$ est un anneau commutatif. Les preuves sont les mêmes que celles faites, dans [L2] pour les suites de Cauchy de réels en prenant « les ε » rationnels (au lieu de réels). L'élément unité est l'image par φ de la suite constante $(a_n = 1)_{n\in\mathbb{N}}$.

L'ensemble $\mathcal{I}$ est évidemment un sous-anneau de $\widetilde{\mathbb{Q}}$. On vérifie que toute suite de Cauchy de $(\mathbb{Q}, |\,|)$ est *bornée*. On en déduit que $\mathcal{I}$ est un *idéal* de $\widetilde{\mathbb{Q}}$. Montrons qu'il est maximal. Pour cela, on va prouver que si $a = (a_n) \in \widetilde{\mathbb{Q}} \setminus \mathcal{I}$, alors il existe $b \in \widetilde{\mathbb{Q}}$ tel que $ab - 1 \in \mathcal{I}$.

Soit $a = (a_n) \in \widetilde{\mathbb{Q}} \setminus \mathcal{I}$. On utilise le lemme 69. Il existe $\delta \in \mathbb{Q}$, $\delta > 0$ et $N \in \mathbb{N}$ tels que $|a_n| \geqslant \delta$ pour $n \geqslant N$. On définit $b = (b_n)$ par :

$$b_n := 0 \ \text{ si } \ n < N \quad \text{ et } \quad b_n := 1/a_n \ \text{ si } \ n \geqslant N.$$

Pour $m, n \geqslant N$, on a $b_n - b_m = \frac{a_m - a_n}{a_n a_m}$, donc $|b_n - b_m| = \frac{|a_m - a_n|}{\delta^2}$, ce qui montre que b est une suite de Cauchy de $\mathbb{Q}$. La suite $ab - 1$ est nulle à partir du rang N, elle tend donc vers 0 : $ab - 1 \in \mathcal{I}$. On en déduit que l'idéal de $\widetilde{\mathbb{Q}}$ engendré par a et $\mathcal{I}$ contient 1 et est donc égal à $\widetilde{\mathbb{Q}}$. Ainsi l'idéal $\mathcal{I}$ est *maximal*. Le quotient $\widetilde{\mathbb{Q}}/\mathcal{I}$ est un *corps*. ∎

On identifie $\mathbb{Q}$ a l'ensemble des suites constantes de $\widetilde{\mathbb{Q}}$ en utilisant l'application :

$$i : a \in \mathbb{Q} \mapsto (a_n = a)_{n\in\mathbb{N}}.$$

On obtient, en passant au quotient une application injective $\varphi \circ i : \mathbb{Q} \to \widehat{\mathbb{Q}}$. Ceci permet de considérer $\mathbb{Q}$ comme un sous-corps de $\widehat{\mathbb{Q}}$, ce que nous ferons dans la suite $\big($en écrivant a au lieu de $\varphi \circ i(a)\big)$.

Nous allons « prolonger la norme $|\ |$ » de $\mathbb{Q}$ à $\widehat{\mathbb{Q}}$. Nous distinguerons deux cas. D'une part : $|\ | = |\ |_\infty$ et d'autre part : $|\ |$ est une norme ultramétrique. Dans le premier cas, nous obtiendrons un corps totalement ordonné complet. Dans le second, un corps complet muni d'une norme ultramétrique (à valeurs dans $\mathbb{Q}_+$).

L'idée du prolongement est de montrer que si $a = (a_n)$ et $b = (b_n)$ sont des suites de Cauchy de $\mathbb{Q}$, avec $a - b \in \mathcal{I}$, alors $|a| = (|a_n|)$ et $|b| = (|b_n|)$ sont aussi des suites de Cauchy de $\mathbb{Q}$ et vérifient $|a| - |b| \in \mathcal{I}$. La norme de la classe de a est alors définie comme la classe de $|a|$.

Le lecteur vérifiera facilement les résultats suivants.

Proposition et définition 71. Soit $a = (a_n) \in \widetilde{\mathbb{Q}}$. Alors $|a| := (|a_n|) \in \widetilde{\mathbb{Q}}$.
Soit $c = (c_n) \in \mathcal{I}$. Alors $|c| := (|c_n|) \in \mathcal{I}$.
Soient $a = (a_n),\ b = (a_n) \in \widetilde{\mathbb{Q}}$ tels que $a - b \in \mathcal{I}$. Alors $|a| - |b| \in \mathcal{I}$.
L'application de $\widetilde{\mathbb{Q}}$ dans lui-même définie par $a \mapsto |a|$ induit une application de $\widehat{\mathbb{Q}}$ dans lui-même. On l'appelle *l'application norme* de $\widetilde{\mathbb{Q}}$.
Elle prolonge l'application norme $|\ | : \mathbb{Q} \to \mathbb{Q}_+$.

L'image de l'application norme de $\widehat{\mathbb{Q}}$ n'est pas toujours contenue dans $\mathbb{Q}_+$.

Nous allons montrer que :

- La norme de $\widehat{\mathbb{Q}}$ vérifie trois assertions similaires (*mutatis mutandis*) à celles de la proposition 20.
- Si $|\ |$ est la valeur absolue usuelle sur $\mathbb{Q}$, la norme prolongée sur $\widehat{\mathbb{Q}}$ est à valeurs dans $\widehat{\mathbb{Q}}$, mais elle n'est pas à valeurs dans $\mathbb{Q}_+$. Nous montrerons que cette norme est la valeur absolue associée à une structure de corps totalement ordonné sur $\widehat{\mathbb{Q}}$.
- Si $|\ |$ est une valeur absolue ultramétrique sur $\mathbb{Q}$, nous montrerons que la norme prolongée sur $\widehat{\mathbb{Q}}$ est à valeurs dans $\mathbb{Q}_+$,

Pour des raisons techniques il est préférable de traiter séparément les deux cas.

Premier cas. On suppose $|\ | = |\ |_\infty$ sur Q. On *définit* le corps des nombres réels par $\mathbb{R} := \widehat{\mathbb{Q}}$. Nous allons prolonger à $\mathbb{R}$ la relation d'ordre de $\mathbb{Q}$ pour en faire un corps totalement ordonné. La valeur absolue de celui-ci coïncidera avec le prolongement de $|\ |_\infty$ défini ci-dessus (*cf.* la proposition-définition 71). Pour allèger les notations on notera $|\ |$ au lieu de $|\ |_\infty$.

Si la norme $|\ |$ sur $\mathbb{Q}$ est la valeur absolue usuelle, on a une version plus précise du lemme 69 (c'est le point clé).

Lemme 72. Soit $a = (a_n) \in \widetilde{\mathbb{Q}} \setminus \mathcal{I}$. Alors, il existe $\delta' \in \mathbb{Q}$, $\delta' > 0$, et $N' \in \mathbb{N}$ tels que l'une des deux assertions suivantes soient vraies :

1. pour tout $n \geqslant N'$, l'on a $a_n \geqslant \delta'$;
2. pour tout $n \geqslant N'$, l'on a $a_n \leqslant -\delta'$.

Démonstration. On applique le Lemme 69. On trouve $\delta > 0$ et $N \in \mathbb{N}$, tels que, pour tout $n \geqslant N$, l'on ait $|a_n| \geqslant \delta$. Puisque la suite (a_n) est de Cauchy, Il existe $N' \geqslant N$ tel que, pour tous $n, m \geqslant N'$, l'on ait $|a_n - a_m| \leqslant \delta/2$. On fixe $m \geqslant N'$. On a $|a_m| > \delta$, donc, par définition de $||$ sur $\mathbb{Q}$, l'on a : ou bien $a_m > \delta$, ou bien $a_m < -\delta$.
Soit $n \geqslant N'$. Dans le premier cas, on a $a_n = a_m + a_n - a_m$, donc :

$$a_n \geqslant a_m - |a_n - a_m| \geqslant a_m - \delta/2 \geqslant \delta/2.$$

Dans le second, on a $a_n \leqslant -\delta/2$. Il suffit donc de prendre $\delta' = \delta/2$. ∎

On va définir la notion de *nombre réel positif*. On note $\widetilde{\mathbb{Q}}_+$ le sous-ensemble de $\widetilde{\mathbb{Q}}$ formé par la réunion de $\mathcal{I}$ et de l'ensemble des suites (a_n) possédant la propriété suivante : il existe $N \in \mathbb{N}$ tel que, pour tout $n \in \mathbb{N}$, $a_n > 0$. *Par définition* un réel positif est un élément x de $\varphi(\widetilde{\mathbb{Q}}_+) \subset \mathbb{R}$. On note $x \geqslant 0$ un réel positif et $\mathbb{R}_+ := \varphi(\widetilde{\mathbb{Q}}_+)$ l'ensemble des réels positifs.

Lemme 73. Pour tout $a \in \widetilde{\mathbb{Q}}$, on a : $\varphi(a) \in \mathbb{R}^+ \Longleftrightarrow a \in \widetilde{\mathbb{Q}}_+$.

Démonstration. L'implication $a \in \widetilde{\mathbb{Q}}_+ \Rightarrow \varphi(a) \in \mathbb{R}^+$ résulte de la définition. Inversement, soit $a = (a_n) \in \widetilde{\mathbb{Q}}$ tel que $\varphi(a) \in \mathbb{R}_+$. On va montrer que $a \in \widetilde{\mathbb{Q}}_+$ en raisonnant par l'absurde. Supposons $a \notin \widetilde{\mathbb{Q}}_+$. Alors $a \notin \mathcal{I}$ et l'on peut appliquer le lemme 72. L'alternative 1 n'est pas possible (on aurait $a \in \widetilde{\mathbb{Q}}_+$). On a donc 2 : il existe $\delta' \in \mathbb{Q}$, $\delta > 0$ et $N' \in \mathbb{N}$ tels que, pour tout $n \geqslant N'$, $a_n \leqslant -\delta'$.
Soit b un représentant de $\varphi(a) \in \mathbb{R}_+$ (c'est-à-dire $\varphi(b) = \varphi(a)$) tel que $b \in \widetilde{\mathbb{Q}}_+$. Puisque $\varphi(a) = \varphi(b)$, on a $a - b \in \mathcal{I}$. Puisque $a \notin \mathcal{I}$, on a $b \notin \mathcal{I}$. Par suite, il existe $\delta'' \in \mathbb{Q}$, $\delta'' > 0$ et $N'' \in \mathbb{N}$ tels que, pour tout $n \geqslant N''$, $b_n \geqslant \delta''$. Par suite, pour tout $n \geqslant \max(N', N'')$, on a $b_n - a_n \geqslant \delta'' + \delta' > 0$. Ceci contredit $b - a \in \mathcal{I}$. ∎
L'étape suivante est d'utiliser $\mathbb{R}_+$ pour définir une relation d'ordre total sur $\mathbb{R}$ qui en fait un corps totalement ordonné.

Lemme 74. Muni de la relation $\leqslant$ définie par : pour tous $x, y \in \mathbb{R}$, $x \leqslant y$ si, et seulement si, $y - x \in \mathbb{R}^+$, $\mathbb{R}$ est un corps commutatif totalement ordonné.

Démonstration. D'après la proposition 67 de la page 492, il suffit de vérifier :

$$\mathbb{R}_+ + \mathbb{R}_+ \subset \mathbb{R}_+,\ \ \mathbb{R}_+.\mathbb{R}_+ \subset \mathbb{R}_+,\ \ \mathbb{R}_+ \cap (-\mathbb{R}_+) = \varnothing,\ \ \mathbb{R}_+ \cup (-\mathbb{R}_+) = \mathbb{R}.$$

Puisque φ est un homomorphisme d'anneaux, il suffit donc, compte tenu du lemme 73, de vérifier :

$$\widetilde{\mathbb{Q}}_+ + \widetilde{\mathbb{Q}}_+ \subset \widetilde{\mathbb{Q}}_+,\ \ \widetilde{\mathbb{Q}}_+.\widetilde{\mathbb{Q}}_+ \subset \widetilde{\mathbb{Q}}_+,\ \ \widetilde{\mathbb{Q}}_+ \cap (-\widetilde{\mathbb{Q}}_+) = \varnothing,\ \ \widetilde{\mathbb{Q}}_+ \cup (-\widetilde{\mathbb{Q}}_+) = \mathbb{R}.$$

Ces vérifications sont sans difficulté. Nous les laissons au lecteur. ∎

On définit la valeur absolue du corps totalement ordonnée $\mathbb{R}$ par $|x| := \max(x, -x)$. On vérifie facilement le résultat suivant.

Proposition 75. Si $a = (a_n) \in \widetilde{\mathbb{Q}}$ représente $x \in \mathbb{R}$ (i. e. $\varphi(a) = x$), on a $|a| := (|a_n|) \in \widetilde{\mathbb{Q}}$ et $|x| = \varphi(|a|)$. En particulier $\varphi(|a|)$ est indépendant du choix du représentant a.

Cette proposition montre que la norme sur $\mathbb{R}$ définie dans la proposition-définition 71 et la valeur absolue associée à la relation d'ordre coïncident.

Lemme 76.

(i) Une suite de réels (x_n) converge vers ℓ dans $\mathbb{R}$ si, et seulement si,, pour tout $\varepsilon \in \mathbb{Q}$, $\varepsilon > 0$, il existe $N \in \mathbb{N}$, tel que, pour tout $n \geqslant N$, $|x_n - \ell| < \varepsilon$.

(ii) Soient $x \in \mathbb{R}$ et $a = (a_n)$ un représentant. Alors la suite (a_n) converge vers x dans $\mathbb{R}$.

Démonstration.

(i) Si la suite converge dans $\mathbb{R}$ la condition est évidemment satisfaite. Inversement, supposons que la condition est satisfaite. Soit $\omega \in \mathbb{R}$, $\omega > 0$. Soit $b = (b_n)$ un représentant de ω. D'après le lemme 72 (on est nécessairement dans le cas 1), il existe $N' \in \mathbb{N}$ et $\varepsilon \in \mathbb{Q}$, $\varepsilon > 0$ tels que, pour tout $n \geqslant N'$, $b_n \geqslant \varepsilon$. ceci implique $\varepsilon \leqslant \omega$. En appliquant la condition, on en déduit qu'il existe $N \geqslant N'$ tel que, pour tout $n \geqslant N$, l'on ait $|x_n - \ell| < \varepsilon \leqslant \omega$.

(ii) C'est une conséquence immédiate de (i). ∎

Théorème 77. Le corps totalement ordonné $\mathbb{R}$ est complet et $\mathbb{Q}$ est dense dans $\mathbb{R}$.

Démonstration. Montrons d'abord la densité. Soit $]x_0, x_1[$ un intervalle ouvert de $\mathbb{R}$, avec $x_0 < x_1$. Soit $y := \frac{x_0+x_1}{2}$. D'après le lemme 76, il existe une suite de rationnels $b := (b_n)$ convergeant vers y dans $\mathbb{R}$. Il existe donc $N \in \mathbb{N}$, tel que, pour tout $n \geqslant N$, l'on ait $|y_n - y| < \frac{x_1-x_0}{2}$. Alors $y_n \in]x_0, x_1[$. On a ainsi montré que tout intervalle ouvert non vide de $\mathbb{R}$ contient un rationnel et $\mathbb{Q}$ est dense dans $\mathbb{R}$.
Montrons que $\mathbb{R}$ est complet. Soit (x_n) une suite de Cauchy de $\mathbb{R}$. Puisque $\mathbb{Q}$ est dense dans $\mathbb{R}$, on peut trouver, pour chaque $n \in \mathbb{N}$, un rationnel $b_n \in \mathbb{Q}$ tel que $x_n - b_n \leqslant 2^{-n}$. Si $n, m > N$, on a :

$$|b_n - b_m| \leqslant |x_n - x_m| + 2^{-n} + 2^{-m} \leqslant 2^{-N}.$$

On en déduit que la suite $b := (b_n)$ est une suite de Cauchy de $\mathbb{Q}$. Elle converge dans $\mathbb{R}$ vers ℓ. On vérifie que la suite (x_n) converge aussi vers ℓ. Ainsi $\mathbb{R}$ est *complet*. ∎

Le corps totalement ordonné $\mathbb{Q}$ est archimédien. Puisque $\mathbb{Q}$ est dense dans $\mathbb{R}$, on en déduit facilement que le corps totalement ordonné $\mathbb{R}$ est archimédien.

Corollaire 78. Le corps $\mathbb{R}$ est archimédien.

Deuxième cas. Soit $|\ |$ une norme ultramétrique sur $\mathbb{Q}$ (par exemple $|\ |_p$ pour un nombre premier p). Nous allons prolonger $|\ |$ en une norme sur $\widehat{\mathbb{Q}}$ à valeurs dans $\mathbb{Q}_p$.

Lemme 79. ,

(i) Soit $a = (a_n) \in \widetilde{\mathbb{Q}} \setminus \mathcal{I}$. Alors la suite $(|a_n|)$ est *constante à partir d'un certain rang*.

(ii) Soient $a, b \in \widetilde{\mathbb{Q}}$. On suppose $a - b \in \mathcal{I}$. Alors les deux suites $(|a_n|)$ et $(|b_n|)$ sont convergentes dans $\mathbb{Q}$ et ont la même limite.

Démonstration.

(i) Soit $a = (a_n) \notin \mathcal{I}$. D'après le lemme 69, il existe $\delta \in \mathbb{Q}$, $\delta > 0$ et $N \in \mathbb{N}$ tels que, pour tout $n \geqslant N$, $|a_n| > \delta$. La suite (a_n) étant une suite de Cauchy de $\mathbb{Q}$, il existe $N' \geqslant N$ tel que, pour tous $n, m \geqslant N'$, l'on ait $|a_m - a_n| < \delta/2$. Par suite $|a_m - a_n| < |a_n|$. Puisque $|\ |$ est ultramétrique, on en déduit $|a_m| = |a_n|$. En particulier, pour tout $m \geqslant N'$, $|a_m| = |a_{N'}|$.

(ii) Si $a \in \mathcal{I}$, on a aussi $b \in \mathcal{I}$ et le résultat est vrai. Si $a \notin \mathcal{I}$, on a aussi $b \notin \mathcal{I}$.
D'après (i), les suites $|a|$ et $|b|$ sont constantes à partir d'un certain rang et, puisque $a - b \in \mathcal{I}$, on a $|a| - |b| \in \mathcal{I}$ et elles ont la même limite. ∎

Soit $x \in \widehat{\mathbb{Q}}$. D'après le lemme 79 ci-dessus, si $x = \varphi(a)$, avec $a = (a_n) \in \widetilde{\mathbb{Q}}$, la limite de la suite $|a|$ est indépendante du choix de a. On peut donc *définir* $|x|$ par $|x| := \lim\limits_{n \to +\infty} |a_n|$.

On vérifie facilement que $|\ |$ ainsi définie est une norme sur $\widehat{\mathbb{Q}}$: la multiplicativié de la norme et l'inégalité ultramétrique passent à la limite.

Théorème 80.
Le corps $\widehat{\mathbb{Q}}$ muni de la norme $|\ |$ définie ci-dessus est complet et $\mathbb{Q}$ est dense dans $\widehat{\mathbb{Q}}$.

Démonstration. Montrons d'abord la densité. Soit $x \in \widehat{\mathbb{Q}}$. Soit $a = (a_n)$ un représentant de x. On vérifie que $\lim\limits_{n \to +\infty} |x - a_n| = 0$, donc la suite a converge vers x dans $\widehat{\mathbb{Q}}$. Soit $r > 0$. Il existe $N \in \mathbb{N}$ tel que, pour $n \geqslant N$, $a_n \in B(x, r)$. Puisque $a_n \in \mathbb{Q}$, $\mathbb{Q}$ est dense dans $\widehat{\mathbb{Q}}$. Pour montrer que $\widehat{\mathbb{Q}}$ et complet, on procède comme dans le cas de $\mathbb{R}$ (*cf.* la preuve du théorème 77). ∎

Soit p un nombre premier. Si $|\ | = |\ |_p$, on dit que $\widehat{\mathbb{Q}}$ est le corps des nombre p-adiques et l'on note $\mathbb{Q}_p := \widehat{\mathbb{Q}}$.

Construction de $\mathbb{R}$ par complétion

Dans le premier cas ci-dessus, nous avons construit un corps totalement ordonné complet archimédien $\mathbb{R}$. Pour montrer qu'il satisfait les axiomes des nombres réels, il reste à prouver que la propriété de la borne supérieure est satisfaite. Le lecteur vérifiera que la preuve que nous détaillons ci-dessous revient essentiellement (en écrivant les choses plus abstraitement) à montrer que si un corps totalement ordonné est archimédien et complet, alors il possède la propriété de la borne supérieure.

Définition 19. Soient A et B deux parties de $\mathbb{R}$, on dit que (A, B) est un couple *d'ensembles adjacents* si :

(i) pour tout $(a, b) \in A \times B$, on a $a \leqslant b$;

(ii) pour tout $\varepsilon > 0$, il existe $(a, b) \in A \times B$ tel que $b - a \leqslant \varepsilon$.

Proposition 81. Soit (A, B) un couple *d'ensembles adjacents*. Il existe $c \in \mathbb{R}$ unique tel que, pour tout $(a, b) \in A \times B$, l'on ait $a \leqslant c \leqslant b$.

Démonstration.
On utilise le fait que $\mathbb{R}$ est complet. Pour tout $n \in \mathbb{N}^*$, il existe $(a_n, b_n) \in A \times B$ tel que $b_n - a_n \leqslant \varepsilon$. Montrons que $(a_n)_{n\in\mathbb{N}^*}$ est une *suite de Cauchy*. Pour tous $n, m \in \mathbb{N}^*$, on a $a_n \leqslant b_m$, donc $a_n - a_m < b_m - a_m \leqslant 1/n$. En échangeant n et m, on en déduit $|a_n - a_m| \leqslant \max(1/n, 1/m)$. Puisque $\mathbb{R}$ est *complet*, la suite de Cauchy $(a_n)_{n\in\mathbb{N}^*}$ admet une limite c. On a $\lim\limits_{n\to+\infty}(b_n - a_n) = 0$, donc $\lim\limits_{n\to+\infty} b_n = c$. Pour tout $(a, b) \in A \times B$, on a, pour tout $n \in \mathbb{N}^*$, $a_n \leqslant b$ et $a \leqslant b_n$. On en déduit $a \leqslant c \leqslant b$. On a ainsi prouvé l'existence de c.
Montrons l'unicité. Soit $c' \in \mathbb{R}$ satisfaisant la condition. Si $c' \neq c$, On peut supposer (quitte à échanger c et c') que $c \leqslant c'$. Alors, pour tout $n \in \mathbb{N}^*$, $a_n \leqslant c \leqslant c' \leqslant b_n$, donc $0 \leqslant c' - c \leqslant 1/n$. Il en résulte, en passant à la limite, $c' = c$. ∎

Exemple. Soit (A, B) une *coupure* de $\mathbb{Q}$ (*cf.* la définition 10 de la page 480). Montrons que c'est une paire d'ensembles adjacents. La condition (i) est satisfaite. Montrons que la condition (ii) l'est. On choisit $a_0 \in A$ et $b_0 \in B$. Si $a_0 = b_0$, c'est teminé. Sinon, soit $c_0 := \frac{a_0+b_0}{2}$. On a $c_0 \in A$ ou $c_0 \in B$. Supposons $c_0 \in B$ (l'autre cas est analogue). On pose $a_1 = a_0$ et $b_1 = c_0$. Alors $a_1 \in A$, $b_1 \in B$ et $0 \leqslant b_1 - a_1 \leqslant \frac{b_0-a_0}{2}$. Si $a_1 = b_1$, c'est teminé, sinon on poursuit ce procédé de dichotomie. Le résultat s'en déduit.
Ainsi, d'après la proposition 81 ci-dessus, on peut associer à une coupure un nombre réel unique. C'est une comparaison entre deux constructions de $\mathbb{R}$.

Proposition 82.
Tout sous-ensemble majoré non vide de $\mathbb{R}$ admet une borne supérieure.

Démonstration. On utilise le fait que $\mathbb{R}$ est archimédien et la proposition 81 ci-dessus. Soit $A \subset \mathbb{R}$ majoré non vide. On note B l'ensemble des majorants (au sens large) de A. par hypothèse B n'est pas vide et pour tous $(a, b) \in A \times B$, on a $a \leqslant b$. Nous allons montrer que (A, B) est un couple d'ensembles adjacents.
Soient $a_0 \in A$ et $b_0 \in B$. Soit $\varepsilon > 0$. On définit $J \subset \mathbb{N}$ par $J = \{n \in \mathbb{N} \mid b_0 - n\varepsilon \in B$. On a $0 \in J$ donc $J \neq \varnothing$. On a pour tout $n \in J$: $a_0 \leqslant b_0 - n\varepsilon$, donc $n \leqslant (b_0 - a_0)/\varepsilon$. Ainsi, en utilisant la proriété d'Archimède, J est *fini* et il admet donc un plus grand élément n_0. Puisque $n_0 + 1 \neq J$, on a $b_0 - n_0\varepsilon - \varepsilon \notin B$. Par suite, si l'on pose $b := b_0 - n_0\varepsilon$, on a $b \in B$ et $b - \varepsilon \notin B$. Il existe donc $a \in A$ tel que $b - \varepsilon < a$, c'est-à-dire $b - a < \varepsilon$. Nous avons donc montré que A et B sont adjacents. D'après la proposition 81 ci-dessus , il existe $M \in \mathbb{R}$ défini par $a' \leqslant M \leqslant b'$, pour tous $(a', b') \in A \times B$. Ce réel M est le plus petit majorant de a, c'est donc la borne supérieure de A. ∎

Nous avons ainsi achevé la construction de $\mathbb{R}$ par complétion de $\mathbb{Q}$ à l'aide des suites de Cauchy de $\mathbb{Q}$, la troisième (et dernière ...) proposée dans ce module.

Constructions des corps p-adiques $\mathbb{Q}_p$ par complétion

Dans tout ce paragraphe p est un nombre premier fixé. Il reste à montrer que le corps $\mathbb{Q}_p$ construit ci-dessus par complétion, dans le deuxième cas, est isomorphe à celui construit algébriquement dans la partie 2. Nous donnons les étapes d'une preuve « explicite » (les démonstrations sont proposées en exercice). Il est aussi possible de donner une preuve « abstraite ». Nous le ferons au paragraphe suivant (*cf.* le corollaire 88). Nous rappelons que, selon nos conventions de lecture, nous avons « oublié » la définition de $\mathbb{Z}_p$. Le premier pas consiste à le définir.

Proposition et définition 83.
Le sous-ensemble $\mathbb{Z}_p \subset \mathbb{Q}_p$ défini par $\mathbb{Z}_p := \{x \in \mathbb{Q}_p \mid |x|_p \geqslant 0\}$ est un sous-anneau de $\mathbb{Q}_p$. On l'appelle *l'anneau des entiers du corps normé* $\mathbb{Q}_p$. Il possède un seul idéal maximal qui est $\mathcal{M} := \{x \in \mathbb{Q}_p \mid |x|_p > 0\}$. Le corps $\mathbb{Z}_p/\mathcal{M}$ est appelé *le corps résiduel* de $\mathbb{Q}_p$.

Démonstration. Le fait que $\mathbb{Z}_p$ soit un sous-anneau et $\mathcal{M}$ un idéal résulte des propriétés de la norme $|\ |_p$. Si $x \in \mathbb{Z}_p$ n'appartient pas à $\mathcal{M}$, on a $|x|_p = 1$, $x \neq 0$, $xx^{-1} = 1$, donc $|x|_p|x^{-1}|_p = |xx^{-1}|_p = 1$ et $|x^{-1}|_p = 1$. Par suite $x^{-1} \in \mathbb{Z}_p$. Ainsi l'idéal $\mathcal{M}$ est maximal (si un idéal de $\mathbb{Z}_p$ contient x, il contient x^{-1}, donc 1). ∎

On a $|\mathbb{Q}_p^*|_p = |\mathbb{Q}^*|_p = \mathbb{Z}$, donc : $\mathcal{M} = \{x \in Q^p \mid |x|_p \geqslant 1\} = p\mathbb{Z}_p$.

On admet le résultat suivant. La preuve de la surjectivité utilise la *densité* de $\mathbb{Q}$ dans Q_p (voir l'exercice III.3.40 de la page 529).

Lemme 84. Pour tout $n \in \mathbb{N}^*$, l'application naturelle $\mathbb{Z}/p^n\mathbb{Z} \to \mathbb{Z}_p/p^n\mathbb{Z}_p$ est un isomorphisme d'anneaux.

On a défini plus haut la limite projective A des $A_n := \mathbb{Z}/p^n\mathbb{Z}$ et on l'a muni d'une structure d'anneau topologique. On admet le résultat suivant. La preuve de la surjectivité utilise le fait que $\mathbb{Q}_p$ est *complet* (voir l'exercice III.3.40 de la page 529).

Proposition 85. L'application qui à $x \in \mathbb{Z}_p$ associe la suite (x_n) de ses images modulo p^n induit un isomorphisme d'anneaux topologiques de $\mathbb{Z}_p$ sur A.

La proposition ci-dessus entraîne l'isomorphisme des *corps* p-adiques obtenus par les deux constructions.

Unicité de corps complets à isomorphisme près

Théorème 86. Soient $(K_1, \| \|_1)$ et $(K_2, \| \|_2)$ deux corps normés complets contenant $\mathbb{Q}$. On suppose que les deux normes induites sur $\mathbb{Q}$ coïncident et que $\mathbb{Q}$ est dense dans K_1 et dans K_2. Alors il existe un unique isomorphisme de corps normés de K_1 sur K_2.

Démonstration. La preuve et facile. Elle est basée sur les suites de Cauchy de $\mathbb{Q}$. Par définition, un isomorphisme de corps normés est un isomorphisme de corps $\theta : K_1 \to K_2$ tel que, pour tout $x \in K_1$, l'on ait $\left\|\theta(x)\right\|_2 = \|x\|_1$. On rappelle qu'un isomorphisme de corps contenant $\mathbb{Q}$ induit l'identité sur $\mathbb{Q}$.

Montrons d'abord *l'unicité*. Soient θ et θ' deux isomorphismes de corps normés. Soit $x \in K_1$. Puisque $\mathbb{Q}$ est dense dans K_1, il existe une suite (x_n) de $\mathbb{Q}$ telle que $x = \lim\limits_{n\to+\infty} x_n$. C'est une suite de Cauchy de $\mathbb{Q}$, donc de K_2 et elle converge dans K_2 vers y, puisque K_2 est complet. On a, pour tout $n \in \mathbb{N}$, $\theta(x_n) = \theta'(x_n) = x_n$.Un isomorphisme de corps normé est *continu*, donc $\theta(x) = y$ et $\theta(x') = y$.

Montrons *l'existence*. Soit $x \in K_1$. Puisque $\mathbb{Q}$ est dense dans K_1, il existe une suite (x_n) de $\mathbb{Q}$ telle que $x = \lim\limits_{n\to+\infty} x_n$. C'est une suite de Cauchy de $\mathbb{Q}$, donc de K_2 et elle converge dans K_2 vers y. Soit (x'_n) une autre suite de Cauchy de $\mathbb{Q}$ telle que $x = \lim\limits_{n\to+\infty} x'_n$. Elle converge dans K_2 vers y'. La suite $(x_n - x'_n)$ est une suite de $\mathbb{Q}$, convergeant vers 0 dans $\mathbb{Q}$. Par suite $y = y'$. On peut donc *définir* une application $\theta : K_1 \to K_2$ en associant à $x \in \mathbb{K}_1$ la limite $\theta(x)$ dans K_2 de l'une quelconque de suites de rationnels qui convergent vers x dans K_1.

Soient (x_n) et (x'_n) des suites de $\mathbb{Q}$ convergeant respectivement vers x et x'.

- Les suites $(\|x_n\|_1)$ et $(\|x_n\|_2)$ sont égales. La suite $(\|x_n\|_1)$ converge vers $\|x\|_1$ dans K_1 et vers $\|\theta(x)\|_2$ dans K_2.

  Par passages à la limite dans K_1 et K_2, on en déduit $\|\theta(x)\|_2 = \|x\|_1$.
- On considère la suite $(x_n + x'_n)$ de $\mathbb{Q}$. Par passages à la limite dans K_1 et K_2, on obtient $\theta(x + x') = \theta(x) + \theta(x')$.
- On considère la suite $(x_n x'_n)$ de $\mathbb{Q}$. Par passages à la limite dans K_1 et K_2, on obtient $\theta(xx') = \theta(x)\theta(x')$.

Pour terminer la preuve, il reste à montrer que θ est *surjective*. Soit $y \in K_2$. Puisque $\mathbb{Q}$ est dense dans K_2, il existe une suite (y_n) de $\mathbb{Q}$ telle que $y = \lim\limits_{n\to+\infty} y_n$. C'est une suite de Cauchy de $\mathbb{Q}$, donc de K_1 et elle converge dans K_1 vers x. D'après la définition de θ, on a $\theta(x) = y$. ∎

La condition de densité de $\mathbb{Q}$ est évidemment essentielle : les corps normés complets $\mathbb{R}$ et $\mathbb{C}$ ne sont pas isomorphes. (Ils ne sont même pas isomorphes en tant que corps !)
Le théorème 86 ci-dessus permet de donner une nouvelle preuve de « l'unicité de $\mathbb{R}$ ».

Corollaire 87. Soient $\mathcal{R}_1$ et $\mathcal{R}_2$ deux corps vérifiant les axiomes des nombres réels. Il existe un unique isomorphisme de corps ordonnés de $\mathcal{R}_1$ sur $\mathcal{R}_2$.

Démonstration. On définit des normes $\| \|_1$ et $\| \|_2$ sur $\mathcal{R}_1$ et $\mathcal{R}_2$ en utilisant la valeur absolue définie par l'ordre. Elles induisent sur $\mathbb{Q}$ la relation d'ordre ordinaire (*cf.* [L1, IV.1, 1.2.1]). En utilisant le théorème 86, on obtient alors un isomorphisme de corps normés $\varphi : \mathcal{R}_1 \to \mathcal{R}_2$. L'ensemble $\mathbb{Q}_+$ est dense dans l'ensemble des éléments positifs de $\mathcal{R}_1$ et dans l'ensemble des éléments positifs de $\mathcal{R}_2$ (cette propriété est vraie dans $\mathbb{R}$, donc dans tout corps vérifiant les axiomes de nombres réels). Donc φ (qui induit l'identité sur $\mathbb{Q}_+$) induit un isomorphisme entre ces deux ensembles. Par suite φ est une application croissante. ∎

On déduit facilement du théorème 86 le résultat suivant.

Corollaire 88. Soit $(K, \| \|)$ un corps normé complet contenant $\mathbb{Q}$. On suppose que $\mathbb{Q}$ est dense dans K et que $| \, |$ induit sur $\mathbb{Q}$ une valuation p-adique $| \, |_p$. Alors il existe un unique isomorphisme de corps normés entre K et $\mathbb{Q}_p$.

4.1.2 Corps normés et complétions

Convention de lecture. Dans cette partie (et dans toute la suite) on « revient à la normale » et l'on suppose que le corps $\mathbb{R}$ et ses propriétés usuelles sont connus. C'est en particulier *nécessaire* pour la définition 20 ci-dessous.

Corps normés

La notion de norme sur un corps sous la forme axiomatique qui suit a été introduite par le mathématicien Hongrois J. Kürschák en 1912.

Définition 20. Soit K un corps. Une *norme* (ou *valeur absolue*) sur K est une application $x \mapsto |x|$ de K dans $\mathbb{R}_+$ vérifiant les trois propriétés suivantes :

(i) $\forall x \in K, \ \ |x| = 0 \Leftrightarrow x = 0$;

(ii) $\forall x, y \in K, \ \ |xy| = |x||y|$;

(iii) $\forall x, y \in K, \ \ |x + y| \leqslant |x| + |y|$.

Une norme et dite *ultramétrique* si elle vérifie la condition

(iii') $\forall x, y \in K, \ \ |x + y| \leqslant \sup\{|x|, |y|\}$.

qui est plus forte que la condition (iii).

Une norme ultramétrique est aussi dite *non-archimédienne*. Une norme qui n'est pas ultramétrique est dite *archimédienne*. Une norme archimédienne $| \, |$, sur le corps K, possède la propriété suivante. Pour tous $x, y \in K$, $x \neq 0$, il existe $n \in \mathbb{N}^*$ tel que $|nx| > |y|$. Une preuve de cette assertion est proposée dans l'exercice III.3.39 de la page 529. Ainsi la terminologie est en accord avec celle adoptée antérieurement pour $\mathbb{R}$.

Exercice 24.

Montrer que toute norme sur un corps fini est triviale.

Solution. Soient $K = F^q$ un corps fini et $| \, |$ une norme sur K. Tout $x \in K$ non nul est racine de l'unité : $x^{q-1} = 1$. Alors $1 = |1| = |x|^{q-1} = |x|^{q-1}$. Puisque $|x| \in \mathbb{R}^+$, on a $|x| = 1$. La norme est donc triviale.

Exercice 25.

Soit $(K, ||)$ un corps normé. Montrer, pour $a \in K$:

(i) $|a| < 1 \ \Rightarrow \lim\limits_{n \to +\infty} a^n = 0$.

(ii) $|a| < 1 \ \Rightarrow \lim\limits_{n \to +\infty} a^n/(1 + a^n) = 0$.

(iii) $|a| > 1 \ \Rightarrow \lim\limits_{n \to +\infty} a^n/(1 + a^n) = 1$.

Solution.

(i) Résulte de $|(|a^n) = |a|^n$.

(ii) En utilisant $1 + a^n = 1 - (-a^n)$ et l'inégalité triangulaire, on obtient :

$$1-|a|^n \leqslant |1+a^n| \leqslant 1+|a|^n \qquad \text{donc} \qquad \lim_{n \to +\infty} |1+a^n| = 1 \quad \text{et} \quad \lim_{n \to +\infty} a^n/(1+a^n) = 0.$$

(iii) On a $1 - \frac{a^n}{1+a^n} = \frac{1}{1+a^n} = \frac{a^{-n}}{1+a^{-n}}$ et $\lim\limits_{n \to +\infty} a^{-n}/(1 + a^{-n}) = 0$, d'après (ii).

Exercice 26.
Soit $(K, |\ |)$ un corps normé. Soit $(x_n)_{n\in\mathbb{N}}$ une suite de Cauchy de K. Montrer que la suite $|x_n|$ converge dans $\mathbb{R}_+$. Montrer que si la norme $|\ |$ est ultramétrique, alors cette suite est constante à partir d'un certain rang.
Solution. On a $||x_{n+m}| - |x_n|| \leqslant |x_{n+m} - x_n|$, donc $(|x_n|)$ est une suite de Cauchy de $\mathbb{R}_+$. Puisque $\mathbb{R}_+$ est complet, elle converge. Si la norme est ultramétrique, on raisonne comme dans la preuve du lemme 79.

Définition 21. Deux normes sur un corps K sont dites équivalentes si elles définissent la même topologie (c'est-à-dire le même ensemble d'ouverts).

Proposition 89. Deux normes $|\ |_1$ et $|\ |_2$ sur un corps K sont *équivalentes* si, et seulement s'il existe $\alpha \in \mathbb{R}_+^*$ tel que $|x|_1 = |x|_2^\alpha$, pour tout $x \in K^*$.

Démonstration. Si $|\ |$ est une norme sur un corps K, alors $|x| < 1$ si, et seulement si, la suite de terme général x^n tend vers 0 quand n tend vers $+\infty$. Si une suite tend vers 0 pour une norme, elle tend vers 0 pour une norme équivalents. Par suite, si les deux normes $|\ |_1$ et $|\ |_2$ sont équivalentes, on a $|x|_1 < 1 \Leftrightarrow |x|_2 < 1$.
Soient $x, y \in \mathbb{K}^*$. On a :

$$\forall n \in \mathbb{Z},\ m \in \mathbb{N}^*, \quad |x|_1^n < |y|_1^m \Leftrightarrow |x^n/y^m|_1 < 1 \Leftrightarrow |x^n/y^m|_2 < 1 \Leftrightarrow |x|_2^n < |y|_2^m.$$

Donc : $n\log|x|_1 < m\log|y|_1 \Leftrightarrow n\log|x|_2 < m\log|y|_2$ et :

$$\left\{r \in \mathbb{Q} \mid r < \frac{\log|y|_1}{\log|x|_1}\right\} = \left\{r \in \mathbb{Q} \mid r < \frac{\log|y|_2}{\log|x|_2}\right\}.$$

Par suite : $\frac{\log|y|_1}{\log|x|_1} = \frac{\log|y|_2}{\log|x|_2}$ (les réels $\frac{\log|y|_1}{\log|x|_1}$ et $\frac{\log|y|_2}{\log|x|_2}$ définissent la même section commençante).
En fixant x, on en déduit que la fonction $y \mapsto \frac{\log|y|_1}{\log|y|_2}$ est *constante* sur K^* : $\log|y|_1 = \alpha\log|y|_2$, avec $\alpha \in \mathbb{R}_+^*$. Ainsi $|y|_1 = |y|_2^\alpha$ ∎

Rappelons que l'anneau premier d'un corps commutatif K est l'image A_K de l'application $\mathbb{Z} \to K$ définie par $n \in \mathbb{Z} \mapsto n.1_K$ (où 1_K est l'élément neutre de la multipplication de K). Il est isomorphe à $\mathbb{Z}$ si K est de caractéristique nulle et il est fini sinon.
Nous admettons le résultat suivant (une preuve fait l'objet de l'exercice III.3.39 de la page 529).

Proposition 90. Soit $(K, |\ |$ un corps normé. Alors $|\ |$ est ultramétrique si, et seulement si, elle est bornée sur l'anneau premier A_K.

Complétion d'un corps normé

Moyennant quelques adaptations la méthode utilisée pour compléter le corps $\mathbb{Q}$ permet de compléter un corps normé quelconque.
Soit $(K, |\ |)$ un corps normé. On note comme dans le cas de $K = \mathbb{Q}$, $\widetilde{K}$ l'ensemble des suites de Cauchy de K et $\mathcal{I}$ l'ensemble des suites tendant vers 0 On construit $\widehat{K} = \widetilde{K}/\mathcal{I}$ en utilisant des variantes des lemmes 69 et 70. Il reste à prolonger la norme et à prouver que le corps obtenu est complet.

On utilise les résultats de l'exercice 26. Soient $a := (a_n)$ et $b = (b_n)$ deux suites de Cauchy de K. Les suites $(|a_n|)$ et $(|b_n|)$ convergent dans $\mathbb{R}_+$ vers c et d. Si $a - b \in \mathcal{I}$, en utilisant $||a_n| - |b_n|| \leqslant |a_n - b_n|$, on montre que $c = d$. On peut donc définir une application $|\ | : \widehat{K} \to \mathbb{R}_+$ par : $|\text{classe de } (a^n)| := \lim\limits_{n\to+\infty} |a_n|$. On vérifie facilement que c'est une norme sur $\widehat{K}$ prolongeant la norme initiale sur K. Si la norme initiale sur K est ultramétrique, on vérifie que cette norme est aussi ultramétrique en procédant comme dans la preuve du lemme 79.

On montre que $(\widehat{K}, |\ |)$ est complet comme dans la preuve du théorème 77.

Proposition 91. Soit $(K, |\ |)$ un corps normé. On prolonge au corps $\widehat{K} = \widetilde{K}/\mathcal{I}$ la norme $|\ |$ comme ci-dessus. Le corps $\widehat{K}$ est *complet* pour cette norme et il contient[a] K comme sous-corps dense.

[a]Plus précisément, il contient un corps normé isomorphe à K.

4.1.3 Compléments sur les normes sur $\mathbb{Q}$

Théorème 92 (Théorème d'Ostrowski). Une norme non triviale sur $\mathbb{Q}$ est équivalente à la norme usuelle $|\ |_\infty$ ou à la norme p-adique $|\ |_p$ pour un nombre premier p.

Les seuls « mondes possibles » sont le « monde réel » et le « monde p-adique ». L'ordre ne règne que dans le premier.

Démonstration. Soit $\big|\ \big|$ une norme *non triviale* sur $\mathbb{Q}$. On considère deux cas.

1. On suppose que, pour tout entier $m \in \mathbb{N}$, $\big|m\big| \leqslant 1$. C'est équivalent à $\big|\ell\big| \leqslant 1$, pour tout nombre premier ℓ. Puisque la norme n'est pas triviale, il exist un nombre premier p tel que $\big|p\big| < 1$. Nous allons montrer que si ℓ est un nombre premier différent de p, alors $\big|\ell\big| = 1$. On raisonne par l'absurde. Supposons qu'il existe un nombre premier $q \neq p$ tel que $\big|q\big| < 1$. D'après le théorème de Bézout, il existe, pour tout $n \in \mathbb{N}$ $a_n, b_n \in \mathbb{Z}$ tel que $a_n p^n + b_n q^n = 1$. On a
$$1 = \big|1\big| = \big|a_n p^n + b_n q^n = 1\big| \leqslant \big|a_n\big|\big|p\big|^n + \big|b_n\big|\big|q\big|^n \leqslant \big|p\big|^n + \big|q\big|^n,$$
ce qui est impossible pour n assez grand, puisque $\lim\limits_{n\to+\infty} \left(\big|p\big|^n + \big|q\big|^n\right) = 0$.

   Soit $x \in \mathbb{Z}$. Si $x = p^m y$, avec $y \in \mathbb{Z}$ non divisible par p, on a $\big|x\big| = \big|p\big|^m$, donc les normes $\big|\ \big|$ et $|\ |_p$ sont équivalentes.
2. On suppose qu'il existe un entier $k \in \mathbb{N}$ tel que $\big|k\big| > 1$. On a $\big|1\big| = 1$, donc $\big|k\big| \leqslant k$ et il existe $\alpha \in]0, 1]$ tel que l'on ait $\big|k\big| = k^\alpha$. On peut alors montrer que, pour tout $x \in \mathbb{Q}$, l'on a $\big|x\big| = |x|_\infty^\alpha$. Nous l'admettrons. Une preuve fait l'objet de l'exercice III.3.46 de la page 530. Ainsi $\big|\ \big|$ est dans ce cas équivalente à la norme usuelle. ∎

Le résultat suivant est une conséquence immédiate de la décomposition d'un nombre entier en facteurs premiers. Il et néanmoins important. Il justifie la normalisation adoptée pour $|\ |_p$.

Théorème 93 (Formule du produit). Si $x \in \mathbb{Q}^*$, alors $|x|_\infty \prod\limits_{p \text{ premier}} |x|_p = 1$.

Dans le produit infini, on convient de n'écrire que les termes différents de 1.

4.2 Comparaison entre les anneaux $\mathbb{C}[X]$ et $\mathbb{Q}_p$

Le lecteur a pu deviner qu'il y a des analogies entre les anneaux $\mathbb{C}[X]$ et $\mathbb{Q}_p$. L'idée remonte à Leopold Kronecker. C'est la base du travail de son élève Hensel et elle a conduit celui-ci à la définition des nombres p-adiques. Nous allons décrire certaines de ces analogies.
Les deux anneaux sont intègres et *principaux*. On note respectivement $Spec\,\mathbb{C}[X]$ et $Spec\,\mathbb{Z}$ l'ensemble de leurs *idéaux maximaux*.

- Les idéaux maximaux de $\mathbb{C}[X]$ sont les $< X-\alpha >= \mathbb{C}[X](X-\alpha)$, où $\alpha \in \mathbb{C}$. Si l'on identifie $< X-\alpha >$ et α, on obtient une identification entre $Spec\,\mathbb{C}[X]$ et $\mathbb{C}$. On peut ajouter à $\mathbb{C}$ un point à l'infini : on obtient la sphère de Riemann. On voit $\mathbb{C}[X]$, et plus généralement $\mathbb{C}(X)$, comme des anneaux de fonctions « sur » la sphère de Riemann :
$$P^1(\mathbb{C}) = Spec\,\mathbb{C}[X] \cup \{\infty\}.$$
- Les idéaux maximaux de $\mathbb{Z}$ sont les $< p >= \mathbb{Z}p$, où p est un nombre premier. On peut ajouter à $Spec\,\mathbb{Z}$ un *nombre premier à l'infini*. On écrit $Spec\,\mathbb{Z} \cup \{\infty\}$. On essaie de voir Z comme « un anneau de fonctions sur une courbe ».

On peut voir *les nombres premiers p comme analogues des polynômes $X-\alpha$*. Un nombre entier n est l'analogue d'un polynôme P. Sa décomposition en facteurs premiers $n = \pm p_1^{m_1} \dots p_r^{m_r}$ est l'analogue de la décomposition $P(X) = \beta(X-\alpha_1)^{m_1} \dots (X-\alpha_r)^{m_r}$ ($\beta \in \mathbb{C}$; $m_1, \dots, m_r \in \mathbb{N}^*$).
En passant aux corps des fractions, l'analogue du corps $\mathbb{Q}$ est le corps des fractions rationnelles $\mathbb{C}(X)$. Si l'on travaille dans le cadre des fonctions analytiques, pour étudier la fraction rationnelle $\frac{1}{1-x}$ au voisinage de $x=0$, on peut utiliser la série entière $1+x+x^2+\dots+x^n+\dots$. Il y a une version purement formelle. On écrit $\frac{1}{1-X} = 1+X+X^2+\dots+X^n+\dots$. Plus généralement, on peut étudier une fraction rationnelle P/Q au voisinage de α en utilisant une série de Laurent ; $\frac{P(x)}{Q(x)} = \sum\limits_{n \geqslant n_0} a_n(x-\alpha)^n$ (n_0 étant l'ordre du pôle en $x=\alpha$). Il y a une version purement formelle, on considère la série formelle $\sum\limits_{n \geqslant n_0} a_n(X-\alpha)^n$ et l'on écrit $\frac{P(X)}{Q(X)} = \sum\limits_{n \geqslant n_0} a_n(X-\alpha)^n$.
On est ainsi conduit à introduire, pour tout $\alpha \in \mathbb{C}$ le complété de $\mathbb{C}(X)$ pour la valuation v_α. C'est le corps des séries de Laurent formelles en $X-\alpha$; le corps $\mathbb{C}(X)$ s'interprète comme sous corps de $\mathbb{C}[[X-\alpha]][(X-\alpha)^{-1}] = \mathbb{C}((X-\alpha))$.
De même pour étudier un nombre rationnel « au voisinage d'un nombre premier p », on a introduit le complété de $\mathbb{Q}$ pour la valuation v_p, c'est-à-dire $\mathbb{Q}_p$; $\mathbb{Q}$ s'interprète comme sous-corps de $\mathbb{Q}_p$.
Pour étudier une fraction rationnelle au voisinage de l'infini, on la développe en série de $1/X$. Algébriquement, on complète $\mathbb{C}(X)$ pour la valuation à l'infini : $\mathbb{C}[[1/X]]$.
De même, pour étudier un nombre rationnel « au voisinage du nombre premier à l'infini », on se place dans $\mathbb{R}$.
L'analogue de l'anneau $\mathbb{A}_p = \mathbb{R} \times \prod\limits_{p \in \mathcal{P}} \mathbb{Z}_p$ (*cf.* la remarque de la page 477) est l'anneau :
$$\mathbb{C}((1/X)) \times \prod_{a \in \mathbb{C}} \mathbb{C}[[X-a]].$$

Exemples.

1. Soit $F(X) = \frac{1}{1-2x} \in \mathbb{C}(X)$. Au voisinage de $X=0$, on a une écriture en série formelle dans $\mathbb{C}((X))$: $F(X) = 1+2X+4X^2+\cdots$

Au voisinage de l'infini, on écrit $F(X) = -\frac{X^{-1}}{2}\,\frac{1}{1-X^{-1}/2}$. On en déduit une écriture en série formelle dans $\mathbb{C}((X^{-1}))$: $F(X) = -\dfrac{1}{2}X^{-1} - \dfrac{1}{4}X^{-2} - \dfrac{1}{8}X^{-3} - \cdots$

2. On considère $1/3 \in \mathbb{Q}$.
   On a, dans $\mathbb{Q}_5$, $1/3 = \ldots 444442 \in \mathbb{Z}_5$ et, dans $\mathbb{R}$, $1/3 = 0,33333\ldots$

On peut aussi considérer les solutions d'équations algébriques. Par exemple $X^2 = 1 - X$ admet une solution série formelle en X : $\sqrt{1-X} = 1 - \frac{X}{2} + \frac{X^2}{8} + \cdots$. De façon analogue, nous avons vu que l'on peut trouver dans $\mathbb{Z}_p$ certaines solutions d'équations $X^2 = a$, $a \in \mathbb{Z}$. On peut ainsi utiliser pour l'étude des *nombres algébriques* des méthodes inspirées par l'étude des *fonctions algébriques*.

4.3 Pour aller plus loin : extensions, clôtures algébriques et complétions

Nous donnons, sans démonstrations, quelques indications sur le problème de l'existence d'une clôture algébrique complète de $\mathbb{Q}_p$. Une étude plus approfondie dépasserait le niveau de nos livres[28]
Une *extension finie* d'un corps commutatif K est un sur-corps L de K tel que L soit un K-espace vectoriel de dimension finie. Son degré, noté $[L : K]$, est, par définition, $\dim_K L$. Une extension algébrique de K est une extension dont tous les éléments sont algébriques sur K. C'est une réunion d'extensions finies. Une clôture algébrique d'un corps commutatif K est une extension algébrique L de K qui est algébriquement close, c'est-à-dire que tout polynôme de degré supérieur ou égal à un de $L[X]$ a au moins une racine dans L. Par exemple, $\mathbb{C}$ est une *clôture algébrique* de $\mathbb{R}$. On l'obtient en « ajoutant » i à $\mathbb{R}$. L'extension $\mathbb{C}/\mathbb{R}$ est donc *finie*.
On prolonge classiquement la valeur absolue sur $\mathbb{R}$ en la valeur absolue de $\mathbb{C}$; $\mathbb{C}$ est *complet* pour cette valeur absolue. Si l'on essaie d'imiter cette démarche pour les corps p-adiques $\mathbb{Q}_p$, on rencontre un certain nombre de difficultés. On a d'abord un résultat « positif ». Des preuves de l'énoncé suivant font l'objet de l'exercice III.3.43 (pour l'unicité) et III.3.44 (pour l'existence).

Théorème 94. Soit p un nombre premier. Soit L une extension finie de $\mathbb{Q}_p$, de degré d. Alors il existe une unique norme sur L prolongeant la norme p-adique $|\ |_p$ sur $\mathbb{Q}_p$. On l'obtient de la façon suivante. À $x \in L$ on associe l'application $\mathbb{Q}_p$-linéaire $\ell_x : L \to L$, définie par $\ell : y \in L \mapsto xy \in L$. On a $\det \ell_x \in \mathbb{Q}_p$. Alors l'application $\|\ \| : L \to \mathbb{R}^+$, définie par $\|\ \| : x \in L \mapsto |\det \ell_x|_p^{1/d}$ est une norme sur L prolongeant $|\ |_p$.

Si $\overline{\mathbb{Q}}_p$ est une clôture algébrique de $\mathbb{Q}_p$, alors on peut montrer que l'extension $\overline{\mathbb{Q}}_p/\mathbb{Q}_p$ est *infinie*. On peut facilement prolonger $|\ |_p$ en une norme sur le corps $\overline{\mathbb{Q}}_p$, de façon unique, en utilisant le théorème ci-dessus, *mais* on peut montrer que $\overline{\mathbb{Q}}_p$ n'est pas complet pour cette norme. On peut le compléter en utilisant la proposition 91. Heureusement le processus s'arrête : le corps complet obtenu, noté C_p, est algébriquement clos. Ce corps « ressemble[29] à $\mathbb{C}$ ». Il lui est abstraitement isomorphe (pour des raisons de cardinal, en employant l'axiome du choix), mais on peut montrer que, en un certain sens, il n'existe pas de « bon isomorphisme ».

[28] Pour une introduction à l'étude des extensions de corps, le lecteur pourra consulter le chapitre VII du livre *Algebra* de S. Lang, Addison Wesley Publishing Company 1967.

[29] Un meilleur analogue de $\mathbb{C}$ est *l'anneau des nombres complexes p-adiques* construit par J.M. Fontaine, en 1982. Il n'y a pas de bon analogue de $2i\pi$ dans C_p et la construction de Fontaine y remédie.

5 Nombres, observabilité et calculabilité

Nous insistons dans cette partie sur l'aspect historique. Nous nous sommes efforcés de donner quelques *repères précis* (textes, dates, références), qui devraient être utiles aux étudiants suivant des cours d'histoire des mathématiques et à ceux qui se destinent à l'enseignement.

5.1 Qu'est-ce qu'un nombre ? Approximations rationnelles et approximations décimales

5.1.1 Qu'est-ce qu'un nombre ?

La notion de nombre entier apparaît dans toutes les civilisations. Les élaborations philosophiques de ce concept sont variées et ont évolué depuis la nuit des temps jusqu'à la théorie des ensembles. Pour l'antiquité grecque, cette élaboration est due aux pythagoriciens. On la retrouve chez Euclide :*« le nombre est une multiplicité faite d'unités »*. Nous n'approfondirons pas cette question et nous partirons de l'idée que la notion de nombre entier est « sans problème » (Kronnecker disait que les entiers étaient « créés par Dieu »). Notons que le *calcul* avec les nombres entiers ne l'est pas. Les égyptiens, les grecs, les romains ne savaient pas bien calculer avec les entiers. Un calcul efficace n'a été permis que par l'invention du zéro et la numération de position que nous avons intensivement utilisée dans ce module. La plupart des civilisations ont utilisé la base $b = 10$, mais il y a des exceptions : les calculateurs babyloniens utilisaient la base $b = 60$.

Partant des entiers, l'étape suivante est historiquement moins évidente. Pour l'antiquité grecque (et dans bien d'autres civilisations) il y a une notion *opératoire* de nombre, liée au « monde concret ». Elle est basée sur l'opération de *mesure des grandeurs* (longueurs, aires, volumes). Par exemple, on essaie de « calculer » une longueur en fonction d'un étalon (ou unité) de longueur. Les entiers ne suffisent pas et l'on essaie de changer d'étalon en divisant l'unité initiale en parties égales. Ainsi émerge le concept de ce que nous appelons les nombres rationnels. Il faut noter que ces nombres n'existent pas chez les grecs, il ne connaissent que les *rapports* (de longueurs, aires, volumes). Si la longueur à mesurer est commensurable à l'étalon tout va bien, mais il y a le problème de la diagonale du carré, qui n'est pas commensurable au côté. On suppose que ce problème (ou plus vraisemblablement une variante en relation avec le nombre d'or) fût un traumatisme pour l'école pythagoricienne.

On retrouve bien plus tard la conception grecque chez Isaac Newton (*Arithmétique universelle*[30], 1707) : *« On entend par nombre, moins une collection de plusieurs unités*[31]*, qu'un rapport abstrait d'une quantité quelconque à une autre de même espèce, qu'on regarde comme l'unité. Le* nombre *est de trois espèces,* l'entier*, le* fractionnaire *et le* sourd[32].

[30] Le texte original *Arithmetica universalis*, publié sans l'accord de Newton, est en latin. Nous citons la traduction française par Noel Beaudeux, chez Bernard, libraire, quai des Augustins, 1802.

[31] C'est une allusion à la définition d'Euclide citée plus haut.

[32] Cette terminologie, un peu étrange, pour les nombres irrationnels semble dater des mathématiques arabes au IXe siècle. Par exemple, Al-Khwārizmī emploie « gidr asamm », qu'au XIIe siècle Gérard de Cremone traduit en latin par « radix surda » (racine sourde). Après lui « nombres sourds » est employé en occident jusqu'au XVIIIe siècle pour désigner les nombres irrationnels (le terme figure dans *l'Encyclopédie*). Dans le texte original de Newton, on trouve « surdus » : ... integer, fractus & surdus : *Integer* quem unitas meritur, *Fractus*, quem unitatis pars submultiplex metitur, & *Surdus*, cui unitas est icommensurabilis.

*L'*entier *est mesuré par l'unité ; le* fractionnaire *par un sous-multiple de l'unité ; le* sourd *est incommensurable avec l'unité. »*
Ce qu'il faut retenir est que, dans le cadre opératoire de la mesure des grandeurs, certains « nombres » existent naturellement, par exemple ceux que nous appelons $\sqrt{2}$, $\sqrt[3]{3}$, π, ou les rapports des longueurs des côtés des polygones réguliers au rayon du cercle circonscrit... Le problème n'est pas leur existence mais leur « calcul ». Mais que veut dire « calculer » ? Dans l'antiquité grecque, il y a une réponse claire chez Archimède. Traduite dans nos concepts, il s'agit de trouver des « valeurs approchées rationnelles », puis de préciser ceci avec des encadrements rationnels avec des écarts aussi petits que possible. On a ainsi une notion *d'approximation rationnelle*. Dans les civilisations disposant du zéro et de la numération de position (Inde, Mésopotamie) apparaît aussi la notion *d'approximation décimale* (ou sa variante en base b). C'est une idée reçue que les deux notions sont équivalentes, mais nous verrons que ce n'est pas si simple.

Exemple. Voici un exemple d'approximation d'un nombre irrationnel en numération en base $b = 60$. Sur une tablette d'argile babylonienne (YBC 7289) conservée à l'université de Yale, est dessiné un carré avec ses diagonales et on y trouve, écrit en caractères cunéiformes, le nombre N :

$$\langle\langle\langle\langle\top\top\ \langle\langle\top\top\top\top\top\ \langle\langle\langle\top\top\top\top\top.$$

et un autre nombre N'. Les babyloniens calculaient en base 60. Il utilisaient donc 59 chiffres (le zéro étant représenté par une place vide). On peut penser que le nombre N est écrit avec trois chiffres (écrits eux de façon décimale). On peut le retranscrire en écrivant les chiffres dans notre système décimal par (42), (25), (35). Son double est N', l'autre nombre écrit sur la tablette est retranscrit en $N' = (1)(24)(51)(10)$. On constate que N' est une approximation de $\sqrt{2}$, écrite en base 60, en plaçant la « virgule » après le (1) (*cf.* l'exercice III.3.4). Ainsi N correspond à l'approximation $\frac{42}{60} + \frac{25}{60^2} + \frac{35}{60^3}$ de $\sqrt{2}/2 = 1/\sqrt{2}$: $N = 0,707106\ldots$.
Il est vraisemblable que, sur le dessin, il s'agit d'un carré de côté $0,(30)$ ($1/2$ pour nous) dont la diagonale mesure environ N ($1/\sqrt{2}$ pour nous). La tablette donne donc un calcul approximatif de la diagonale d'un carré de côté $1/2$.
Ainsi les babyloniens utilisaient la numération de position avec « des chiffres après la virgule » et étaient capables de calculer des approximations de $1/\sqrt{2}$ et $\sqrt{2}$ avec une grande précision (6 et 5 décimales exactes dans notre système décimal, *cf.* l'exercice III.3.4). Ils employaient vraisemblablement la première étape de la méthode de Héron, avant Héron ....

Depuis l'antiquité la physique s'est considérablement développée. Il est tentant aujourd'hui d'étendre le concept de « nombre opératoire » des anciens en remplaçant l'observation du monde géométrique par celle du monde physique. Nous y reviendrons au paragraphe suivant.

L'étape suivante est une *nouvelle conception opératoire du nombre*. Au lieu d'observer le monde qui nous entoure, on va observer « le monde des calculs des mathématiciens ». Depuis le moyen age (et même depuis Archimède ...) les mathématiciens considèrent certains objets que nous appelons « suites » et « séries ». Il faut ici prendre garde à ne pas confondre des concepts qui ont *radicalement changé* au cours de l'histoire et en particulier, savoir

« oublier la théorie des ensembles ». Pour nos lointains prédécesseurs, suite ou série veut essentiellement dire ce que nous appelons une suite ou une série dont le terme général est « un rationnel engendré par une loi explicite » (par exemple une récurrence, plus généralement un programme). C'est donc « un objet calculable ». Même s'ils ne formalisent pas les choses comme nous (avec des $\forall$, $\exists$, ε, α), les mathématiciens du XVIIIe siècle savent parfaitement ce qu'est une suite (ou une série) convergente. Ils considèrent ainsi la limite d'une suite (ou la somme d'une série), *les notions de suite (ou de série) étant prises en leur sens*, comme *un nombre* bien défini. Ceci étant, ils veulent en général « le calculer ». Ceci veut dire qu'ils veulent l'exprimer en fonction de « nombres déjà connus » (par exemple $\sum_{n\in\mathbb{N}^*} 1/n^2 = \pi^2/6$; Euler 1741) ou, à défaut, en donner de « bonnes approximations rationnelles », ou en calculer un certain nombre de décimales.

5.1.2 Des rapports au nombre abstrait

L'idée d'une définition abstraite des nombres (rationnels et certains irrationnels), *indépendamment de la théorie des proportions*, apparaît chez des mathématiciens du monde arabo-musulman au moyen-age. Elle est clairement formulée, bien plus tard, au milieu du XVIIIe siécle par le mathématicien allemand Abraham Gotthelf Kästner[33] C'est une rupture avec la vision grecque que l'on trouve encore chez Newton et dans l'Encyclopédie de Diderot et d'Alembert. Kästner écrit (en VF) : *"On peut considérer le nombre irrationnel comme étant composé de deux parties : l'une, le commencement, est rationnelle et peut être prolongée à volonté de telle sorte que l'autre partie, la fin, qui reste en toute rigueur toujours inconnue, devienne plus petite que toute grandeur donnée... »*. Citons aussi[34] : *« ... on peut enfermer un nombre irrationnel entre deux rationnels qui peuvent se rapprocher l'un de l'autre d'aussi près qu'on le désire »*.

On voit ainsi apparaître l'idée d'une définition *abstraite* des réels par approximation rationnelle ou par coupures. Kästner a aussi l'idée de définir les opérations sur ses nombres en prolongeant celles sur les rationnels. Il y a eu ensuite diverses tentatives de définition des réels (ou tout au moins d'éclaircissement de la notion de nombre irrationnel), en relation avec des réflexions sur la rigueur en analyse. On peut citer Cauchy (1821) : *« ... un nombre irrationnel est la limite des diverses fractions qui en fournissent des valeurs de plus en plus approchées »*, le mathématicien pragois Bernhard Bolzano (vers 1830) et Weierstraß (dont la définition des réels n'a jamais été publiée). Les constructions vraiment rigoureuses de $\mathbb{R}$ sont dues à Dedekind, Meray et Cantor (puis Heine). Pour leur présentation historique, nous renvoyons le lecteur à l'introduction de la partie IV.1.3 du livre [L1]. En France, l'idée de ... *dégager l'idée de nombre irrationnel de son origine géométrique.* a été popularisée par Jules Tannery (*Introduction à la théorie des fonctions d'une variable*, 1886). Tannery semble avoir élaboré une théorie des coupures avant de connaître les travaux de Dedekind.

5.1.3 Approximations rationnelles et calculs de décimales

Nous allons préciser les notions d'approximations rationnelles d'une part et de calcul des décimales d'autre part, et les comparer. La comparaison est plus délicate (et plus intéres-

[33] *Anfangsgründe der Arithmetik, Geometrie, ebenen und sphärischen Trigonometrie und Perspectiv*, 1758.

[34] Dans *l'Encyclopédie*, d'Alembert n'acceptait comme nombres que les rationnels, il écrivait que les irrationnels ne sont pas des nombres, mais qu'ils sont approchables par ceux-ci *« autant qu'on veut »*.

sante !) que ce qu'il est généralement admis. Elle nous permettra de prendre conscience d'un problème de « calculabilité », sur lequel nous reviendrons au paragraphe suivant en relation avec les travaux d'Alan Turing.

1. *Approximations rationnelles.* On considère deux suites (a_n) et (b_n) de nombres rationnels calculées par un algorithme (un programme), telles que, pour tout $n \in \mathbb{N}$, $a_n \leqslant a_{n+1} \leqslant b_{n+1} \leqslant b_n$ et $b_n - a_n \leqslant 2^{-n}$. On dit que ces suites sont *calculables*. D'après les propriétés de $\mathbb{R}$ (théorème des segments emboités), il existe un unique réel x tel que, pour tout $n \in \mathbb{N}$, $a_n \leqslant x \leqslant b_n$. Quitte à extraire des suites, on peut remplacer 2^{-n} par 10^{-n}

   En pratique, on a la situation inverse, on « connaît x » (par exemple $x = \sqrt{3}, \mathrm{e}, \pi$, $\ln 2, \ldots$) et l'on essaie d'exploiter cette connaissance pour chercher un programme calculant les deux suites.

2. *Calcul des décimales.* On considère une suite de chiffres engendrée par un algorithme définissant une écriture décimale propre (ce dernier point n'est pas nécessairement facile à vérifier ...). Elle définit un réel x.

   En pratique, on a la situation inverse, on « connaît x » et l'on essaie d'exploiter cette connaissance pour chercher un programme calculant son écriture décimale (propre).

Dans les deux cas, une fois défini x, on le « calcule ». Dans un très grand nombre de textes contemporains les auteurs considèrent implicitement que les deux notions de calcul sont équivalentes. C'est vrai *théoriquement* mais *c'est faux du point de vue du calcl effectif*[35]. Bien sûr, la connaissance des décimales donne des encadrements explicites. C'est le *calcul effectif* des décimales à partir d'encadrements qui peut poser un problème, parfois « insoluble ».

Exemple. Un « très bon » encadrement d'un réel par des rationnels : $a \leqslant x \leqslant b$ ne donne pas nécessairement beaucoup de décimales exactes de x. Il ne donne même pas nécessairement la connaissance exacte de la partie entière de x.
Soient $x \in \mathbb{R}$, $a = 0,99999999$ et $b = 1$. On a $b - a = 10^{-8}$, mais l'encadrement $a \leqslant x \leqslant b$ donne pour la troncature x_8 *deux* possibilités :

- $x_8 = 0,99999999$,
- $x_8 = 1,00000000$ (qui équivaut à $x = 1$).

Ainsi la partie entière de x est 0 *ou* 1. On ne sait *rien* de plus.

Alan Turing dit[36] que la proposition suivante est fausse : *Il existe une procédure générale qui permet de fabriquer un programme de calcul de la suite des décimales à partir des programmes qui calculent les suites (a_n) et (b_n)*. Toujours selon Turing, les approximations définissent *théoriquement* un réel, mais ne permettent pas nécessairement de calculer la suite de ses décimales bien que l'on puisse prouver *théoriquement* que cette suite est calculable[37].

[35] Même Alan Turing a apparemment fait l'erreur dans sa première version de son travail sur les nombres calculables. Il est revenu sur le probléme six mois plus tard. Nous en reparlerons dans le paragraphe suivant.

[36] Dans *On Computable Numbers, with an Application to the Entscheidungsproblem. A Correction*. Proc. London Math. Soc., 2e série, 43, (1938), p. 544–546.

[37] Turing explique que ceci est lié à l'usage du principe du tiers exclu dans la preuve théorique, *cf.* la note 39.

Pour comprendre le problème, nous allons revenir à l'exemple des approximations de $\sqrt{2}$ par les suites de Héron, que nous conseillons au lecteur de (re)lire. Nous avions constaté sur cet exemple une possible difficulté pour déduire le calcul des décimales de $\sqrt{2}$ de la connaissance d'encadrements. C'est le coeur du problème.

Exemple : calcul des décimales de $\sqrt{2}$ par la méthode de Héron. Nous renvoyons le lecteur à [L1, IV.1, 2.3.2]. On montre que $\sqrt{2}$ est la limite d'une suite récurrente strictement décroissante de rationnels (c'est un cas particulier de la méthode de Newton). En passant à une suite extraite, on trouve une suite de rationnels (y_k) telle que $0 < y_k - \sqrt{2} < 10^{-k}$. D'où $y_k - 10^{-k} < \sqrt{2} < y_k$. On a ainsi des encadrement par des rationnels aussi bons que l'on veut du type 1. On peut écrire un programme engendrant les y_k. La question est alors : peut-on utiliser ces encadrements pour calculer exactement autant de décimales de $\sqrt{2}$ que l'on veut ? La réponse est positive, mais ce n'est pas évident ! Nous allons donner l'idée de la difficulté et du moyen de la surmonter. Les détails font l'objet de l'exercice III.3.4.

On peut écrire facilement un programme engendrant exactement les écritures décimales des y_k, qui sont périodiques de période calculable (*cf.* [L1, IV.1, 3.1.2]).

Notons : $\sqrt{2} = 1,\alpha_1\alpha_2\ldots\alpha_{k-1}\alpha_k\ldots$ l'écriture décimale cherchée. On obtient , à la $k^{ème}$ étape, un encadrement :

$$(1,\beta_{k,1}\beta_{k,2}\ldots\beta_{k,k-1}\beta_{k,k}\ldots) - 10^{-k} \leqslant 1,\alpha_1\alpha_2\ldots\alpha_{k-1}\alpha_k\ldots \leqslant 1,\beta_{k,1}\beta_{k,2}\ldots\beta_{k,k-1}\beta_{k,k}\ldots,$$

Si la k-ieme décimale $\beta_{k,k}$ de y_k *n'est pas nulle*, alors :

$$1,\beta_{k,1}\ldots\beta_{k,k-1}(\beta_{k,k}-1)\ldots \leqslant 1,\alpha_1\ldots\alpha_{k-1}\alpha_k\ldots \leqslant 1,\beta_{k,1}\ldots\beta_{k,k-1}\beta_{k,k}\ldots,$$

et l'on peut calculer $(k-1)$ décimales exactes de $\sqrt{2}$: $\alpha_1 = \beta_{k,1},\ldots,\alpha_{k-1} = \beta_{k,k-1}$. C'est ce que nous avons fait dans le livre [L1] pour quelques valeurs de k. Mais, *si* $\beta_{k,k}$*=0*, alors l'encadrement *ne donne qu'une information incomplète sur les décimales de* $\sqrt{2}$ (*cf.* l'exercice III.3.2). Le remède semble simple : on peut remplacer k par $k+1$ et tester la $(k+1)^{ème}$ décimale $\beta_{k+1,k+1}$ de y_{k+1}. Si elle n'est pas nulle, on obtient k décimales de $\sqrt{2}$, donc a fortiori $(k-1)$. Si c'est encore 0, on continue. Le problème est que le processus pourrait ne pas s'arrêter. On devine qu'alors $\sqrt{2}$ serait un décimal (c'est vrai, mais la preuve n'est pas facile...) et l'on sait que ce n'est pas le cas ($\sqrt{2}$ est irrationnel). On peut donc calculer les décimales de $\sqrt{2}$ *à un ordre arbitraire*. Nous pouvons ainsi surmonter la difficulté signalée par Turing grâce au fait que *nous savons que* $\sqrt{2}$ *est irrationnel*.

La méthode que nous venons d'esquisser pour $\sqrt{2}$ se généralise. Si l'on sait que le réel x n'est pas un décimal (en particulier s'il est *irrationnel*), alors la proposition fausse selon Turing devient vraie : on peut transformer les programmes engendrant les suites (a_n) et (b_n) de 1, page 512, en un programme calculant les décimales de x. Mais, *si l'on ne sait rien sur* x, ce nouveau programme pourrait tourner indéfiniment et l'on ne peut pas savoir a priori si le calcul s'arrête ou pas. Turing donne l'exemple de la constante d'Euler-Mascheroni γ. On ne sait pas si γ est ou non un décimal. On peut prouver qu'il existe un programme qui calcule ses décimales *à un ordre arbitraire*, mais on ne sait pas en écrire un[38].

[38] Ceci n'empêche pas d'écrire des programmes et d'obtenir beaucoup de décimales exactes. Il existe par exemple

On peut énoncer le résultat suivant.

Proposition 95. Soient $x \in \mathbb{R}$, $x > 0$ et $(b_n)_{n\in\mathbb{N}}$ une suite de rationnels. On suppose que, pour tout $n \in \mathbb{N}$, on a $b_n - 10^{-n} \leqslant x \leqslant b_n$ et que la suite (b_n) *est calculée par un programme* P.

(i) Si x n'est pas un décimal, on peut *écrire, à partir de* P, un programme P' qui calcule la partie entière et la suite des décimales de x.

(ii) Si l'on ne sait rien sur x, on peut montrer *qu'il existe* un programme P'' qui calcule la partie entière et la suite des décimales de a, mais on ne sait pas en écrire un à partir de P.

Démonstration.

(i) On construit par récurrence une suite *décroissante* $(\tilde{b}_n)_{n\in\mathbb{N}}$ telle que, pour tout $n \in \mathbb{N}$, on a $\tilde{b}_n - 10^{-h} \leqslant x \leqslant \tilde{b}_n$.

On pose $\tilde{b}_0 = b_0$, puis, pour $n \in \mathbb{N}$, $\tilde{b}_{n+1} := \min(\tilde{b}_n, b_{n+1})$. On vérifie que la suite $(\tilde{b}_n)$ convient. On peut, à partir de P, écrire facilement un programme $\widetilde{P}$, qui calcule la suite de rationnels $(\tilde{b}_n)$. On est alors ramené aux hypothèses de l'exercice III.3.4 de la page 521 qui propose une preuve du résultat.

(ii) On a une alternative, mais on ne sait pas dans quel cas on est.

* Ou bien x n'est pas un décimal et l'on peut applique le résultat de (i).
* Ou bien x est un décimal et la preuve du résultat est évidente.

On prouve ainsi *l'existence*[39] d'un programme P'' qui calcule les décimales de x, *mais* on ne sait pas écrire, à partir de P, un tel programme. ■

5.2 Les réels sont-ils réels ? Alan Turing et la calculabilité

Nous terminons ce module consacré aux nombres réels par quelques remarques historiques sur les fondements des mathématiques et l'émergence de l'informatique, prolongeant l'introduction du module I.1 du livre [L1]. Nous parlerons aussi des relations entre les nombres réels des mathématiciens, et le « monde réel », dans la ligne des considérations de la partie 5.1.1. Dedekind s'était déjà intéressé à cette question en créant sa théorie des coupures à partir de l'intuition d'une droite *continue* : *« Accepter cette proposition de la ligne n'est rien de plus qu'un axiome, . . . , par lequel nous pensons la ligne comme continue. Si l'espace a quelque existence réelle, il n'est pas nécessairement continu . . . ".*

En 1998 le mathématicien français Roger Godement a écrit [40] : *Les nombres prétendument « réels » — il est trop tard pour changer la terminologie — ne se rencontrent pas vraiment dans la réalité physique ; ils sont nés dans le cerveau des mathématiciens.*

un programme calculant plus de 400 milliards de décimales de la constante d'Euler-Mascheroni γ, même si l'on ne sait pas s'il pourrait les calculer toutes.

[39] En utilisant le principe du tiers exclu.

[40] Analyse mathématique I, Springer

Cette affirmation est à rapprocher des idées de L. Kronecker, *cf.* la note 4 de la page 430.
Nous allons voir que l'on peut être moins « pessimiste » que R. Godement. *Certains* nombres réels méritent mieux leur nom et, parmi eux, il y en a dont on peut dire raisonnablement qu'ils « se rencontrent (ou peuvent se rencontrer) dans la réalité physique ». On peut parvenir à cette conclusion en suivant une approche mathématique et logique comme l'a fait le mathématicien anglais Alan Mathieson Turing au milieu des années trente. On peut aussi tenter de suivre une autre voie en étudiant les relations entre les mathématiques et la physique jusque dans ses développements récents.
Dans ce paragraphe, l'interprétation des réels par les écritures décimales est importante.
Au début du XXe siècle[41] David Hilbert s'était proposé, dans le prolongement des travaux de Bertrand Russell, Giuseppe Peano et Gottlog Frege, de donner des bases solides aux mathématiques (plus précisément à l'arithmétique). Son défi était de donner un système d'axiomes *consistant* et *complet* suffisamment riche pour permettre l'expression et la preuve de toutes les propositions arithmétiques fondamentales. En 1928, Hilbert est revenu sur la question en formulant un deuxième défi : le *problème de la décision* (*Entscheidungsproblem*).
Avant de poursuivre, nous allons décrire sommairement les notions de *consistance*, *complétude* et *décidabilité*.

- Un système d'axiomes est dit *consistant* s'il n'est pas possible de prouver à la fois une proposition P *et* son contraire $\neg P$ (c'est-à-dire « non » P, *cf.* [L1, I.1, 7.1, page 47]).
- Un système d'axiomes est dit *complet* si pour toute proposition P exprimable dans le système, il est possible de prouver P *ou* $\neg P$.
- Une proposition P est dite *décidable* dans un système d'axiomes si on peut la démontrer ou démontrer sa négation dans le cadre de ce système. Sinon on dit qu'elle est *indécidable* (logiquement) ou, de préférence, qu'elle *indépendante* du système d'axiomes.
- Un problème est *décidable* (algorithmiquement) s'il existe une « procédure de décision effective » (un algorithme qui se termine en un nombre fini d'étapes) permettant de répondre par oui ou par non à la question posée par le problème. Sinon il est indécidable (algorithmiquement).

Il ne faut pas confondre l'indécidabilité logique et l'indécidabilité algorithmique.

Exemples.

1. *Le cinquième axiome d'Euclide.* Le système axiomatique d'Euclide pour la géométrieplane est formé de 5 axiomes ou postulats [42]. Le cinquième postulat (P_5) est celui dit des parallèles : *étant donnés un point et une droite ne passant pas par ce point, il existe une et une seule droite passant par ce point et parallèle à la première*. Durant plusieurs siècles les mathématiciens ont essayé de déduire ce postulat des 4 premiers. Ce n'est qu'au XIXe siècle que C. F. Gauss, Janos Bolyai, Nicolai Lobachevsky, Henri Poincaré... ont découvert des modèles de géométries, dites non-euclidiennes, dans lesquels (P_5) est faux. Ainsi le système axiomatique défini par les 4 premiers postulats *n'est pas complet*. Le cinquième postulat d'Euclide est *indépendant* du système des 4 premiers.

[41] En 1900 puis dans les années 20.
[42] Euclide, *Elements*, Livre I, vers 300 avant J.-C.

2. *L'hypothèse du continu.* Cantor avait conjecturé que tout ensemble contenu dans $\mathbb{R}$ et contenant $\mathbb{N}$ est dénombrable ou a le même cardinal que $\mathbb{R}$. C'est *l'hypothèse du continu : HC*. Kurt Gödel a montré en 1938 que l'on ne pouvait pas réfuter HC avec les axiomes standard de la théorie des ensembles et Paul Cohen a montré en 1963 que l'on ne pouvait pas non plus démontrer HC avec ceux-ci. Ainsi, HC est *indépendante* des axiomes ZFC[43] (Zermelo-Fraenkel+choix).
3. *Le dixième problème de Hilbert.* La question de Hilbert sur le problème de la décision est liée à un problème qu'il avait posé en 1900 à Paris (au congrès international des mathématiciens) sur les *équations diophantiennes* (le 10ème problème).

   Une équation (portant sur plusieurs variables $x_1, \ldots, x_n$, les inconnues) est dite diophantienne si elle est de la forme $D(x_1, \ldots, x_n) = 0$, où D est un polynôme à plusieurs variables à coefficients dans $\mathbb{Z}$. Hilbert demandait de trouver une méthode *générale* permettant de dire, au moyen d'un nombre fini d'opérations, si une équation diophantienne admet ou non une solution *en nombres entiers*. Youri Matiiassevitch a montré en 1970 qu'une telle méthode n'existe pas. C'est un résultat d'indécidabilté algorithmique.
4. *Le problème de l'arrêt*. Le problème de l'arrêt est un problème de *décision* : déterminer à partir de la description d'un algorithme (programme) s'il s'arrête ou non. Alan Turing a montré en 1936 que c'est impossible.

Le premier espoir de Hilbert a été ruiné[44] par le mathématicien allemand Kurt Gödel en 1931. Gödel a montré deux théorèmes d'incomplétude. Le premier dit essentiellement que dans tout système d'axiomes assez puissant pour formaliser l'arithmétique il existe une proposition P telle que ni P ni $-P$ ne soit prouvable à partir des axiomes. Le second, qui résulte du premier, dit qu'aucun système d'axiomes *consistant* assez puissant pour formaliser l'arithmétique ne peut permettre de prouver sa propre consistance. Le point de départ de la preuve de Gödel est le paradoxe d'Epiménide le Crétois, qui affirmait « tous les Crétois sont menteurs ». Gödel le reformule en « la présente assertion est indémontrable », qu'il remplace ensuite par une assertion sur les nombres entiers positifs.
Il restait le problème de la décision. En 1936 le mathématicien anglais A.M. Turing ruina le deuxième espoir de Hilbert en donnant une réponse négative à ce problème[45] dans son article *Sur les nombres calculables*[46] . C'est ici que nous retrouvons les nombres réels et en particulier leur expression décimale. En effet, pour élaborer le *concept de calculabilité*, Turing se ramène au concept de *nombres* « calculables ». Son article commence par (en VF) : *On peut définir sommairement les nombres « calculables » comme étant les réels dont l'expression décimale est calculable avec des moyens finis*[47]. Aujourd'hui on dirait qu'il existe

[43] P. Cohen a montré que l'axiome du choix n'était pas une conséquence de ZF.

[44] Un diable est caché dans le paradis Cantorien. André Weil : *« Dieu existe parce que les mathématiques sont cohérentes, et le diable existe parce-que nous ne pouvons pas le prouver »*.

[45] Alonzo Church avait obtenu un résultat équivalent, mais formulé différemment, quelques mois auparavant.

[46] *On Computable Numbers, with an Application to the Entscheidungsproblem*, 1937.

[47] Six mois après son article Turing a constaté qu'il y avait une difficulté avec cette définition. Elle vient de la différence « subtile » entre l'approximation par deux suites de *rationnels* d'une part et le calcul des décimales d'autre part, que nous avons étudiée dans le paragraphe précédent.

un *programme* qui affiche la suite des chiffres de l'expression décimale. Mais, quand Turing écrit son article, il n'existe ni programme ni ordinateur. Cet article est en fait *un pas essentiel* vers l'invention des ordinateurs et le développement de l'informatique théorique.
Pour formaliser les notions de calculabilité avec des moyens finis (en termes d'algorithme...), Turing introduit un concept de *machine* (en particulier de *machine universelle*). On dit aujourd'hui *machine de Turing*. Turing écrit : *Selon ma définition, un nombre est calculable si sa représentation décimale peut être écrite par une machine*. Turing montre que les nombres réels calculables forment un *sous-corps* de $\mathbb{R}$, contenant évidemment $\mathbb{Q}$, mais aussi le corps $\overline{\mathbb{Q}}$ des nombres algébriques. Un point important est que « la plupart des nombres réels » ne sont pas calculables : *Bien que l'ensemble des nombres calculables soit très grand, et comparable à bien des égards à celui des réels, il est pourtant dénombrable*. Alors que les nombres réels non-calculables sont « les plus nombreux », il est bien difficile d'en *définir*[48] un !
Au XVIIIe siècle un certain nombre de mathématiciens (principalement Leibniz et Newton) ont inventé le calcul différentiel et intégral. Cette invention était largement motivée par les sciences expérimentales et l'étude du mouvement des astres, c'est à dire de la réalité physique. Dès cette époque apparaissent en calcul différentiel et intégral des nombres réels irrationnels « définis explicitement », comme $\ln 2$, π, $\sqrt{\pi}$, e, $\zeta(3)$,... en utilisant certaines séries ou intégrales. On peut, pour simplifier, parler des nombres réels irrationnels figurant dans les œuvres complètes de Leonhard Euler $(1707-1783)$. En suivant Turing, on peut montrer que ces nombres sont calculables. Récemment la définition de ces nombres a été formalisée avec la théorie des *périodes* (et plus généralement des *périodes exponentielles*), un domaine de recherche très actif aux confins de l'arithmétique et de la physique théorique[49]. L'article de Kontsevitch et Zagier (cité dans le *Un peu d'histoire* de la page 177) se termine par : *"Then all classical constants are periods in an appropriate sense"*. On peut raisonnablement penser que, parmi « les réels pouvant se rencontrer dans la réalité physique », il y a les périodes (en un sens approprié...).
Pour conclure par une interrogation, il est traditionnellement admis, comme « évident », que les « constantes fondamentales » (sans dimension) des physiciens, comme la constante de structure fine α (qui régit la force électromagnétique en mécanique quantique relativiste), sont représentables par des nombres réels. Ce n'est pas si sûr, car certaines d'entres elles varient ou pourraient varier. Ainsi, l'ensemble $\mathbb{R}$, « né dans le cerveau des mathématiciens », peut, quand on le voit du point de vue du « monde réel », paraître « bien trop gros », mais il est aussi fort possible qu'il soit « trop petit ».
Nous conseillons au lecteur intéressé par les thèmes évoqués dans ce paragraphe la lecture du très intéressant article de Gregory Chaitin[50] : *How real are real numbers ?* Il contient des prolongements vers les théories des probabilités et de l'information [51] et des considérations sur la possible nature digitale de l'univers physique.

[48] C'est l'objet de travaux en théorie algorithmique de l'information, en particulier par Gregory Chaitin. Le nombre Ω de Chaitin, la probabilité d'arrêt, est défini mais non calculable.
[49] Le lecteur pourra consulter *Un peu d'histoire* du module intégration, page 177 ? ?.
[50] arXiv :math/0411418v3, 2004
[51] Le nombre Ω de Chaitin, la probabilité d'arrêt, est défini à partir du problème de l'arrêt et de la notion de complexité d'un programme. Ce nombre est dans la ligne d'idées d'Émile Borel en particulier sur la notion de

Exercice type corrigé

Le but de l'exercice est de construire des homéomorphismes entre les espaces métriques $(\mathbb{Z}_p, d_p)$ et certains compacts de $\mathbb{R}$ (des exemples de fractales) et de mieux faire comprendre les différences entre les espaces topologiques $\mathbb{Z}_b$ et $[0,1]$. On verra que les écritures p-adiques à droite *impropres* sont « au coeur du problème ».

III.3.0 On note sous forme de série de Hensel $\sum_{n\in\mathbb{N}} \alpha_n p^n$ les éléments de l'anneau des entiers p-adiques $\mathbb{Z}_p$.

1. a) On définit une application $\varphi : \mathbb{Z}_2 \to \mathbb{R}$ par $\varphi : \sum_{n\in\mathbb{N}} \alpha_n 2^n \mapsto \sum_{n\in\mathbb{N}} \frac{2\alpha_n}{3^{n+1}}$. Décrire l'image C de φ ? Montrer que φ est continue et induit un homéomorphisme sur son image munie de la topologie induite par celle de $\mathbb{R}$.

   b) On définit une application $\psi : \mathbb{Z}_2 \to \mathbb{R}$ par $\psi : \sum_{n\in\mathbb{N}} \alpha_n 2^n \mapsto \sum_{n\in\mathbb{N}} \frac{\alpha_n}{2^{n+1}}$. Décrire l'image C' de ψ ? Montrer que ψ est continue et surjective mais n'est pas injective. Que peut-on dire des réels dont l'image inverse par ψ n'est pas réduite à un point.

   c) Montrer qu'il existe une unique application $f : C \to C'$ telle que $\psi = f \circ \varphi$ et qu'elle est continue et surjective. La décrire.

2. Dans cette question on choisit un nombre premier p. Soit b un nombre réel, $b \geqslant p$.
   On pose $\kappa_b := \frac{b-1}{p-1}$. On définit une application $\varphi_b : \mathbb{Z}_p \to \mathbb{R}$ par

$$\varphi_b : \sum_{n\in\mathbb{N}} \alpha_n p^n \mapsto \kappa_b \sum_{n\in\mathbb{N}} \frac{\alpha_n}{b^{n+1}}.$$

   a) Montrer que les applications φ et ψ de la question 1 sont des cas particuliers de φ_b.

   b) Montrer que l'application φ_b est continue et que son maximum est 1.

   c) On suppose $b > p$. Montrer que φ_b est injective et définit un homéomorphisme de $\mathbb{Z}_p$ sur son image.

   d) On suppose $b = p$. Montrer que φ_b est une surjection de $\mathbb{Z}_p$ sur le segment $[0,1]$ et qu'elle n'est pas injective.

3. On considère $\mathbb{R}^n$ muni de sa structure euclidienne canonique. Pour les notions de sous-ensemble *cubable* et de sous-ensemble *de mesure nulle* de $\mathbb{R}^n$, le lecteur se reportera au module III.1 *Intégration*. On dit qu'un sous-ensemble compact $C \subset \mathbb{R}^n$ est *autosimilaire* de type : $(\lambda, k) \in (\mathbb{R}_+^*, \mathbb{N}^*)$ s'il existe k *similitudes*, S_i $(i \in [\![1,k]\!])$ de l'espace affine euclidien $\mathbb{R}^n$, toutes de rapport $1/\lambda$, telles que les $S_i(C)$ soient *deux à deux disjoints* et que : $C = \bigcup_{i\in[\![1,k]\!]} S_i(C)$. Si $(\lambda, k) \neq (1,1)$, on dit que C est une fractale. On dit qu'un compact $C \subset \mathbb{R}^n$ est *quasi-autosimilaire* si les $S_i(C)$, au lieu d'être deux à deux disjoints, sont d'intersections deux à deux *de mesure nulle*.

   a) Donner des exemples de compacts cubables *quasi-autosimilaires*. Soit C un tel compact de type (λ, k) $(\lambda \neq 1)$ et de volume non nul. Calculer $\log k / \log \lambda$.

« nombre oracle » qu'il a formulée (ironiquement ...) en 1926. Un nombre oracle serait un nombre réel qui pourrait répondre à toute question possible du type *oui/non*, fabriqué en ordonnant toutes les questions possibles écrites en français.

b) Montrer que, pour tout p premier et tout $b > p$, le compact $C := \varphi_b(\mathbb{Z}_p)$ est autosimilaire. Quel est son type ?

c) Le résultat de la question 3. a) suggère une « notion de dimension » pour un ensemble compact auto-similaire (dite *dimension fractale*). Quelle est-elle pour $C = \varphi_b(\mathbb{Z}_p)$?

Solution.

1. a) L'image de φ est l'ensemble triadique de Cantor : $C = K_3$ (*cf.* [L2, II.1 3.2, proposition 53]). Soient $x, y \in \mathbb{Z}_2$, $x = \sum_{n\in\mathbb{N}} \alpha_n 2^n$ et $y = \sum_{n\in\mathbb{N}} \beta_n 2^n$.

On note m le plus petit entier tel que $\alpha_n \neq \beta_n$. On a $d_2(x, y) = 2^{-m}$.

On suppose $\alpha_m = 1$ et $\beta_m = 0$. Alors :

$$\frac{2}{3^{m+1}} \leqslant \varphi(x) \leqslant \frac{1}{3^m} \quad \text{et} \quad 0 \leqslant \varphi(y) \leqslant \frac{2}{3^{m+2}} \sum_{n\geqslant 0} \frac{1}{3^n} = \frac{1}{3^{m+1}}.$$

On en déduit : $|\varphi(x) - \varphi(y)| = \varphi(x) - \varphi(y)$ et l'encadrement :

$$3^{-m-1} \leqslant |\varphi(x) - \varphi(y)| \leqslant 3^{-m}.$$

Ceci entraîne l'injectivité et la continuité de φ. Puisque $\mathbb{Q}_2$ est *compact*, l'application continue φ induit un *homéomorphisme* de $\mathbb{Q}_2$ sur $\mathbb{K}_3$ (*cf.* [L2, II.1, 3.2, Corollaire 52]).

b) Montrons que l'image de ψ est $C' = [0, 1]$. Tout réel t de $[0, 1[$ admet une écriture dyadique propre $0, \alpha_1\alpha_2 \ldots$ et s'écrit donc $x = \sum_{n\in\mathbb{N}} \frac{\alpha_n}{2^{n+1}}$ ($\alpha_n = 0$ ou 1). Par ailleurs 1 admet une écriture dyadique impropre $0, 11\ldots$ et s'écrit donc $1 = \sum_{n\in\mathbb{N}} \frac{1}{2^{n+1}}$. Par suite $C' = [0, 1]$. On montre la continuité de ψ comme celle de φ dans 1, *mutatis mutandis*. L'application φ n'est pas injective car, pour tout $m \in \mathbb{N}^*$, $1/2^m$ et $\sum_{n>m} \frac{1}{2^{n+1}}$ sont distincts et ont la même image par ψ. Ceci correspond à des écritures dyadiques impropres.

c) On définit $f : K_3 \to [0, 1]$ par $f := \psi \circ \varphi{-1}$, ce qui prouve l'existence (l'unicité est évidente). Cette application,est continue et surjective. Elle identifie les extrémités des intervalles ouverts que l'on a enlevés pour fabriquer l'ensemble de Cantor en les envoyant sur un rationnel de la forme $r/2^h$ ($r, h \in \mathbb{N}$).

2. a) On suppose $p := 2$. Alors $\varphi = \varphi_3$ et $\psi = \varphi_2$.

b) Le maximum correspond à $\alpha_n = p - 1$ pour tout $n \in \mathbb{N}$. Il est égal à :

$$\kappa_b(p-1) \sum_{n\in\mathbb{N}} \frac{1}{b^{n+1}} = \frac{b-1}{p-1}(p-1)\frac{1}{b-1} = 1.$$

Soient $x, y \in \mathbb{Z}_p$. On note m l'ordre de $x - y$. Alors :

$$|\varphi_b(x) - \varphi_b(y)| \leqslant \kappa_b(p-1) \sum_{n\geqslant m} \frac{1}{b^{n+1}} = \frac{b-1}{b^m}\frac{1}{1-b} = b^{-m}.$$

La continuité de φ_b en résulte.

c) Soient $x, y \in \mathbb{Z}_p$. On vérifie facilement qu'une condition *nécessaire* pour que $\varphi_b(x) = \varphi_b(y)$ est que (à permutation près) x et y soient de la forme $x = p^m$, $y = \sum_{n \geqslant m} \frac{p-1}{p^{n+1}}$ ($m \in \mathbb{N}^*$). Calculons alors les images par φ_b. On a $\varphi_b(x) = \kappa_b b^{-m-1}$ et $\varphi_b(y) = \kappa_b(p-1)b^{-m-1} \sum_{n \geqslant m} \frac{1}{b^{n+1}} = b^{-m-1}$. Par suite l'application est injective si, et seulement si, $\kappa_b = 1$, c'est-à-dire si, et seulement si, $b > p$. Si $b > p$, l'application continue φ_b définit un *homéomorphisme* sur son image, car $\mathbb{Z}_p$ est *compact*.

d) On suppose $b = p$. Nous avons vu ci-dessus que φ_b est continue et n'est pas injective. On montre que son image est le segment $[0, 1]$ par le même argument, *mutatis mutandis* que dans le cas $p = 2$ ($\varphi_b = \psi$).

3. a) Soit $m \in \mathbb{N}^*$. Par dichotomie on peut diviser un segment de $\mathbb{R}$ en 2^m segments égaux deux à deux d'intersection réduite à un point, un carré fermé de $\mathbb{R}^2$ en 4^m carrés égaux, d'intersection réduite à un côté ...

Soit C un compact quasi-autosimilaire de type (λ, k). On a, pour tous $i, j \in [\![1, k]\!]$, $\mathrm{vol}\,(S_i(C)) = \mathrm{vol}\,(S_j(C))$ et $\mathrm{vol}\,(S_i(C)) = \lambda^{-n}\,\mathrm{vol}(C)$, donc :

$$\mathrm{vol}(C) = vol\,(S_1(C)) + \cdots + \mathrm{vol}\,(S_k(C)) = k\,\mathrm{vol}\,(S_1(C)) = k\lambda^{-n}\,\mathrm{vol}(C).$$

Puisque $\mathrm{vol}\,C \neq 0$, on obtient $k\lambda^{-n} = 1$ et, en prenant les logarithmes : $\log k = n \log \lambda$, c'est-à-dire $n = \log k / \log \lambda$.

b) On suppose $b > p$. On a $\varphi_b(\alpha_n p^n) = \kappa_b \alpha_n / b^{n+1}$, donc $\varphi_b(p\mathbb{Z}_p) = b^{-1}\varphi_b(\mathbb{Z}_p)$.

On a $\mathbb{Z}_p = (0 + p\mathbb{Z}_p) \cup (1 + p\mathbb{Z}_p) \cup (2 + p\mathbb{Z}_p) \cup \ldots \cup (p - 1 + p\mathbb{Z}_p)$ et cette réunion est *disjointe* (*cf.* (5) de l'exercice 11 de la page 455). C'est une propriété « d'autosimilarité » du compact $\mathbb{Z}_p$ de $\mathbb{Q}_p$. En appliquant φ_b, on obtient :

$$C_b := \varphi_b(\mathbb{Z}_b) = b^{-1}\varphi_b(\mathbb{Z}_b) \cup \left(\kappa_b \frac{1}{b} + b^{-1}\varphi_b(\mathbb{Z}_b)\right) \cup \ldots \cup \left(\kappa_b \frac{b-1}{b} + b^{-1}\varphi_b(\mathbb{Z}_b)\right).$$

Posons, pour $i \in [\![1, p]\!]$, $C_{b,i} := \kappa_b \frac{i-1}{b} + b^{-1}\varphi_b(\mathbb{Z}_b)$. On a $C_b = \bigcup_{i \in [\![1,p]\!]} C_{b,i}$. La réunion est disjointe (φ_b est bijective) et les $C_{b,i}$ sont semblables à C_b, le rapport de similitude étant b^{-1}, donc C_b est *autosimilaire de type* (b, p).

c) On suppose $b > p$. D'après le résultat de la question 3 a), on a $n = \log k / \log \lambda$. Il est donc naturel, compte tenu du résultat de la question 3 b), de *définir* la dimension fractale de $C_b := \varphi_b(\mathbb{Z}_b)$ par $\log p / \log b$. Par exemple la dimension fractale de l'ensemble de Cantor K_3 est $\log 2 / \log 3$.

Quand b tend vers p par valeurs strictement supérieures, la fractale C_b « tend vers le segment $[0, 1]$ » qui n'est plus une fractale et n'est plus homéomorphe à $\mathbb{Z}_p$. La dimension fractale tend vers 1 par valeurs strictement inférieures. Quand b prend la valeur 1, les composantes connexes de $[0, 1] \setminus C_p$ « disparaissent » et leurs extrémités sont « recollées en un point », ce qui fait disparaître l'injectivité (*cf.* la question 1 c)).

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 sous-section 1.1

III.3.1

On admet que toute progression arithmétique de la forme $(1+k10^{s+1})_{k\in\mathbb{N}^*}$ (où $s \in \mathbb{N}$) contient un nombre premier. On définit un nombre réel $x \in]0,1[$ en concaténant, après la virgule, les écritures décimales des nombres premiers écrits dans l'ordre croissant : $x := 0,23571113171923\ldots$
Montrer que x est *irrationnel*.

III.3.2 Soient $a,\ v,\ w \in \mathbb{R}^+$, et $k \in \mathbb{N}^*$.

(i) On suppose que $v - 10^{-k} \leqslant a \leqslant v$ et que l'on connaît la partie entière de v et ses k premières décimales. Trouver une condition sur ces décimales permettant de calculer la partie entière de a et, si $k \geqslant 2$, ses $k-1$ premières décimales. Étudier le cas où cette condition n'est pas réalisée.

(ii) On suppose que $w \leqslant a \leqslant w + 10^{-k}$ et que l'on connaît la partie entière de w et ses k premières décimales. Trouver une condition sur ces décimales permettant de calculer la partie entière de a et, si $k \geqslant 2$, ses $k-1$ premières décimales. Étudier le cas où cette condition n'est pas réalisée.

Les deux exercices suivants sont plus difficiles.

III.3.3 $*$ Cet exercice porte sur le problème de *l'effectivité* de la somme de deux réels donnés sous forme décimale. Soient $x,\ y \in [0,1/2[$. On note $z := x+y$ et, pour tout $k \in \mathbb{N}^*$, $x_k := 0,\alpha_1\alpha_2\ldots\alpha_k$ et $y_k := 0,\beta_1\beta_2\ldots\beta_k$ les troncatures à l'ordre k respectives de x et y.

(i) Soit $m \in \mathbb{N}^*$. On suppose que $\alpha_m + \beta_m \neq 9$. Montrer qu'il existe un réel w, tel que $0 \leqslant w < 10^{-m}$ et $z = x_m + y_m + w$ ou $z = x_m + y_m + 10^{-m} + w$. En déduire que z_{m-1} est la troncature à l'ordre $m-1$ de $x_m + y_m$.

(ii) Soit $m \in \mathbb{N}^*$. On suppose que $\alpha_m + \beta_m \leqslant 8$. Montrer que $z_{m-1} = x_{m-1} + y_{m-1}$.

(iii) Soit $x := 0,0541375\ldots$. Calculer la première décimale de $x+y$ pour $y := 0,0458622\ldots$, puis pour $y := 0,0458625\ldots$.

(iv) Trouver des nombres réels x et y tels qu'il soit nécessaire et suffisant de connaître leurs troncatures à l'ordre 1000002 pour calculer la première décimale de $x+y$.

(v) On suppose que $x+y$ n'est pas un nombre décimal.
Donner une méthode de calcul de $z := x+y$ (c'est-à-dire de la suite de ses décimales). En déduire un programme de calcul (P_n) de z_n à partir des troncatures de x et y (en ne posant pas de problème pour le stockage de ces troncatures...).

(vi) Que peut-il se passer si l'on exécute le programme (P_n) sans faire d'hypothèse sur x et y ?

III.3.4 $**$ *Calcul des décimales exactes de* $\sqrt{2}$ *et de* e *à un ordre arbitraire.*

(i) Soient $a \in \mathbb{R},\ a > 0$, et $(v_h)_{h\in\mathbb{N}}$ une suite *décroissante* de réels convergeant vers a. On suppose que, pour tout $h \in \mathbb{N}$, on a $v_h - 10^{-h} \leqslant a$. Pour tous $h, p \in \mathbb{N}$, on note $(v_h)_p \in \mathcal{D}$ la troncature de v_h à l'ordre p.

a) Soient $k \in \mathbb{N}^*$ et $p \in \mathbb{N}$, avec $p \geqslant k$. On suppose que la $p^{\text{ème}}$ décimale de v_p n'est pas nulle. Montrer que v_p et a ont la même troncature à l'ordre $k-1$: $(v_h)_{k-1} = a_{k-1}$.

b) On suppose qu'il existe $k \in \mathbb{N}$ tel que, pour tout $h \in \mathbb{N}$, tel que $h > k$, la $h^{ème}$ décimale de v_h est nulle. Montrer que a est un nombre décimal.

c) On suppose que a n'est pas un nombre décimal et que l'on connaît (par un programme), pour tout $h \in \mathbb{N}$, la troncature à l'ordre h de v_h. Donner une méthode pour calculer la partie entière et un nombre arbitraire de décimales de a et l'écrire sous forme de programme. Peut-on prévoir a priori le nombre d'étapes du calcul ?

d) Que peut-on dire si l'on connaît, pour tout $h \in \mathbb{N}$, la troncature à l'ordre h de v_h, mais si l'on ne sait a priori rien sur a.

(ii) Soient $a \in \mathbb{R}$, $a > 0$, et (u_n) une suite décroissante de réels convergeant vers a.
Montrer qu'il existe une suite extraite $(v_h := u_{n_h})$ de (u_n) telle que pour tout $h \in \mathbb{N}$, l'on ait $v_h - 10^{-h} \leqslant a$.

(iii) a) Expliquer comment on peut calculer, *à tout ordre*, les décimales de $\sqrt{2}$, en justifiant la méthode esquissée dans l'exemple de la page 513.

b) Calculer $1/\sqrt{2}$ et $\sqrt{2}$ avec 6 décimales exactes. Comparer avec les nombres N et N' de la tablette d'argile babylonienne de l'université de Yale décrits dans l'exemple de la page 510.

(iv) Donner un procédé permettant de calculer un nombre arbitraire de décimales de e.
Calculer 5 décimales de e en utilisant ce procédé.

Exercices sur la sous-section 1.3

III.3.5 Calculer l'écriture décadique de $1/21$, puis celle de $1/42$.

III.3.6 Calculer dans $\mathbb{Z}_3$ l'écriture à gauche de $24/17$ avec 18 chiffres. En déduire l'écriture complète. Calculer avec 9 chiffres l'écriture de $24/51$ et $24/153$ dans $\mathbb{Q}_3$.

III.3.7 Soit p un nombre premier. Soit $x = \ldots \alpha_n \ldots \alpha_1 \alpha_0 \in \mathbb{Z}_p$. On suppose $\alpha_0 \neq 0$. Donner un algorithme simple pour calculer de proche en proche les décimales de l'inverse de x dans $\mathbb{Z}_p$. Calculer 7 décimales de l'inverse de $x = \ldots 1223\underline{1223}$ dans $\mathbb{Z}_7$.

Exercices sur la section 2

III.3.8 On note v_2 la valuation dyadique sur $\mathbb{Q}$. On note $E(r)$ la partie entière d'un nombre réel r. Soient $n \in \mathbb{N}$, $n \geqslant 2$ et $m := E(\log_2 n)$.
Montrer l'égalité : $v_2\left(\sum_{i=1}^{n} \frac{1}{i}\right) = -m$. Montrer que la somme $\sum_{i=1}^{n} \frac{1}{i}$ n'est pas un entier.

III.3.9 Soit p un nombre premier. On note $\mathbb{Z}[[X]]$ l'anneau des séries formelles à coefficients dans $\mathbb{Z}$ (*cf.* [L2, I.1, 4]). Montrer que l'on peut définir une application $\mathbb{Z}[[X]] \to \mathbb{Z}_p$ par $\sum_{i\in\mathbb{N}} a_i X^i \mapsto \sum_{i\in\mathbb{N}} a_i p^i$. Montrer que c'est un homomorphisme d'anneaux. En déduire un isomorphisme d'anneaux : $\mathbb{Z}[[X]]/(X-p) \to \mathbb{Z}_p$, où $(X-p)$ est l'idéal principal engendré par le polynôme $X-p$ dans $\mathbb{Z}[[X]]$.

III.3.10 Calculer l'écriture 5-adique de $1/9$ en utilisant l'indicatrice d'Euler.

III.3.11 Donner une preuve de la compacité de $\mathbb{Z}_p$ en utilisant des recouvrements ouverts.

III.3.12 Étudier les suites (n), $(1/n)_{n\in\mathbb{N}^*}$, (p^n), (q^n) (où q est un nombre premier différent de p) dans $\mathbb{Q}_p$.

III.3.13 Soit K un corps commutatif. On note $K[[X]]$ (resp. $K((X))$) l'anneau (resp. le corps) des séries formelles à coefficients dans $\mathbb{Z}$ (*cf.* [L2, I.1, 4]). On munit $K((X))$ de la valeur absolue $|f| := 2^{-\omega(F)}$, où $\omega(f)$ est l'ordre de la série de Laurent $f \in K((X))$ (*cf.* [L2, I.1, 4, proposition et définition 49]) et de la distance correspondante d.
Quelle est la boule unité fermée de $(K((X)), d)$? Montrer que $(K((X)), d)$ est *complet*. Montrer que $K[[X]]$ est *compact* si, et seulement si, K est un corps fini.

III.3.14 Soit p un nombre premier. Soit $c \in \mathbb{N}^*$ non divisible par p. Montrer que la suite (c^{p^n}) converge dans $\mathbb{Z}_p$. Soit γ sa limite. Montrer que $\gamma \equiv c \mod p$ et $\gamma^{p-1} = 1$. En déduire une nouvelle preuve du fait que le polynôme $x^{p-1} - 1$ a $(p-1)$ racines distinctes dans $\mathbb{Z}_p$.

III.3.15 *Formule de Legendre pour* $v_p(n!)$.
Soit p un nombre premier. On note v_p la valuation p-adique sur $\mathbb{N}$. On note $E(x)$ la partie entière d'un rationnel x. Pour $v \in \mathbb{N}^*$, on note m_v le nombre des multiples de p^v dans $[\![1, n]\!]$.

1. Soient $n \in \mathbb{N}^*$.
   Pour tout $v \in \mathbb{N}$, on note a_v le nombre d'entiers $k \in [\![1, n]\!]$ tels que $v_p(k) = v$. Montrer que $v_p(n!) = \sum\limits_{v\geqslant 1} va_v$. Montrer que $m_v := \sum\limits_{i\geqslant v} a_v$.
2. Montrer que $v_p(n!) = \sum\limits_{v\geqslant 1} m_v$.
3. Montrer que $m_v = E(n/p^v)$. En déduire que $v_p(n!) = \sum\limits_{v\geqslant 1} E(n/p^v)$.
   C'est *la formule de Legendre*.
4. On note $s_p(n)$ la somme des chiffres de l'écriture de n en base p.
   Montrer que $v_p(n!) = \dfrac{n - s_p(n)}{p-1}$.
5. Montrer[52] que la plus grande puissance de 2 divisant $n!$ est 2^{n-s} où s est le nombre de chiffres 1 dans l'écriture de n en base 2.
6. Par combien de zéros se termine l'écriture décimale de $100!$?
7. Montrer que $v_p(n!) \leqslant \frac{n}{p-1}$. Montrer que $v_p(n!) \sim \frac{n}{p-1}$, quand $n \to +\infty$.
8. Montrer que la série entière exponentielle $\exp(X) := 1 + X + X^2/2! + \cdots + X^n/n! + \cdots$ converge si, et seulement si, $|x|_p < p^{1/(p-1)}$. Montrer que, pour $p \neq 2$ et $x \in \mathbb{Q}_p$, cette condition est équivalente à $|x|_p < 1$, c'est-à-dire $x \in p\mathbb{Z}_p$. Montrer que, pour $p = 2$, elle est équivalente à $|x|_p < 1/2$, c'est-à-dire $x \in 4\mathbb{Z}_2$.

[52] La réponse est dans le livre de Legendre *Essai sur la théorie des nombres*, 1808 !

III.3.16 *Produit de deux séries convergentes.*
Soient $\sum_{n\in\mathbb{N}} a_n$ et $\sum_{n\in\mathbb{N}} b_n$ deux séries convergentes de $\mathbb{Q}_p$, de sommes respectives A et B. Montrer que la série produit de terme général $c_n = \sum_{i=0}^{n} a_i b_{n-i}$ est convergente et que sa somme est AB.

III.3.17 $*$ *Permutation des termes d'une série*
Montrer que si l'on permute les termes d'une série convergente de $\mathbb{Q}_p$ la série obtenue est encore convergente et a la même somme.

III.3.18 $*$ Soit p un nombre premier.

1. Soit $(b_{m,n})_{(m,n)\in\mathbb{N}^2}$ une *suite double* de $\mathbb{Q}_p$. montrer que les conditions suivante sont équivalentes.
   (i) Pour tout $m \in \mathbb{N}$, $\lim_{n\to+\infty} b_{m,n} = 0$ et $\lim_{m\to+\infty} b_{m,n} = 0$ uniformément en n
   (ii) Pour tout $\varepsilon > 0$, il existe $N \in \mathbb{N}$ tel que si $\max(m,n) \geqslant N$, alors $|b_{m,n}|_p < \varepsilon$.
   (iii) On a $\lim_{m\to+\infty} b_{m,n} = 0$ uniformément en n et $\lim_{n\to+\infty} b_{m,n} = 0$ uniformément en m
2. Montrer la proposition 26 de la page 461.

III.3.19 Montrer la formule de Hadamard pour le rayon de convergence d'une série entière p-adique (*cf.* (7), page 463).

III.3.20 Montrer les résultats sur le module de croissance $M_{\hat{f}}$ d'une série formelle p-adique $\hat{f}$ énoncés dans la remarque de la page 464. On dit qu'un rayon $r > 0$ est critique s'il existe deux indices *distincts* i et j tels que $M_{\hat{f}}(r) = |a_i|_p r^n = |a_j|_p r^n$. On utilisera ces rayons (s'il en existe) pour les preuves.

III.3.21 Montrer le lemme 33 de la page 465.

III.3.22 Terminer la preuve de la proposition 36 de la page 466.

III.3.23

1. Soit $a \in \mathbb{Z}_p$. Montrer que, pour tout $n \in \mathbb{N}$, $\binom{a}{n} := \frac{a(a-1)\dots(a-n+1)}{n!} \in \mathbb{Z}_p$. Montrer que la série entière $\sum_{n\in\mathbb{N}} \binom{a}{n} X^n$ converge pour $|x|_p < 1$.
2. Soit K un corps commutatif.
   On considère la série entière formelle : $\widehat{S} := \sum_{n\in\mathbb{N}} \binom{1/2}{n} X^n \in \mathbb{K}[[X]]$. Calculer son carré dans $K[[X]]$. On suppose que $x \in \mathbb{Z}_p$, avec $|x|_p < 1$. On note $S(x)$ la somme de la série convergente $\widehat{S}(x)$. Calculer $(S(x))^2$.
3. Soient $p := 7$, $a = 1/2$ et $x = 7/9$.
   Montrer que, dans $\mathbb{R}$, $(1+x)^{1/2} = 4/3$. Montrer que, dans $\mathbb{Z}_7$, $(1+x)^{1/2} - 1 \in 7\mathbb{Z}_7$ et que $(1+x)^{1/2} \neq 4/3$. Montrer que dans $\mathbb{Z}_7$, $(1+x)^{1/2} = -4/3$.

III.3.24 *Logarithme p-adique.*
Soit p un nombre premier. On munit $\mathbb{Q}_p$ de la distance p-adique.
On considère la série entière de $\mathbb{Q}_p$:

$$x - \frac{x^2}{2} + \cdots + (-1)^{n+1}\frac{x^n}{n} + \cdots \tag{10}$$

1. Montrer que $\lim\limits_{n\to+\infty} |n|_p^{1/n} = 1$ et que le domaine de convergence est $B(0,1)$. Si la série (10) converge , on note $f(x) = \log_p(1+x)$ sa somme.
   Calculer la dérivée de la fonction $f : x \mapsto \log_p(1+x)$.
2. Soit $y \in p\mathbb{Z}_p$, fixé. On définit une fonction :
   $$g : x \in p\mathbb{Z}_p \to \log_p\left((1+x)(1+y)\right) = f\left(y + (1+y)x\right) \in \mathbb{Z}_p.$$
   Montrer que $g' = f'$. En déduire que :
   $$\forall x, y \in \mathbb{Q}_p,\ \ |x|_p < 1,\ |y|_p < 1\ : \log_p\left((1+x)(1+y)\right) = \log_p(1+x) + \log_p(1+y).$$
3. On suppose $p := 2$. Montrer que $\log_2(-1)$ a un sens. Montrer que $\log_2(-1) = 0$.
4. Montrer qu'il existe une suite d'entiers naturels (N_n) tels que, pour tout $n \in \mathbb{N}^*$, $2 + \frac{2^2}{2} + \cdots + \frac{2^n}{n}$ soit divisible par 2^{N_n} (et pas par 2^{N_n+1}) et $\lim\limits_{n\to+\infty} N_n = +\infty$.
   Calculer N_5, N_6, N_7, N_{10}.
   Trouver une suite d'entiers naturels (N'_n) minorant la suite (N_n) et telle que $\lim\limits_{n\to+\infty} N'_n = +\infty$.

III.3.25 $*$ On utilise des notations et des résultats des exercices III.3.15 et III.3.24.
On note $\exp_p$ la fonction définie par la série formelle exponentielle dans son domaine de convergence $B\left(0, p^{-1/(p-1)}\right)$.

1. Montrer que, pour tous $x, y \in B\left(0, p^{-1/(p-1)}\right)$, on a $\exp_p(x+y) = \exp_p(x)\exp_p(y)$.
2. Soit $x \in B\left(0, p^{-1/(p-1)}\right)$. Montrer que $\left|\exp_p(x) - 1\right|_p < 1$.
   Soit $n \geqslant 2$. On note s la somme des chiffres de n écrit en base p.
   Montrer que $v_p(x^{n-1}/n!) > \frac{s-1}{p-1} \geqslant 0$. En déduire que, pour $n \geqslant 2$, $|x^n/n!|_p < |x|_p$.
   Montrer que $\left|\exp_p(x) - 1\right|_p = |x|_p$.
3. Montrer que, si $x \in B\left(0, p^{-1/(p-1)}\right)$ on a $\log_p \circ \exp_p(x) = x$.
4. Soit $x \in B\left(1, p^{-1/(p-1)}\right)$.
   Montrer que $\left|\log_p(1+x)\right|_p < p^{-1/(p-1)}$ et que $\exp_p \circ \log_p(1+x) = 1 + x$.
5. Montrer que les fonctions $\log_p$ et $\exp_p$ donnent des isomorphismes réciproques entre le groupe multiplicatif formé par la boule ouverte $B\left(1, p^{-1/(p-1)}\right)$ et le groupe additif formé par la boule ouverte $B\left(0, p^{-1/(p-1)}\right)$.
6. Soient $p := 2$. Comparer $\exp_2 \circ \log_2(-1)$ et -1. Commenter le résultat.

7. Montrer que si $p \neq 2$, alors $B\left(0, p^{-1/(p-1)}\right) = B'(0,1) = p\mathbb{Z}_p$. Montrer que si $p = 2$, alors $B\left(0, p^{-1/(p-1)}\right) = 4\mathbb{Z}_2$.
   Soit $n \in \mathbb{N}^*$ Montrer que :
   - si $p \neq 2$, alors $\exp_p$ et $\log_p$ induisent des isomorphismes réciproques de groupes $\exp_{p,n} : (p\mathbb{Z}/p^n\mathbb{Z}, +) \to (1+p\mathbb{Z}/p^n\mathbb{Z}, \times)$ et $\log_{p,n} : (1+p\mathbb{Z}_p, \times) \to (p\mathbb{Z}/p^n\mathbb{Z}, +)$;
   - si $p = 2$, alors $\exp_2$ et $\log_2$ induisent des isomorphismes réciproques de groupes $\exp_{2,n} : (4\mathbb{Z}/2^n\mathbb{Z}, +) \to (1+4\mathbb{Z}/2^n\mathbb{Z}, \times)$ et $\log_{2,n} : (1+p\mathbb{Z}_p, \times) \to (4\mathbb{Z}/2^n\mathbb{Z}, +)$.
8. Montrer que, dans $\mathbb{Z}_3$, $\exp_3(3) = 1 + 3 + \frac{3^2}{2!} + \frac{3^2}{2!} + \frac{3^3}{3!} + \frac{3^4}{4!} \mod 81$.
   En déduire que $\exp_3(3) = -14 \mod 81$
   On note $(\mathbb{Z}/81\mathbb{Z})^*$ le groupe multiplicatif des éléments inversibles de $\mathbb{Z}/81\mathbb{Z}$. Montrer que c'est est un groupe cyclique. Quel est son ordre ?
   Montrer que le groupe $1 + 3\mathbb{Z}_3/81\mathbb{Z}_3$ est engendré par $\exp_3(3)$. En déduire que le groupe cyclique $(\mathbb{Z}/81\mathbb{Z})^*$ est engendré par 14.

III.3.26 Soient p un entier impair et $a \in \mathbb{U} \subset \mathbb{Z}_p$. Montrer que a est un carré dans $\mathbb{U}$ si, et seulement si, c'est un carré modulo p.
Montrer que le résultat n'est pas vrai pour $p = 2$.

III.3.27 $*$ *Structure du groupe multiplicatif* $\mathbb{Q}_p^*$.
Soit p un nombre premier. Pour tout $n \in \mathbb{N}^*$, on pose $\mathbb{U}_n := 1 + p^n\mathbb{Z}_p$.

1. Montrer que les $\mathbb{U}_n$ forment une suite décroissante de sous-groupes ouverts de $\mathbb{U}$. On définit $\varprojlim \mathbb{U}/\mathbb{U}_n$ de façon similaire à $\varprojlim \mathbb{Z}/p^n\mathbb{Z}$. Vérifier que $\mathbb{U} = \varprojlim \mathbb{U}/\mathbb{U}_n$. Montrer que l'on a des isomorphismes $\mathbb{U}_n/\mathbb{U}_{n+1} \to F_p$. Quel est l'ordre de $\mathbb{U}_1/\mathbb{U}_n$? Quel est l'ordre de $\mathbb{U}/\mathbb{U}_n$?
2. Soit $n \geqslant 1$ si $p \neq 2$ et $n \geqslant 2$ si $= 2$. Soit $x \in \mathbb{U}_n \setminus \mathbb{U}_{n+1}$. Montrer que $x^p \in \mathbb{U}_{n+1} \setminus \mathbb{U}_{n+2}$.
   Soit $a := 1 + p \in \mathbb{U}_1 \setminus \mathbb{U}_2$. On note a_n l'image de a dans $\mathbb{U}_1/\mathbb{U}_n$. Montrer que $\mathbb{U}_1/\mathbb{U}_n$ est un groupe cyclique engendré par a_n.
3. Soit $p \neq 2$. Montrer que $\mathbb{U}_1$ est isomorphe à $\mathbb{Z}_p$ (on utilisera les homomorphismes $\theta_n : \mathbb{Z}/p^{n-1}\mathbb{Z} \to \mathbb{U}_1/\mathbb{U}_n$ définis par $y \mapsto a_n^y$).
4. Soit $p = 2$. Définir un isomorphisme $\mathbb{Z}_2 \to \mathbb{U}_2$, appliquant $2\mathbb{Z}_2$ sur $\mathbb{U}_3$.
   Montrer que $\mathbb{U}_1 = \{\pm 1\} \times \mathbb{U}_2 \approx \{\pm 1\} \times \mathbb{Z}_2$.
5. Montrer que le groupe multipicatif $\mathbb{Q}_p^*$ est isomorphe à $\mathbb{Z} \times \mathbb{Z}_p \times \mathbb{Z}/(p-1)\mathbb{Z}$ si $p \neq 2$ et à $\mathbb{Z} \times \mathbb{Z}_2 \times \mathbb{Z}/2\mathbb{Z}$ si $p = 2$.
6. Comparer avec la question 7 de l'exercice III.3.25.

III.3.28 Soit p un nombre premier *impair*.

1. Montrer qu'un élément $x \in \mathbb{Q}_p^*$ est un carré si, et seulement si, $x = p^{2n}u$, avec $n \in \mathbb{Z}$ et $u \in \mathbb{U}$, l'image modulo p de u dans F_p^* étant un carré (on utilisera l'exercice III.3.26).
2. Donner une autre preuve de 1 en utilisant l'exercice III.3.27 ci-dessus.
3. Montrer que $\mathbb{Q}_p^*/(\mathbb{Q}_p^*)^2$ est d'ordre 4. Soit $c \in \mathbb{Z}_p^*$ un élément dont l'image modulo p dans F_p^* n'est pas un carré de F_p^*. Montrer que l'on a des représentants de la forme $1, p, c, cp$ des classes modulo $(\mathbb{Q}_p^*)^2$ dans $\mathbb{Q}_p^*$.

III.3.29 On utilise les notations et les résultats de l'exercice III.3.27. Soit $p = 2$.

1. Montrer que l'ensemble des carrés de $\mathbb{U}_2$ est égal à $\mathbb{U}_3$. En déduire que l'ensemble des carrés de $\mathbb{U}$ est égal à $\mathbb{U}_3$.
2. Montrer qu'un élément $x \in \mathbb{Q}_2^*$ est un carré si, et seulement si, $x = p^{2n}u$, avec $n \in \mathbb{Z}$ et u congru à 1 modulo 8.
3. Montrer que $\mathbb{Q}_2^*/(\mathbb{Q}_2^*)^2$ est d'ordre 8. Montrer que ± 1, ± 5, ± 2, ± 10 sont des représentants des classes modulo $(\mathbb{Q}_2^*)^2$ dans $\mathbb{Q}_2^*$.

III.3.30

1. Montrer la proposition 43 de la page 472 (on utilisera deux fois une formule de Taylor).
2. Soient $r \in \mathbb{N}^*$ et $m, k \in \mathbb{N}$ tels que $0 \leqslant 2k < m$. Soient $f \in \mathbb{Z}_p[X_1, \ldots, X_r]$, $x = (x_i) \in (\mathbb{Z}_p)^r$ et $j \in [\![1, r]\!]$. On suppose que $f(x) \equiv 0 \mod p^m$ et $v_p\left(\frac{\partial f}{\partial X_j}(x)\right) = k$. Montrer qu'il existe un zéro $y \in (\mathbb{Z}_p)^r$ de f qui est congru à x modulo p^{m-k}.
3. Un point $x = (x_i) \in (\mathbb{Z}_p)^r$ est dit *primitif* si l'un des x_i n'est pas divisible par p.
   Soit $f := \sum a_{i,j} X_i X_j$ une forme quadratique sur $(\mathbb{Z}_p)^r$ dont le discriminant est inversible dans $\mathbb{Z}_p$. Soient $a \in \mathbb{Z}_p$ et $x \in (\mathbb{Z}_p)^r$ *primitif* tel que $f(x) \equiv a \mod p$. Montrer qu'il existe $y \in (\mathbb{Z}_p)^r$ tel que $f(y) = a$ et $y \equiv x \mod p$.

III.3.31 On considère l'isomorphisme d'anneaux $\varpi : Z_{10} \to \mathbb{Z}_5 \times \mathbb{Z}_2$ (*cf.* la proposition 47 de la page 476). On identifie $\mathbb{Z} \times \mathbb{Z}$ à un sous-anneau de $\mathbb{Z}_5 \times \mathbb{Z}_2$.

1. Soit $(a, b) \in \mathbb{Z}^2$. On note $x := \varpi^{-1}((a, b))$. Donner une méthode de calcul d'un nombre arbitraire de chiffres de x.
2. Calculer les 5 chiffres de droite de l'écriture décadique de x pour $a := 2$ et $b := 3$. Retrouver le résultat en utilisant l'algorithme des restes chinois.

III.3.32 *Automorphismes de corps des* $\mathbb{Q}_p$. Soit p un nombre premier.

1. Montrer que, pour $a \in \mathbb{Q}_p$ les deux conditions suivantes sont équivalentes :
   (i) $a \in \mathbb{U}$;
   (ii) a^{p-1} admet une racine $n^{\text{ème}}$ pour *une infinité* de valeurs de $n \in \mathbb{N}^*$.
2. Montrer que le seul automorphisme de corps de $\mathbb{Q}_p$ est l'identité.
3. Que peut-on dire d'un automorphisme de corps de $\mathbb{R}$?

III.3.33 Soit p un nombre premier. Montrer que les corps $\mathbb{R}$ et $\mathbb{Q}_p$ ne sont pas isomorphes.

III.3.34 *Caractères du groupe additif* $\mathbb{Q}_p$.
Soit p un nombre premier. Un caractère de $(\mathbb{Q}_p, +)$ est, par définition, un homomorphisme *continu* du groupe additif $\mathbb{Q}_p$ dans le groupe multiplicatif des nombres complexes $\mathbb{C}^*$. On note U le groupe multiplicatif des nombres complexes de module 1.
Soit χ un caractère de $(\mathbb{Q}_p, +)$.

1. Montrer que l'image de χ est un sous-groupe du groupe U des nombres complexes de module 1.

2. Montrer qu'il existe un voisinage ouvert $\mathcal{V}$ de 1 dans U ne contenant pas de sous-groupe non trivial de U. En déduire que le noyau de χ contient un sous-groupe de la forme $p^k\mathbb{Z}_p$ avec $k \in \mathbb{N}^*$.
3. Montrer que l'image de χ est contenue dans le sous-groupe des racines de l'unité d'ordre une puissance arbitraire de p.
4. Soit θ un homomorphisme *continu* du groupe multiplicatif $\mathbb{Z}_p^*$ dans U. Montrer que son image est contenue dans U et que son noyau est un sous-groupe ouvert de $\mathbb{Z}_p^*$.

Exercice sur la section 3

III.3.35 Soit $\mathcal{S} \subset \mathcal{P}(\mathbb{Q})$ l'ensemble des sections commençantes de $\mathbb{Q}$. On a une inclusion stricte $\mathbb{R} \subset \mathcal{S}$. Montrer que l'on peut définir sur $\mathcal{S}$ une relation d'ordre total et une addition induisant la relation d'ordre et l'addition sur $\mathbb{R}$. Montrer que $\mathcal{S}$ possède la propriété de la borne supérieure et que l'addition est associative et commutative et possède un élément neutre, mais qu'un élément de $\mathcal{S}$ n'a pas nécessairement d'opposé.

Exercices sur la section 4

III.3.36 *Axiomes des nombres réels.*
Soit K un corps totalement ordonné contenant $\mathbb{Q}$ comme sous-corps. Montrer que les conditions suivantes sont équivalentes.

(i) La propriété de la borne supérieure est vérifiée.
(ii) Le corps K est archimédien et il possède la propriété des intervalles emboîtés.
(ii) Le corps K est archimédien et complet.
(iii) Le corps K est complet et $\mathbb{Q}$ est dense dans K.
(iv) Tout segment de K est compact.
(v) Les compacts de K sont les sous-ensembles fermés bornés.
(vi) Le corps K est localement compact.
(vii) Le corps K est connexe.

III.3.37 *Un corps ordonné non archimédien.*
On considère l'anneau $\mathbb{R}[X]$. Pour $R \in \mathbb{R}[X]$, on dit que $R > 0$ si $R \neq 0$ et si son coefficient dominant est positif. Pour $P,\ Q \in \mathbb{R}[X]$, on dit que $P \leqslant Q$ si $P = Q$ ou $Q - P > 0$. Montrer que $(\mathbb{R}[X], +, \times, \leqslant)$ est un *anneau totalement ordonné non archimédien*. En déduire une structure de *corps totalement ordonné non archimédien* sur $\mathbb{R}(X)$.

III.3.38 $*$ *Corps ordonnables.*
Soit K un corps commutatif. On dit que la partie $P \subset K$ satisfait la propriété $(*)$ si les conditions suivantes sont satisfaites : $P + P \subset P,\ \ P.P \subset P,\ \ -1 \notin P,$ et $\forall x \in K,\ x^2 \in P$.
Montrer que les trois conditions suivantes sont équivalentes (pour prouver (iii) $\Rightarrow$ (i), on pourra utiliser $(*)$ et le lemme de Zorn ; *cf.* [L1] pour l'énoncé de celui-ci).

(i) Le corps K est ordonnable.
(ii) Soient $x_1, \ldots, n \in K$. Si $x_1^2 + \cdots = x_n^2 = 0$, alors $x_1 = \ldots = x_n = 0$.
(iii) L'élément $-1 \in K$ n'est pas une somme de carrés de K.

III.3.39 Soit K un corps. On note A *l'anneau premier* de K, c'est-à-dire l'image de $\mathbb{Z}$ dans K par l'application définie par $n \mapsto n1_k$.

1. Montrer qu'une norme $|\ |$ sur K est ultramétrique si, et seulement si, pour tout $z \in K$, on a $|z+1| \leqslant \max(|z|, 1)$.
2. Montrer qu'une norme $|\ |$ sur K est non-archimédienne si, et seulement si, pour tout $a \in A$, on a $|a| \leqslant 1$.
3. On note $|A| := \{|a| \mid a \in A\}$. Montrer qu'une norme $|\ |$ sur K est non archimédienne si, et seulement si, $|A|$ est *majoré* dans $\mathbb{R}$. Montrer que si K est *de caractéristique non-nulle*, alors toute norme sur K est ultramétrique.
4. Montrer qu'une norme $|\ |$ sur K est archimédienne si, et seulement si, l'on a la propriété (P) suivante : étant donnés $x, y \in K$, $x \neq 0$, il existe $n \in \mathbb{N}^*$ tel que $|nx| > |y|$.

III.3.40 Montrer le lemme 84 et la proposition 85 de la page 502.

III.3.41 Montrer que $\mathbb{C}$ est le complèté de $\mathbb{Q}(i)$ pour la norme $|a+ib| = \sqrt{a^2+b^2}$. Montrer que si L est un corps *normé complet* et si $K \subset L$ est un sous-corps *dense* de L, alors L est le complété de K pour la norme induite.

III.3.42 $*$ *Normes sur* $K(X)$.
Soit K un corps commutatif. Soit $\mathcal{P}$ l'ensemble des polynômes irréductibles unitaires de $K[X]$.

1. Montrer que $f \mapsto |f|_\infty := 2^{\deg f}$ $(f \neq 0)$ définit une norme ultramétrique $|\ |_\infty$ sur $K(X)$.
2. Montrer que si $P \in \mathcal{P}$ et $f \neq 0$, il existe un unique $v_P(f) \in \mathbb{Z}$ tel que $f = P^{v_P(f)} \frac{Q_1}{Q_2}$ avec $Q_1, Q_2 \in K[X]$ premiers avec P. Montrer que $f \mapsto |f|_P := 2^{-(\deg P)\, v_P(f)}$ définit une norme ultramétrique $|\ |_P$ sur $K(X)$.
3. Donner une interprétation simple des normes $|\ |_\infty$ et $|\ |_P$ quand $K = \mathbb{C}$ en utilisant la droite projective $P^1(\mathbb{C})$. La traduire en termes de résidus.
4. Montrer que toute norme non triviale sur $K(X)$, triviale sur K, est équivalente à $|\ |_\infty$ ou à $|\ |_P$ pour un $P \in \mathcal{P}$ (on utilisera l'exercice III.3.39).
5. Soit K un corps *fini*. Montrer que toute norme non triviale sur $K(X)$ est équivalente à $|\ |_\infty$ ou à $|\ |_P$ pour un $P \in \mathcal{P}$.
6. On suppose $K := \mathbb{Q}$. Montrer qu'il existe une norme sur $\mathbb{Q}[X]$ qui n'est équivalente ni à $|\ |_\infty$ ni à $|\ |_P$ pour un $P \in \mathcal{P}$. On pourra utiliser un nombre transcendant pour interpréter $\mathbb{Q}[X]$ comme sous-anneau de $\mathbb{R}$ (ou de $\mathbb{Q}_p$).
7. Montrer que tout $f \in K(X)$ non nul vérifie une formule du produit : $|f|_\infty \prod_{P \in \mathcal{P}} |f|_P = 1$.

   Comparer avec une formule de résidus si $K = \mathbb{C}$.

III.3.43 *Normes sur un K-espace vectoriel de dimension finie.*
Soit $(K, |\ |)$ un corps normé *complet*

1. Soit V un K-espace vectoriel *de dimension finie*. Montrer que toutes les normes $\|\ \|$ sur V, telles que $\|\lambda x\| = |\lambda| \|x\|$, si $x \in V$ et $\lambda \in K$, sont *équivalentes* et que V est complet pour n'importe laquelle d'entre-elles.

2. Soit L un corps qui est une *extension finie* de K. On suppose qu'il existe une norme sur L dont la restriction à K est $|\ |$. Montrer que toutes les normes sur L dont la restriction à K est $|\ |$ sont équivalentes.
3. Soit p un nombre premier. Soient L une extension finie de $\mathbb{Q}_p$ et $\|\ \|_1$, $\|\ \|_2$ deux normes sur le corps L qui prolongent la norme p-adique $|\ |_p$ sur $\mathbb{Q}_p$. Montrer que $\|\ \|_1 = \|\ \|_2$.

III.3.44 $*$ Soit p un nombre premier. Soit L une extension finie de $\mathbb{Q}_p$, de degré d. À $x \in L$ on associe l'application $\mathbb{Q}_p$-linéaire $\ell_x : L \to L$, définie par $\ell : y \in L \mapsto xy \in L$.
On a $N(x) := \det \ell_x \in \mathbb{Q}_p$. Montrer que l'application $f : L \to \mathbb{R}^+$, définie par

$$f : x \in L \mapsto |N(x)|_p^{1/d}$$

est une norme sur L prolongeant $|\ |_p$.

III.3.45 $*$ *Théorème d'approximation d'Artin-Whapples.*
Soient K un corps et : $|\ |_1, \ldots, |\ |_s$ des normes sur K *deux à deux non équivalentes*.
Soient $x_1, \ldots, x_s \in K$ et $\varepsilon > 0$.

1. Soit $i \in [\![1, s]\!]$. Montrer qu'il existe $z_i \in K$ tel que $|z_i|_i > 1$ et $|z_i|_j < 1$ pour $j \neq i$ (on pourra utiliser les suites (x^n) et $(x^n/(1+x^n))$).
2. Montrer qu'il existe $x \in K$ tel que, pour tout $i \in [\![1, s]\!]$, $|x - x_i|_i < \varepsilon$.
   C'est le *théorème d'approximation d'Artin-Whapples*.
3. Le théorème d'Artin-Whapples ressemble à un résultat bien connu, lequel ?

III.3.46 Terminer la preuve du théorème 92 de la page 506 (théorème d'Ostrowski).

Polynômes orthogonaux

III.4

Dans ce chapitre, nous donnons une présentation assez complète de la théorie des polynômes orthogonaux et de nombreuses applications fondamentales de celle-ci, en utilisant « de façon croisée » les outils *élémentaires* d'algèbre et d'analyse présentés dans les divers modules du livre [L2]. C'est une exposition un peu inhabituelle du sujet, qui est d'habitude traité avec des techniques plus sophistiquées. Nous avons fait délibérément ce choix ; le lecteur pourra constater que, s'il permet d'aller assez loin, le caractère élémentaire de l'exposition se « paie », pour un certain nombre de résultats délicats, par des calculs un peu lourds ou des hypothèses un peu techniques. Il comprendra ainsi l'utilité d'outils plus puissants et plus abstraits. Par exemple, les techniques hilbertiennes sont plus efficaces et plus élégantes que les techniques préhilbertiennes utilisées ici (mais elles nécessitent une théorie de l'intégration plus approfondie, dans le prolongement du module III.1). Les formes plus générales de la formule de Cauchy, en théorie des fonctions analytiques, ou l'utilisation de certaines fonctions spéciales (comme les fonctions hypergéométriques) permettraient d'alléger certains calculs et, surtout, de mieux comprendre certaines structures qui apparaissent ici "en filigrane", mais ceci sortirait de notre cadre élémentaire.

Nous insistons sur le cas[1] des polynômes orthogonaux dits « classiques » (cas de Legendre, Jacobi, Laguerre, Hermite,...), qui sont particulièrement importants pour les applications.

Certaines applications des polynômes orthogonaux classiques (fonctions sphériques, oscillateur harmonique, atome d'hydrogène) sont plus particulièrement destinées aux étudiants intéressés par la physique. En relation avec ces applications, nous étudions quelques exemples fondamentaux d'équations aux dérivées partielles, complétant le module *Équations aux dérivées partielles*.

1 Introduction

Les *polynômes orthogonaux* apparaissent dans de nombreux domaines des mathématiques (théorie de l'approximation, fonctions spéciales, combinatoire, probabilités et statistique, équations aux dérivées partielles, approximants de Padé, fractions continues, équations intégrales, représentations des groupes...), de la physique (vibrations et propagation des ondes, mécanique quantique, mécanique statistique...) et des sciences de l'ingénieur. Bien qu'as-

[1] Nous donnons des variantes "discrètes" des polynômes orthogonaux classiques sous forme d'exercices.

sez élémentaire, ce sujet a maheureusement quasiment disparu des cours de mathématiques au niveau de la Licence dans les universités françaises, survivant seulement dans de fort nombreux problèmes d'examens ou de concours ! Nous pensons que c'est une grave erreur. Son étude reste toujours importante pour la physique, elle est par ailleurs en plein renouveau en mathématiques avec de nombreuses applications récentes[2]. Par ailleurs, l'étude de ce module permettra à l'étudiant d'illustrer et de mieux comprendre d'autres parties du cours de mathématiques de Licence (algèbre linéaire et bilinéaire, probabilités, analyse réelle et complexe, équations différentielles, analyse numérique, algorithmique...) et leurs *interactions*. Elle devrait, de surcroît, l'aider à réussir examens et concours...

On peut étudier les polynômes orthogonaux (pour un *poids* ou, plus généralement[3] pour une forme linéaire sur $\mathbb{R}[T]$) sous différents angles. Conformément à l'usage « moderne », nous commencerons par exposer la *théorie générale* en partant du procédé d'orthogonalisation de Gram-Schmidt, vu dans le module *Espaces préhilbertiens* du livre [L2]. Nous appliquerons ensuite cette théorie aux *polynômes orthogonaux classiques* : les polynômes de Jacobi (et en particulier de Legendre, Tchebychev...), de Laguerre (généralisés) et d'Hermite. Nous mettrons en œuvre des techniques variées pour les étudier en insistant sur leurs étroites relations avec les *équations différentielles linéaires* et nous donnerons deux applications importantes en physique (l'oscillateur harmonique et l'atome d'hydrogène en mécanique quantique). Nous reviendrons ensuite au cas général, en terminant par quelques *compléments* sur des aspects « anciens » du sujet revenus récemment sur le devant de la scène[4]. La théorie *générale* des polynômes orthogonaux s'est développée historiquement (d'Euler à Stieljes) de façon assez différente de sa présentation moderne, en relation étroite avec l'étude des *fractions continues*, en particulier avec des travaux, au XIXe siècle, de Laguerre (1879), Laplace (1882), Gauss, Jacobi, Heine, Tchebychev, Stieltjes (1894)... La relation entre polynômes orthogonaux et fractions continues est importante et source de nombreuses applications. Elle a récemment réapparu dans les développements contemporains du sujet. En accord avec ce contexte, nous la traiterons en utilisant la *transformation de Stieltjes* qui fait intervenir des fonctions holomorphes (*cf.* le module *Fonctions analytiques* du livre [L2]).

2 Généralités

2.1 Introduction

2.1.1 Définitions

Dans tout ce module, conformément à la convention faite dans le module *Polynômes et fractions rationnelles* du livre [L1] (section *Polynômes sur les corps* $\mathbb{R}$ *et* $\mathbb{C}$, page 310), nous identifierons la $\mathbb{R}$-algèbre $\mathbb{R}[T]$ des polynômes à une indéterminée à coefficients réels et la $\mathbb{R}$-algèbre des fonctions polynomiales réelles sur $\mathbb{R}$.

Soit (α, β), avec $\alpha \in \mathbb{R} \cup \{-\infty\}, \beta \in \mathbb{R} \cup \{+\infty\}$, un intervalle donné de $\mathbb{R}$, nous supposerons toujours $\alpha < \beta$ et nous identifierons l'algèbre $\mathbb{R}[T]$ à l'algèbre obtenue par restriction à (α, β), donc à une sous-algèbre de l'espace vectoriel réel $\mathcal{C}(\alpha, \beta)$ des fonctions continues

[2] En relation, par exemple, avec la théorie des matrices aléatoires.

[3] Cette approche est essentiellement algébrique. Par contre, la notion de poids nécessite des notions d'intégrale ou, plus généralemnt de mesure.

[4] *Cf.* par exemple le livre récent de Percy Deift [Dei].

sur (α, β). L'intervalle (α, β) sera par ailleurs quelconque (ouvert, fermé, semi-ouvert, borné ou non). Nous considérerons sur un tel intervalle des *fonctions poids* w (l'initiale de « weight »); ces fonctions seront toujours supposées continues *positives* et non identiquement nulles sur l'intervalle (α, β). *Nous supposerons toujours* que les *moments*[5], notés μ_n, associés à w sont « *tous finis* », c'est-à-dire que les intégrales $\mu_n := \displaystyle\int_\alpha^\beta t^n w(t)\,\mathrm{d}t$ sont *absolument* convergentes pour tout $n \in \mathbb{N}$, ces intégrales étant prises *au sens des intégrales de Riemann*, éventuellement *impropres* (*cf.* le module *Intégration* du livre [L2]). Cette condition est évidemment vérifiée si l'intervalle $(\alpha, \beta) = [\alpha, \beta]$ est fermé et borné, c'est-à-dire un segment. Quitte à remplacer w par $\frac{w}{\mu_0}$, on peut supposer que $\mu_0 = 1$, c'est-à-dire que $w(t)\,\mathrm{d}t$ définit une mesure de *probabilité* sur (α, β).
Si tous les moments sont finis, pour tout polynôme $p(t) := a_n t^n + \cdots + a_0 \in \mathbb{R}[T]$, l'intégrale $\displaystyle\int_\alpha^\beta p(t)w(t)\,\mathrm{d}t$ est absolument convergente et l'on a $\displaystyle\int_\alpha^\beta p(t)w(t)\,\mathrm{d}t = \sum_{i=0}^n a_i\mu_i$.
Si $p \in \mathbb{R}[T]$ et $\displaystyle\int_\alpha^\beta p(t)^2 w(t)\,\mathrm{d}t = 0$, on a, en utilisant le lemme 1 ci-dessous, $p^2 = 0$, donc $p = 0$. Ainsi, $(\,p \mid q\,)_w = \displaystyle\int_\alpha^\beta p(t)q(t)w(t)\,\mathrm{d}t$ définit un *produit scalaire* sur $\mathbb{R}[T]$.
Si $(\,p \mid q\,)_w = 0$, nous dirons que p et q sont *orthogonaux* relativement au poids w. S'il n'y a pas de confusion possible, nous écrirons parfois $(\ \mid\ )$ au lieu de $(\ \mid\ )_w$.
Pour tout $n \in \mathbb{N}$, on a $\mu(t^{2n}) = (\,t^n \mid t^n\,)_w > 0$.
Si nécessaire, nous interpréterons le poids w comme une fonction sur $\mathbb{R}$ en le prolongeant par 0 sur le complémentaire de (α, β). Avec cette convention, pour tous $c,\ d \in \mathbb{R}$, les intégrales $\displaystyle\int_c^d w(t)\,\mathrm{d}t$ sont convergentes.

Lemme 1. Soit w un poids sur $]\alpha, \beta[$, vérifiant les conditions ci-dessus. Soit $q \in \mathbb{R}[T]$ tel que $q \geqslant 0$ sur $]\alpha, \beta[$. Si $\displaystyle\int_\alpha^\beta q(t)w(t)\,\mathrm{d}t = 0$, alors $q = 0$.

Démonstration. La fonction wq est continue et positive (au sens large) sur $]\alpha, \beta[$. On a donc $wq = 0$ sur cet intervalle. La fonction $t \mapsto q(t)$ est identiquement nulle sur tout intervalle ouvert sur lequel la fonction w est strictement positive (il en existe, puisque cette dernière est continue et non identiquement nulle), donc nulle puisqu'elle est polynomiale. ∎

Remarque. Si l'on dispose de la théorie de l'intégrale de Lebesgue (*cf.* le module III.1), on peut développer une théorie plus générale en prenant pour poids w une fonction intégrable au sens de Lebesgue sur l'intervalle (α, β), positive presque partout, non presque partout nulle et à moments tous finis. Si l'on dispose de la théorie de la mesure (*cf.* le module III.1) et de la théorie de l'intégration correspondante (*cf.* par exemple [Ru]), on peut, plus généralement, remplacer la mesure de densité w

[5] Le terme moment vient de la mécanique. Si l'on a une distribution de masses de densité locale w sur le segment $[\alpha, \beta]$, μ_2 est le moment du segment par rapport à l'origine.

par une *mesure borélienne* (c'est-à-dire une mesure sur la tribu de Borel de $\mathbb{R}$, *cf.* le module III.1) positive de support infini. On peut enfin utiliser la notion, *a priori* encore plus générale[6], de *polynômes orthogonaux associés à une forme linéaire strictement positive* (*cf.* ci-dessous page 536), qui nous permettra d'étendre facilement un certain nombre de résultats obtenus dans le cas d'un poids en restant dans un cadre élémentaire.

Définition 1 (Polynômes orthogonaux). Une suite de *polynômes orthogonaux*, relativement au poids w et à l'intervalle (α, β) (vérifiant les conditions ci-dessus), est une suite de polynômes réels $(p_n)_{n\in\mathbb{N}}$ telle que $\deg p_n := n$, pour tout $n \in \mathbb{N}$, et $(p_n \mid p_m)_w = 0$, pour tous $n, m \in \mathbb{N}$, $n \neq m$.

Comme nous l'avons vu dans le module *Algèbre bilinéaire* du livre [L2], on obtient une suite $(u_n)_{n\in\mathbb{N}}$ de polynômes orthogonaux en appliquant le procédé d'orthogonalisation de Gram-Schmidt au système libre $(t^n)_{n\in\mathbb{N}}$ des monômes de $\mathbb{R}[T]$. On remarque que les polynômes u_n obtenus sont tous *unitaires* : $u_n(t) = t^n + \cdots$, pour tout $n \in \mathbb{N}$.
Si $(p_n)_{n\in\mathbb{N}}$ est une suite de polynômes orthogonaux, pour tout $n > 0$, p_n est orthogonal à tous les polynômes de degré strictement inférieur à n, c'est-à-dire à $\mathbb{R}_{n-1}[T]$. La droite engendrée par p_n est l'orthogonal de $\mathbb{R}_{n-1}[T]$ dans $\mathbb{R}_n[T]$. Elle est *indépendante* du choix initial de la suite de polynômes orthogonaux. Pour un intervalle et un poids donné, chaque suite $(p_n)_{n\in\mathbb{N}}$ de polynômes orthogonaux est déterminée, de manière unique, par le choix des coefficients des termes de plus haut degré des p_n qui est *arbitraire*. Nous choisirons ces coefficients *positifs* (sauf dans quelques exemples particuliers de polynômes classiques) et nous noterons $p_n(t) := \gamma_n t^n + \cdots$.

Normalisation.. Un choix « systématique » des γ_n sera appelé « normalisation » de la suite de polynômes orthogonaux. Il n'y a malheureusement pas de choix vraiment privilégié. On peut choisir tous les polynômes *unitaires* : $\gamma_n := 1$, pour tout $n \in \mathbb{N}$; nous noterons en général $(u_n)_{n\in\mathbb{N}}$ une telle suite. On peut aussi *normer* la suite, en remplaçant p_n par $q_n := p_n/\|p_n\|$; dans ce cas, la nouvelle valeur de γ_n, que nous noterons ξ_n ($q_n(t) = \xi_n t^n + \cdots$), est $\xi_n = \gamma_n/\|p_n\| = 1/\|u_n\|$. Dans le cas des polynômes orthogonaux classiques, d'autres normalisations se sont souvent imposées pour des raisons diverses (par exemple $P_n(1) = 1$ pour les polynômes de Legendre P_n). Selon les auteurs, le domaine d'application..., il peut même exister plusieurs normalisations. Il faut traiter ces questions de normalisation avec soin ; elles sont fréquemment une source d'erreurs de calcul... Nous verrons sur des exemples que la comparaison de diverses normalisations peut être délicate.

Remarque. Si $\alpha = -\beta$ et si le poids w est une fonction *paire*, les polynômes orthogonaux de degré pair sont pairs et les polynômes orthogonaux de degré impair sont impairs. En effet, si $(p_n)_{n\in\mathbb{N}}$ est une suite de polynômes orthogonaux ($p_n(t) = \gamma_n t^n + \cdots$), $(p_n(-t))_{n\in\mathbb{N}}$ est aussi une suite de polynômes orthogonaux (on change t en $-t$ dans les relations d'orthogonalité). On a donc $p_n(-t) = \lambda_n p_n(t)$ ($n \in \mathbb{N}^*$, $\lambda_n \in \mathbb{R}$), $\lambda_n(-1)^n\gamma_n = \gamma_n$ et $\lambda_n = (-1)^n$: $p_n(-t) = (-1)^n p_n(t)$.

[6] D'après un théorème difficile, dû à Favard et Shohat, cette généralisation n'est qu'apparente.

Proposition 2. Soit (p_n) une suite de polynômes orthogonaux. Si $k > m$, p_k est orthogonal à $\mathbb{R}_m[T]$: $(\,p_k \mid p\,) = \displaystyle\int_\alpha^\beta p_k(t)p(t)w(t)\,\mathrm{d}t = 0$, pour tout polynôme p de degré au plus m. En particulier, si $k > 0$, $(\,p_k \mid 1\,) = \displaystyle\int_\alpha^\beta p_k(t)w(t)\,\mathrm{d}t = 0$.

Exemples. Les exemples suivants seront étudiés en détail dans la suite.

(i) Polynômes de *Legendre*. L'intervalle est $[-1, 1]$ et le poids $w(t) := 1$ (*cf.* l'exercice 5 de la page 198 du module *Espaces préhilbertiens* du livre [L2]). On est dans le cas de la remarque ci-dessus, donc les p_n sont pairs, pour n pair, impairs sinon. Oa a $\mu_n = 0$ si n est impair, $\mu_n = \displaystyle\int_{-1}^1 t^n\,\mathrm{d}t = \frac{2}{n+1}$ si n est pair et $(\,t^n \mid t^m\,) = \mu_{m+n}$.

Le procédé d'orthogonalisation de Gram-Schmidt donne, nous l'avons vu :

$$u_0(t) = 1, \qquad u_1(t) = t, \qquad u_2(t) = t^2 - \frac{1}{3} \quad \text{et} \quad u_3(t) = t^3 - \frac{3}{5}t.$$

La normalisation traditionnelle est $P_n(1) = 1$, soit :

$$P_0(t) = 1, \qquad P_1(t) = t, \qquad P_2(t) = \frac{1}{2}(3t^2 - 1) \quad \text{et} \quad P_3(t) = \frac{1}{2}(5t^3 - 3t).$$

(ii) Polynômes de *Laguerre*. L'intervalle est $[0, +\infty[$ et le poids $w(t) = \mathrm{e}^{-t}$. On a :

$$\mu_n = \int_0^{+\infty} t^n \mathrm{e}^{-t}\,\mathrm{d}t = n!.$$

Les premiers polynômes de Laguerre sont :

$$L_0(t) = 1 \qquad\qquad L_1(t) = -t + 1$$

$$L_2(t) = t^2 - 4t + 2 \qquad\qquad L_3(t) = -t^3 + 9t^2 - 18t + 6.$$

Nous normaliserons en choisissant $(-1)^n$ comme coefficient du terme de plus haut degré de L_n (on trouve aussi fréquemment le choix $(-1)^n/n!$ pour ce coefficient).

(iii) Polynômes *d'Hermite* . L'intervalle est $\mathbb{R} =]-\infty, +\infty[$ et le poids $w(t) = e^{-t^2}$ (certains auteurs, en particulier en probabilités et statistiques, choisissent plutôt le poids $w(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$). La normalisation traditionnelle consiste à choisir 2^n comme coefficient du terme de plus haut degré. On est dans le cas de la remarque ci-dessus, donc les p_n sont pairs, pour n pair, impairs sinon.

On a $\mu_n = \displaystyle\int_{-\infty}^{+\infty} t^n \mathrm{e}^{-t^2}\,\mathrm{d}t = \int_0^{+\infty} u^{\frac{n+1}{2}-1}\mathrm{e}^{-u}\,\mathrm{d}u = \Gamma\left(\frac{n+1}{2}\right)$ (Γ désigne la fonction Gamma, *cf.* le module *Intégration* du livre [L2]).

Les premiers polynômes d'Hermite sont :

$$H_0(t) = 1, \qquad H_1(t) = 2t, \qquad H_2(t) = 4t^2 - 2 \quad \text{et} \quad H_3(t) = 8t^3 - 12t.$$

Pour les cas (ii) et (iii) le lecteur vérifiera les relations d'orthogonalité. Il pourra aussi à titre d'exercice appliquer à ces cas le procédé d'orthogonalisation de Gram-Schmidt.

On retrouve des variantes des poids définis dans ces trois exemples dans le module IV.3. Ces poids sont des densités de probabilité de lois usuelles (loi uniforme, loi exponentielle, loi gaussienne).

Polynômes orthogonaux relativement à une forme linéaire sur $\mathbb{R}[T]$

À un poids w sur (α, β) dont tous les moments sont finis, on peut associer la forme linéaire sur $\mathbb{R}[T]$: $\mathscr{M}_w : p \in \mathbb{R}[T] \mapsto \displaystyle\int_\alpha^\beta p(t)w(t)\,\mathrm{d}t \in \mathbb{R}$.
On a, pour tous $p,\ q \in \mathbb{R}[T]$, $(\,p \mid q\,)_w = \mathscr{M}(pq)$ et $\|p\|_w^2 = \mathscr{M}(p^2)$.
On peut se placer dans un *cadre plus général* en se donnant une forme linéaire *arbitraire* $\mathscr{M}$ sur $\mathbb{R}[T]$ et définir une forme bilinéaire symétrique $(p, q) \mapsto (\,p \mid q\,)_{\mathscr{M}} := \mathscr{M}(pq)$ sur $\mathbb{R}[T]$.
On dira que la forme $\mathscr{M}$ est *positive* si, pour tout $p \in \mathbb{R}[T]$ positif, c'est-à-dire tel que $p(t) \geqslant 0$, on a $\mathscr{M}(p) \geqslant 0$. On dira qu'elle est *strictement positive* si elle est positive et si, pour p positif, $\mathscr{M}(p) = 0$, entraîne $p = 0$.
Si $\mathscr{M}$ est positive (respectivement strictement positive), la forme $(\,.\mid .\,)_{\mathscr{M}}$ est positive (respectivement strictement positive).
D'après le lemme 1, dans le cas d'un poids w, la forme $\mathscr{M}_w$ est strictement positive.
Si $\mathscr{M}$ est strictement positive, $(\,.\mid .\,)_{\mathscr{M}}$ est un produit scalaire. On note $\|.\|_{\mathscr{M}}$ la norme associée : $\|p\|_{\mathscr{M}}^2 := (\,p \mid p\,)_{\mathscr{M}}$.

Définition 2. Soit $\mathscr{M}$ une forme linéaire sur $\mathbb{R}[T]$. On dit qu'une suite $(p_n)_{n\in\mathbb{N}}$ de $\mathbb{R}[T]$ est une *suite de polynômes orthogonaux pour* $\mathscr{M}$ (ou $\mathscr{M}$-orthogonaux) si :

1. pour tout $n \in \mathbb{N}$, $\deg p_n = n$;
2. pour tous $n,\ m \in \mathbb{N}$, $n \neq m$, $(\,p_n \mid p_m\,)_{\mathscr{M}} = 0$.

Notons que la connaissance d'une forme linéaire $\mathscr{M}$ sur $\mathbb{R}[T]$ est équivalente à celle de la suite $\big(\mathscr{M}(t^n)\big)_{n\in\mathbb{N}}$.
Si la forme $\mathscr{M}$ est *strictement positive*, on obtient une suite de polynômes $\mathscr{M}$-orthogonaux à partir de la suite $(t^n)_{n\in\mathbb{N}}$ par le procédé d'orthogonalisation de Gram-Schmidt.

Exercice 1.

Soit $\mathscr{M}$ une forme linéaire sur $\mathbb{R}[T]$. Les conditions suivantes sont équivalentes.

(i) La forme $\mathscr{M}$ est strictement positive.

(ii) Pour tout $p \in \mathbb{R}[T]$, $p \neq 0$, on a $\mathscr{M}(p^2) > 0$.

(iii) II existe une suite $(p_n)_{n\in\mathbb{N}}$ de polynômes $\mathscr{M}$-orthogonaux telle que, pour tout $n \in \mathbb{N}$, $\mathscr{M}(p_n^2) > 0$.

Solution. On a évidemment (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii).
Montrons (iii) $\Rightarrow$ (ii). Les polynômes $\{p_n\}_{n\in\mathbb{N}}$ forment une base de $\mathbb{R}[T]$. Donc tout polynôme p s'écrit $p = \lambda_0 p_0 + \cdots + \lambda_m p_m$ et l'on a $\mathscr{M}(p^2) = \lambda_0^2\mathscr{M}(p_0^2) + \cdots + \lambda_m^2\mathscr{M}(p_m^2)$.
Si $p \neq 0$, l'un des λ n'est pas nul et $\mathscr{M}(p^2) > 0$.
Montrons (ii) $\Rightarrow$ (i). Soit p positif, $p \neq 0$. Le degré de p est nécessairement pair (sinon p tend vers $-\infty$ quand t tend vers $+\infty$ ou vers $-\infty$). Si p est est un polynôme positif de degré $2m$, il existe des polynômes q et r de degré m tels que $p = q^2 + r^2$

(*cf.* une application du théorème de d'Alembert-Gauß page 310 du livre [L1]) et l'on a $\mathscr{M}(p) = \mathscr{M}(q^2) + \mathscr{M}(r^2)$. Si $p \neq 0$, on a $q \neq 0$ ou $r \neq 0$, d'où le résultat.

Favard (1935) et Shohat (1935) ont montré que si la forme $\mathscr{M}$ est *strictement positive* les polynômes orthogonaux correspondants sont définis par un « poids généralisé » (la forme $\mathscr{M}$ est ce que l'on appelle une intégrale de Stieltjes par rapport à une fonction à variation bornée, *cf.* le module *Intégration* du livre [L2]).

2.2 Formules de récurrence et de Darboux-Christoffel

2.2.1 Formules de récurrence

Point Méthode

Il est important de remarquer que, pour tous $q,\ p_1,\ p_2 \in \mathbb{R}[T]$, on a :

$$(\,qp_1 \mid p_2\,)_w = (\,p_1 \mid qp_2\,)_w = \int_\alpha^\beta q(t)p_1(t)p_2(t)w(t)\,\mathrm{d}t.$$

Ainsi, pour $q \in \mathbb{R}[T]$ fixé, l'opérateur linéaire $p \mapsto qp$ de $\mathbb{R}[T]$ dans $\mathbb{R}[T]$ est *autoadjoint*. Les *formules de récurrence pour les polynômes orthogonaux* découlent de cette propriété, quand on prend $q(t) := t$.
Plus généralement, si $\mathscr{M}$ est une forme linéaire sur $\mathbb{R}[T]$, on a :

$$\text{pour tous } q,\ p_1,\ p_2 \in \mathbb{R}[T], \quad (\,qp_1 \mid p_2\,)_{\mathscr{M}} = (\,p_1 \mid qp_2\,)_{\mathscr{M}}.$$

La formule de récurrence donnée ci-dessous est la principale propriété des familles « générales » de polynômes orthogonaux. Elle jouera un rôle fondamental dans l'approche par fractions continues introduite en fin de module.

Théorème 3. Soit $(p_n)_{n\in\mathbb{N}}$ une suite de polynômes orthogonaux associée au poids w sur l'intervalle (α,β) $(p_n(t) := \gamma_n t^n + \cdots)$.

(i) La suite $(p_n)_{n\in\mathbb{N}}$ vérifie, pour $n \geqslant 2$, la relation de récurrence :

$$p_n - (A_n t + B_n)p_{n-1} + C_n p_{n-2} = 0, \tag{1}$$

avec $A_n = \dfrac{\gamma_n}{\gamma_{n-1}}$, $B_n = -\dfrac{\gamma_n}{\gamma_{n-1}}\dfrac{(\,tp_{n-1} \mid p_{n-1}\,)}{\|p_{n-1}\|^2}$, $C_n = \dfrac{\gamma_n\gamma_{n-2}}{\gamma_{n-1}^2}\dfrac{\|p_{n-1}\|^2}{\|p_{n-2}\|^2}$.
Cette relation reste vraie pour $n = 1$, en convenant que $p_{-1} = 0$ et que C_1 est arbitraire.

(ii) Si l'intervalle (α,β) est *symétrique* par rapport à l'origine et si $p_{n-1}^2 w$ est une fonction *paire*, alors $B_n = 0$.

(iii) La suite $(u_n)_{n\in\mathbb{N}}$ de polynômes orthogonaux *unitaires* vérifie, pour $n \geqslant 2$, la relation de récurrence :

$$u_n - (t + B'_n)u_{n-1} + C'_n u_{n-2} = 0, \tag{2}$$

avec $B'_n = -\dfrac{(\,tu_{n-1} \mid u_{n-1}\,)}{\|u_{n-1}\|^2}$, $C'_n = \dfrac{\|u_{n-1}\|^2}{\|u_{n-2}\|^2}$.

Nous verrons que pour plusieurs familles de polynômes orthogonaux classiques (Legendre, Tchebychev, Gegenbauer, Hermite), la condition de symétrie de (ii) est vérifiée pour tout $n \in \mathbb{N}$; ainsi $B_n = 0$, ce qui simplifie la formule de récurrence. Notons que, quel que soit $n \in \mathbb{N}^*$, A_n et C_n sont *non nuls*.

Démonstration. Prouvons l'assertion (iii), on en déduira ensuite la relation générale (1). Soit $n \in \mathbb{N}, n \geqslant 2$, le polynôme $u_n - tu_{n-1}$ est de degré au plus $n-1$, par suite :

$$u_n - tu_{n-1} = \sum_{h=0}^{n-1} \lambda_h u_h$$

et l'on a :

$$\forall h \geqslant 0,\ \|u_h\|^2 \lambda_h = (\, u_n - tu_{n-1} \mid u_h \,) = -(\, tu_{n-1} \mid u_h \,).$$

On a : $(\, tu_{n-1} \mid u_h \,) = (\, u_{n-1} \mid tu_h \,) = 0$ si $\deg\,(tu_h) \leqslant n-2$ (c'est-à-dire si $\deg u_h \leqslant n-3$). Ainsi :

$$u_n - tu_{n-1} = -\frac{(\, tu_{n-1} \mid u_{n-1} \,)}{\|u_{n-1}\|^2} u_{n-1} - \frac{(\, tu_{n-1} \mid u_{n-2} \,)}{\|u_{n-2}\|^2} u_{n-2}.$$

On a aussi $(\, tu_{n-1} \mid u_{n-2} \,) = (\, u_{n-1} \mid tu_{n-2} \,)$ et $tu_{n-2} = u_{n-1} + r$ avec $r \in \mathbb{R}_{n-2}[T]$, donc $(\, u_{n-1} \mid tu_{n-2} \,) = \|u_{n-1}\|^2$ et :

$$u_n - tu_{n-1} = -\frac{(\, tu_{n-1} \mid u_{n-1} \,)}{\|u_{n-1}\|^2} u_{n-1} - \frac{\|u_{n-1}\|^2}{\|u_{n-2}\|^2} u_{n-2}.$$

Revenons au cas général. Pour tout $n \in \mathbb{N}, n \geqslant 2$, on a $p_n = \gamma_n u_n$, il suffit donc, pour montrer (1), de remplacer dans la relation (2) : u_n par p_n/γ_n, u_{n-1} par p_{n-1}/γ_{n-1} et u_{n-2} par p_{n-2}/γ_{n-2}.

Si les conditions de (ii) sont satisfaites, la fonction $t \mapsto tp_{n-1}(t)^2 w(t)$ est *impaire*, donc $(\, tp_{n-1} \mid p_{n-1} \,) = \displaystyle\int_\alpha^\beta tp_{n-1}(t)^2 w(t)\,\mathrm{d}t = 0$ et $B_n = 0$. ■

Remarques.

1. Les C'_n apparaissant dans la relation de récurrence (3) sont *strictement positifs*.
2. On convient de $C'_1 := \mu_0$. La relation de récurrence (2) permet de calculer les $\|u_n\|$ $(n \in \mathbb{N})$ à partir de la suite $(C'_n)_{n\in\mathbb{N}^*}$: $\|u_n\|^2 = C'_{n+1} \cdots C'_1$

On peut aussi écrire la relation de récurrence (1) en utilisant les polynômes orthogonaux orthonormés $(q_n)_{n\in\mathbb{N}}$.

Corollaire 4. Soit $(q_n)_{n\in\mathbb{N}}$ une suite de polynômes orthogonaux *orthonormés* $(q_n = \xi_n t^n + \cdots)$. Elle vérifie, pour $n \geqslant 2$, la relation de récurrence :

$$q_n - (A_n t + B_n) q_{n-1} + C_n q_{n-2} = 0, \tag{3}$$

avec $A_n = \frac{\xi_n}{\xi_{n-1}}$, $B_n = -\frac{\xi_n}{\xi_{n-1}}(\, tq_{n-1} \mid q_{n-1} \,)$, $C_n = \frac{\xi_n \xi_{n-2}}{\xi_{n-1}^2}$.

Cette relation reste vraie pour $n = 1$, en convenant que $q_{-1} = 0$ et que C_1 est arbitraire.

Il est souvent utile d'écrire la relation de récurrence (3) sous la forme équivalente :

$$tq_{n-1} = \tilde{A}_n q_n - \tilde{B}_n q_{n-1} + \tilde{C}_n q_{n-2}, \tag{4}$$

avec $\tilde{A}_n = \frac{\xi_{n-1}}{\xi_n}$, $\tilde{B}_n = -(\,tq_{n-1} \mid q_{n-1}\,)$, $\tilde{C}_n = \frac{\xi_{n-2}}{\xi_{n-1}}$. On constate que $\tilde{C}_n = \tilde{A}_{n-1}$.
Par ailleurs $q_n = \xi_n u_n$, donc :

$$1 = \frac{\|q_{n-1}\|}{\|q_n\|} = \frac{\xi_{n-1}}{\xi_n}\frac{\|u_{n-1}\|}{\|u_n\|} = \frac{\tilde{C}_{n+1}}{\sqrt{C'_{n+1}}} \quad \text{et donc} \quad \tilde{C}_{n+1} = \sqrt{C'_{n+1}},$$

et :

$$\tilde{B}_n = -(\,tq_{n-1} \mid q_{n-1}\,) = -\xi_{n-1}^2(\,tu_{n-1} \mid u_{n-1}\,) = -\frac{(\,tu_{n-1} \mid u_{n-1}\,)}{\|u_{n-1}\|^2} = B'_n.$$

On peut donc passer facilement de la relation de récurrence (3) à la relation de récurrence (4) et inversement, en utilisant les relations : $\tilde{B}_n = B'_n$, $\tilde{C}_n = \sqrt{C'_n}$, $\tilde{A}_n = \tilde{C}_{n+1}$.
On peut associer à l'opérateur de multiplication par t sur l'espace vectoriel $\mathbb{R}[T]$ exprimé dans la base $(q_n)_{n\in\mathbb{N}}$ des polynômes orthogonaux *orthonormés* la *matrice infinie symétrique* tridiagonale :

$$\begin{pmatrix} -\tilde{B}_1 & \tilde{C}_2 & 0 & \dots & 0 & \dots \\ \tilde{C}_2 & -\tilde{B}_2 & \tilde{C}_3 & \dots & 0 & \dots \\ 0 & \tilde{C}_3 & -\tilde{B}_3 & \dots & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\tilde{B}_n & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \end{pmatrix}$$

On l'appelle *matrice de Jacobi*. On peut montrer que les *polynômes caractéristiques* des matrices *symétriques* finies (n, n) extraites (également appelées matrices de Jacobi) :

$$\begin{pmatrix} -\tilde{B}_1 & \tilde{C}_2 & 0 & \dots & 0 \\ \tilde{C}_2 & -\tilde{B}_2 & \tilde{C}_3 & \dots & 0 \\ 0 & \tilde{C}_3 & -\tilde{B}_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\tilde{B}_n \end{pmatrix}$$

sont proportionnels aux $\{q_n\}$ (*cf.* l'exercice III.4.3). Il en résulte que les zéros des polynômes orthogonaux sont *réels*. Nous retrouverons et préciserons ce résultat dans la partie 2.3.

Remarque. On vérifie facilement que les résultats du théorème 3 restent vrais pour des polynômes orthogonaux pour une forme linéaire $\mathscr{M}$ *strictement positive*, à condition de remplacer dans l'énoncé $(\,.\mid .\,) = (\,.\mid .\,)_w$ par $(\,.\mid .\,)_{\mathscr{M}}$ et $\|.\| = \|.\|_w$ par $\|.\|_{\mathscr{M}}$. Les résultats ci-dessus sur les matrices de Jacobi s'étendent également.

On peut associer à un poids w d'autres suites de polynômes, les *polynômes de seconde espèce* définis ci-dessous. Nous les utiliserons de façon essentielle au paragraphe 4.2.2.

Soit $p \in \mathbb{R}[T]$. Il existe un polynôme de deux variables $P \in \mathbb{R}[T,\tau]$ tel que, pour tous $t,\tau \in \mathbb{R}$, on ait $p(t) - p(\tau) = (t-\tau)P(t,\tau)$. Par suite, pour tout $t \in \mathbb{R}$ fixé, on peut définir l'application $\tau \in \mathbb{R} \mapsto P(t,\tau)$. Elle est polynomiale, par abus de langage on la notera $\tau \mapsto \dfrac{p(t)-p(\tau)}{t-\tau}$.

Proposition et définition 5 (Polynômes de seconde espèce).
Soit $(p_n = \gamma_n t^n + \cdots)_{n\in\mathbb{N}}$ une suite de polynômes orthogonaux associés au poids w. Il existe une suite de polynômes $(r_n)_{n\in\mathbb{N}}$ vérifiant la relation de récurrence (1), avec les conditions initiales : $r_0 = 0$ et $r_1(t) = \gamma_1\mu_0$. On a, pour tout $n \in \mathbb{N}$:

$$r_n(t) = \int_\alpha^\beta \frac{p_n(t)-p_n(\tau)}{t-\tau} w(\tau)\,\mathrm{d}\tau$$

et $\deg r_n = n-1$, pour tout $n \in \mathbb{N}^*$.
Les polynômes $\{r_n\}_{n\in\mathbb{N}}$ sont appelés *polynômes de seconde espèce* associés au poids w.

Chez certains auteurs anglo-saxons les polynômes de seconde espèce sont appelés *numerator polynomials*. Nous expliquerons cette terminologie dans la partie 4.1.2.

Démonstration.

Pour $n \in \mathbb{N}^*$, on définit, pour tout $t \in \mathbb{R}$, $R_n(t) := \displaystyle\int_\alpha^\beta \frac{p_n(t)-p_n(\tau)}{t-\tau} w(\tau)d\tau$. On a $R_0(t) = 0$ et $R_1(t) = \gamma_1\mu_0$ puisque $p_1(t) = \gamma_1 t + \gamma_1'$. Montrons que les fonctions $t \in \mathbb{R} \mapsto R_n(t)$ vérifient la relation de récurrence (1). Pour $n \geqslant 2$, on considère :

$$\begin{aligned} &R_n - (A_n t + B_n)R_{n-1} + C_n R_{n-2} \\ &\quad = \int_\alpha^\beta \frac{p_n(t)-p_n(\tau) - (A_n t + B_n)\big(p_{n-1}(t)-p_{n-1}(\tau)\big) + C_n\big(p_{n-2}(t)-p_{n-2}(\tau)\big)}{t-\tau} w(\tau)\,\mathrm{d}\tau. \end{aligned}$$

On a, en utilisant la relation de récurrence (1) pour la suite $(p_n)_{n\in\mathbb{N}}$:

$$\begin{aligned} &p_n(t)-p_n(\tau) - (A_n t + B_n)\big(p_{n-1}(t)-p_{n-1}(\tau)\big) + C_n\big(p_{n-2}(t)-p_{n-2}(\tau)\big) \\ &\quad = -(A_n\tau + B_n)p_{n-1}(\tau) + (A_n t + B_n)p_{n-1}(\tau) = A_n(t-\tau)p_{n-1}(\tau). \end{aligned}$$

On en déduit :

$$R_n - (A_n t + B_n)R_{n-1} + C_n R_{n-2} = A_n \int_\alpha^\beta q_{n-1}(\tau)w(\tau)\,\mathrm{d}\tau = 0.$$

Les suites (R_n) et (r_n) sont donc égales. Le terme de plus haut degré de r_n est $\mu_0\gamma_n t^{n-1}$. ∎

Remarque. On peut reprendre l'énoncé de la proposition-définition 5 pour une suite de polynômes orthogonaux pour une forme linéaire *strictement positive* $\mathscr{M} : \mathbb{R}[T] \to \mathbb{R}$, en définissant les polynômes de seconde espèce $\{r_n\}_{n\in\mathbb{N}}$ par $r_n(t) := \mathscr{M}^{(\tau)}\big(\frac{p_n(t)-p_n(\tau)}{t-\tau}\big)$ (où $\mathscr{M}^{(\tau)}$ signifie que l'on applique, à t fixé, $\mathscr{M}$ à un polynôme en τ).

La preuve est similaire en remplaçant μ_0 par $\mathscr{M}(1)$.

Exercice 2.

Soient deux suites de nombres réels $(B'_n)_{n\in\mathbb{N}^*}$ et $(C'_n)_{n\in\mathbb{N}^*}$. On définit par récurrence une suite $(p_n)_{n\in\mathbb{N}}$ de polynômes *unitaires* et de degré n :

$$p_0 = 1,\ p_1 = t + B'_1 \qquad \text{et} \quad \text{si} \quad n \geqslant 2, \quad p_n = (t + B'_n)p_{n-1} - C'_n p_{n-2}.$$

(i) Montrer qu'il existe une *unique forme linéaire* $\mathscr{M} : \mathbb{R}[T] \to \mathbb{R}$ telle que les polynômes $\{p_n\}_{n\in\mathbb{N}}$ soient $\mathscr{M}$*-orthogonaux* et que $\mathscr{M}(1) = C'_1$.

(ii) Montrer que, si $\mathscr{M}$ est définie comme en (i), pour tout $n \in \mathbb{N}$:

$$\mathscr{M}(p_n^2) = C'_1 \dots C'_{n+1}. \tag{5}$$

et :

$$\mathscr{M}(t^{n+1}p_n) = -C'_1 \dots C'_{n+1}(B'_1 + \cdots + B'_{n+1}). \tag{6}$$

(iii) Montrer que la forme linéaire $\mathscr{M}$ définie en (i) est strictement positive si, et seulement si, pour tout $n \in \mathbb{N}^*$, $C'_n > 0$.

Solution. Les $\{p_n\}_{n\in\mathbb{N}}$ forment une base de $\mathbb{R}[T]$. S'ils sont orthogonaux pour une forme linéaire $\mathscr{M}$, on a $\mathscr{M}(p_n) = \mathscr{M}(p_0 p_n) = 0$ pour tout $n \in \mathbb{N}^*$. La condition $\mathscr{M}(1) = C'_1$ détermine donc $\mathscr{M}$, d'où l'unicité. Notons que $\mathscr{M}$ est triviale si, et seulement si, $C'_1 = 0$.
On montre l'existence en définissant $\mathscr{M}$ sur la base $\{p_n\}_{n\in\mathbb{N}}$ par les conditions :

$$\mathscr{M}(1) = C'_1 \quad \text{et} \quad \forall\, n \in \mathbb{N}^*, \quad \mathscr{M}(p_n) = 0.$$

Montrons que, si $\mathscr{M}$ est ainsi définie, pour tout $n \in \mathbb{N}$, p_n est $\mathscr{M}$-orhogonal à $\mathbb{R}_{n-1}[T]$.
On fait la preuve par récurrence sur un entier h. L'hypothèse de récurrence est :
(H_h) : pour tout $n > h$, $\mathscr{M}(t^h p_n) = 0$.
On a :
(H_0) : pour tout $n > 0$, $\mathscr{M}(p_n) = 0$, qui est vraie par hypothèse.
Supposons (H_h) vraie et montrons (H_{h+1}).
Soit $n > h+1$. D'après (H_h) : $\mathscr{M}(t^h p_{n+1}) = \mathscr{M}(t^h p_n) = \mathscr{M}(t^h p_{n-1}) = 0$. Par hypothèse $p_{n+1} = (t + B'_{n+1})p_n - C'_{n+1}p_{n-1}$, donc :

$$t^h p_{n+1} = t^{h+1}p_n + B'_{n+1}t^h p_n - C'_{n+1}t^h p_{n-1},$$

donc $0 = \mathscr{M}(t^{h+1}p_n) + B'_{n+1}\mathscr{M}(t^h p_n) - C'_{n+1}\mathscr{M}(t^h p_{n-1}) = \mathscr{M}(t^{h+1}p_n)$. Ce qui prouve (H_{h+1}).
Ainsi, pour tous $h, n \in \mathbb{N}$, $h < n$, $\mathscr{M}(t^h p_n) = 0$ et la suite $(p_n)_{n\in\mathbb{N}}$ est $\mathscr{M}$-orthogonale.

Montrons les relations (5). Pour tout $n \in \mathbb{N}^*$, on a :

$$0 = \mathscr{M}(t^{n-1}p_{n+1}) = \mathscr{M}(t^n p_n) + B'_{n+1}\mathscr{M}(t^{n-1}p_n) - C'_{n+1}\mathscr{M}(t^{n-1}p_{n-1}),$$

donc $\mathscr{M}(t^n p_n) = C'_{n+1}\mathscr{M}(t^{n-1}p_{n-1})$. On en déduit $\mathscr{M}(p_n^2) = \mathscr{M}(t^n p_n) = C'_1 \dots C'_{n+1}$.
Montrons les relations (6). Pour tout $n \in \mathbb{N}^*$, on a :

$$0 = \mathscr{M}(t^n p_{n+1}) = \mathscr{M}(t^{n+1}p_n) + B'_{n+1}\mathscr{M}(t^n p_n) - C'_{n+1}\mathscr{M}(t^n p_{n-1}),$$

d'où :

$$\begin{aligned}\mathscr{M}(t^{n+1}p_n) &= -B'_{n+1}\mathscr{M}(t^n p_n) + C'_{n+1}\mathscr{M}(t^n p_{n-1})\\ &= -C'_1 \dots C'_{n+1}B'_{n+1} + C'_{n+1}\mathscr{M}(t^n p_{n-1}).\end{aligned}$$

Ce qui permet de prouver le résultat par récurrence.
Montrons (iii).
Si $\mathscr{M}$ est strictement positive, on a, pour tout $n \in \mathbb{N}$, $C'_1 \dots C'_{n+1} = \mathscr{M}(p_n^2) > 0$, donc les C'_n sont tous strictement positifs.
Si, pour tout $n \in \mathbb{N}^*$, $C'_n > 0$, alors, pour tout $n \in \mathbb{N}$, $\mathscr{M}(p_n^2) > 0$. On en déduit que $\mathscr{M}$ est strictement positive en utilisant l'exercice 1 de la page 536.

Exercice 3.
Soient $(\mathbf{b}_n)_{n\in\mathbb{N}^*}$, $(\mathbf{c}_n)_{n\in\mathbb{N}^*}$ des suites de polynômes de $\mathbb{R}[T]$. On suppose que, pour tout $n \in \mathbb{N}^*$, $\mathbf{c}_n \neq 0$. On considère la relation de récurrence :

$$\forall\, n \geqslant 2, \quad p_n = \mathbf{b}_n p_{n-1} - \mathbf{c}_n p_{n-2}. \tag{7}$$

On définit deux suites de polynômes $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$, par cette relation de récurrence et les conditions initiales $u_0 := 1$, $u_1 := \mathbf{b}_1$ et $v_0 = 0$, $v_1 = \mathbf{c}_1$.
On définit, pour tout $n \in \mathbb{N}^*$, le polynôme $W_n = u_n v_{n-1} - v_n u_{n-1}$. Montrer que, pour tout $n \geqslant 2$, $W_n = \mathbf{c}_n W_{n-1}$. En déduire $W_n = -\mathbf{c}_1 \dots \mathbf{c}_n$ et $W_n \neq 0$.
Solution. On a, en utilisant la relation de récurrence (7), pour tout $n \geqslant 2$:

$$u_n v_{n-1} - v_n u_{n-1} = \begin{vmatrix} u_n & u_{n-1} \\ v_n & v_{n-1} \end{vmatrix} = -\begin{vmatrix} \mathbf{c}_n u_{n-2} & u_{n-1} \\ \mathbf{c}_n v_{n-2} & v_{n-1} \end{vmatrix} = \mathbf{c}_n \begin{vmatrix} u_{n-1} & u_{n-2} \\ v_{n-1} & v_{n-2} \end{vmatrix},$$

donc $W_n = \mathbf{c}_n W_{n-1}$. On en déduit par récurrence $W_n = \mathbf{c}_n \dots \mathbf{c}_2 W_1$.
On a $W_1 = -v_1 u_0 = -\mathbf{c}_1$, d'où $W_n = -\mathbf{c}_n \dots \mathbf{c}_1$. Ainsi W_n est produit de polynômes non nuls, donc non nul.

2.2.2 Formule de Darboux-Christoffel

Soit (q_n) la suite de polynômes orthonormés associée à un poids w. On va étudier, dans ce paragraphe, les fonctions de deux variables définies, sur $\mathbb{R}^2$, pour tout $n \in \mathbb{N}$, par :

$$(t, \tau) \mapsto K_n(t, \tau) := \sum_{k=0}^{n} q_k(t) q_k(\tau).$$

Pour des raisons que nous donnerons plus loin (*cf.* la remarque de la page 547), ces fonctions sont appelés *noyaux*.

Proposition 6 (Formule de Darboux-Christoffel).

(i) $$K_n(t,\tau) = \sum_{k=0}^{n} q_k(t)q_k(\tau) = \frac{\xi_n}{\xi_{n+1}}\Big[\frac{q_n(\tau)q_{n+1}(t) - q_n(t)q_{n+1}(\tau)}{t-\tau}\Big] \text{ (pour } t \neq \tau);$$

(ii) $$K_n(t,t) = \sum_{k=0}^{n} q_k^2(t) = \frac{\xi_n}{\xi_{n+1}}\big[q_n(t)q'_{n+1}(t) - q'_n(t)q_{n+1}(t)\big] \geqslant 0.$$

Démonstration. Montrons (i). On prouve la relation par récurrence en utilisant le corollaire 4 de la page 538. On a $q_0(t) = \xi_0 = 1$ et $q_1(t) = \xi_1 t - \xi_1\mu_0^{-1}\displaystyle\int_\alpha^\beta tw(t)\,\mathrm{d}t$, donc :

$$K_0(t,\tau) = q_0^2(t) = \frac{\xi_0}{\xi_1}\Big[\frac{q_0(\tau)q_1(t) - q_0(t)q_1(\tau)}{t-\tau}\Big].$$

Soit $n > 0$. supposons :

$$K_{n-1}(t,\tau) = \sum_{k=0}^{n-1} q_k(t)q_k(\tau) = \frac{\xi_{n-1}}{\xi_n}\Big[\frac{q_{n-1}(\tau)q_n(t) - q_{n-1}(t)q_n(\tau)}{t-\tau}\Big].$$

On a, pour tout $n > 0$, $q_{n+1} = (A_{n+1}t + B_{n+1})q_n - C_{n+1}q_{n-1}$, donc :

$$\begin{aligned}
&\frac{\xi_n}{\xi_{n+1}}\Big[\frac{q_n(\tau)q_{n+1}(t) - q_n(t)q_{n+1}(\tau)}{t-\tau}\Big] \\
&\qquad = \frac{\xi_n}{\xi_{n+1}}A_{n+1}q_n(t)q_n(\tau) + \frac{\xi_n}{\xi_{n+1}}C_{n+1}\Big[\frac{q_{n-1}(\tau)q_n(t) - q_{n-1}(t)q_n(\tau)}{t-\tau}\Big].
\end{aligned}$$

Mais $\dfrac{\xi_n}{\xi_{n+1}}A_{n+1} = 1$ et $\dfrac{\xi_n}{\xi_{n+1}}C_{n+1} = \dfrac{\xi_n}{\xi_{n+1}}\dfrac{\xi_{n+1}\xi_{n-1}}{\xi_n^2} = \dfrac{\xi_{n-1}}{\xi_n}$, donc :

$$\begin{aligned}
&\frac{\xi_n}{\xi_{n+1}}\Big[\frac{q_n(\tau)q_{n+1}(t) - q_n(t)q_{n+1}(\tau)}{t-\tau}\Big] \\
&\qquad = q_n(t)q_n(\tau) + \frac{\xi_{n-1}}{\xi_n}\Big[\frac{q_{n-1}(\tau)q_n(t) - q_{n-1}(t)q_n(\tau)}{t-\tau}\Big] \\
&\qquad = q_n(t)q_n(\tau) + K_{n-1}(t,\tau) \\
&\qquad = K_n(t,\tau).
\end{aligned}$$

L'assertion (ii) résulte de l'assertion (i) par passage à la limite quand τ tend vers t. ■

L'exercice suivant donne une interprétation du noyau K_n indépendante des polynômes orthogonaux (nous en donnerons plus loin une autre : *cf.* la remarque de la page 547).

Exercice 4.

Soit w un poids sur le segment $[\alpha,\beta]$. Soit $\tau \in [\alpha,\beta]$ un point donné. Montrer que parmi les polynômes u de norme 1 et de degré n, il en existe deux, et deux seulement, maximisant $|u(\tau)|$ et qu'il sont donnés par $u(t) = \pm\frac{K_n(t,\tau)}{\sqrt{K_n(\tau,\tau)}}$.

Solution. Soient q_n, $n \in \mathbb{N}$, les polynômes orthonormés associés à w.

Si u est un polynôme de norme 1 et de degré n, il s'écrit $u(t) = \sum_{k=0}^{n} a_k q_k(t)$.

Comme $\|u\|_w = 1$, on a $\sum_{k=0}^{n} a_k^2 = 1$. D'après l'inégalité de Cauchy-Schwarz (dans $\mathbb{R}^n$), on a :

$$u(\tau)^2 \leqslant \left(\sum_{k=0}^{n} a_k^2\right)\left(\sum_{k=0}^{n} q_k^2(\tau)\right) = \sum_{k=0}^{n} q_k^2(\tau).$$

Le maximum de $u(\tau)^2$ est donc atteint quand l'inégalité est une égalité, c'est-à-dire quand les vecteurs $(a_1, \ldots, a_n)$ et $\big(q_1(\tau), \ldots, q_n(\tau)\big)$, de $\mathbb{R}^n$, sont colinéaires :

$$(a_1, \ldots, a_n) = \lambda\big(q_1(\tau), \ldots, q_n(\tau)\big).$$

Alors $u(t) = \lambda \sum_{k=0}^{n} q_k(\tau) q_k(t) = \lambda K_n(t, \tau)$.

Par ailleurs $1 = \sum_{k=0}^{n} a_k^2 = \lambda^2 \sum_{k=0}^{n} q_k(\tau)^2 = \lambda^2 K_n(\tau, \tau)$. Finalement $u(t) = \pm \dfrac{K_n(t, \tau)}{\sqrt{K_n(\tau, \tau)}}$.

Remarque. Soit $\mathscr{M}$ une forme linéaire *strictement positive* sur $\mathbb{R}[T]$. On vérifie facilement (en reprenant la preuve *mutatis mutandis*) que les formules de Christoffel-Darboux restent vraies pour les polynômes $\mathscr{M}$-orthogonaux orthonormés

2.3 Les zéros des polynômes orthogonaux

Ce paragraphe est consacré à l'étude des propriétés des zéros des polynômes orthogonaux. Soient $(p_n)_{n\in\mathbb{N}}$ (respectivement $(q_n)_{n\in\mathbb{N}}$) les polynômes orthogonaux (respectivement orthonormés) associés à un poids w sur l'intervalle (α, β). On rappelle que les p_n ont pour coefficient dominant 1.Les polynômes p_n et q_n sont proportionnels, ils ont donc les mêmes zéros. Nous ferons, dans toute la suite, un abus de langage et parlerons des zéros « du » $n^{ème}$ polynôme orthogonal. Pour les preuves, on peut donc se limiter au cas des polynômes orthonormés, ce que nous ferons systématiquement. Le polynôme q_n est de degré n, il a donc n zéros (comptés avec leur multiplicité) dans le corps $\mathbb{C}$ des nombres complexes. Le résultat suivant est plus précis.

Théorème 7. Pour tout $n \in \mathbb{N}^*$, le $n^{ème}$ polynôme orthogonal p_n a tous ses zéros *simples*, ces zéros sont *réels* et appartiennent à l'intervalle *ouvert* $]\alpha, \beta[$.

Démonstration. Soit $n > 0$. On considère l'ensemble X des zéros de p_n *de multiplicité impaire* appartenant à l'intervalle ouvert $]\alpha, \beta[$. Supposons que X soit vide. Alors p_n ne change pas de signe sur $]\alpha, \beta[$. On peut supposer $p_n \geqslant 0$ (on le remplace sinon par $-p_n$) et lui appliquer le lemme 1 de la page 533, puisque $0 = (\,1 \mid p_n\,) = \displaystyle\int_\alpha^\beta p_n(t) w(t)\,\mathrm{d}t = 0$, on a donc $p_n = 0$, ce qui est impossible. Ainsi $X \neq \varnothing$.

Soit $k \geqslant 1$ le nombre d'éléments de X. Si $k = n$, le théorème est vrai (dans ce cas X contient tous les zéros complexes de p_n et ces zéros sont nécessairement simples), sinon $k < n$. Montrons que ce dernier cas est impossible. Notons $t_1, \ldots, t_k$, les éléments de X et posons $p(t) = (t - t_1) \cdots (t - t_k)$. Le polynôme p est de degré $k < n$, donc $0 = (p_n \mid p) = \displaystyle\int_\alpha^\beta (p_n p)(t) w(t)\,\mathrm{d}t$. Par ailleurs tous les zéros du polynôme $p_n p$ appartenant à $]\alpha, \beta[$ sont *de multiplicité paire*, donc ce polynôme ne change pas de signe sur cet intervalle. Supposons que $p_n p \geqslant 0$ (on le remplace sinon par $-p_n p$), on lui applique le lemme 1 de la page 533, et l'on obtient $p_n p = 0$, et $p = 0$ ou $p_n = 0$, ce qui est impossible. ∎

Le résultat suivant concerne la comparaison entre les zéros de deux polynômes orthogonaux successifs p_n et p_{n+1}. Il est illustré par la figure III.4.1 de la page suivante.

Théorème 8. Soit $n \in \mathbb{N}^*$. Les polynômes orthogonaux p_n et p_{n+1} n'ont pas de zéro commun, de plus les zéros de p_n et de p_{n+1} sont *alternés* sur $]\alpha, \beta[$.

Démonstration. On peut supposer les polynômes orthonormés : $p_n = q_n$. Si q_n et q_{n+1} avaient un zéro commun t_0, ce serait aussi, d'après la formule de récurrence (*cf.* le théorème 4 de la page 538), un zéro de q_{n-1}. En répétant (si nécessaire) cet argument, on obtiendrait $0 = q_{n+1}(t_0) = q_n(t_0) = q_{n-1}(t_0) = \cdots = q_0(t_0) = \xi_0$, d'où une contradiction, puisque $\xi_0 \neq 0$.

Soient t_r et t_{r+1} deux zéros successifs de q_n. D'après la proposition 6 de la page 543, on a :

$$K_n(t_r, t_r) = \frac{\xi_n}{\xi_{n+1}} \left[-q_n'(t_r) q_{n+1}(t_r)\right] = \sum_{k=0}^{n} q_k^2(t_r) \geqslant 0,$$

donc $q_n'(t_r) q_{n+1}(t_r) \leqslant 0$, et, de même $q_n'(t_{r+1}) q_{n+1}(t_{r+1}) \leqslant 0$. Puisque t_r et t_{r+1} sont deux zéros successifs de q_n et que ces zéros sont simples, $q_n'(t_r)$ et $q_n'(t_{r+1})$ sont non nuls et *de signes opposés*, donc $q_{n+1}(t_r)$ et $q_{n+1}(t_{r+1})$, non nuls, sont aussi *de signes opposés*. On en déduit que le polynôme q_{n+1} s'annule au moins une fois sur l'intervalle $]t_r, t_{r+1}[$. On montrerait de même que q_n s'annule au moins une fois entre deux zéros successifs de q_{n+1}. Le résultat s'en déduit. ∎

On peut reprendre (avec quelques modifications simples) les preuves du théorème et du théorème 8 (en utilisant pour le second la remarque de la page 544) dans le cadre des polynômes orthogonaux pour une forme linéaire *strictement positive* $\mathscr{M}$. On obtient le résultat suivant.

Théorème 9. Soient $\mathscr{M}$ une forme linéaire *strictement positive* sur $\mathbb{R}[T]$ et $(p_n)_{n\in\mathbb{N}}$ une suite de polynômes $\mathscr{M}$-orthogonaux. Alors, pour tout $n \in \mathbb{N}^*$:

(i) le $n^{\text{ème}}$ polynôme orthogonal p_n a tous ses zéros *simples et réels* ;

(ii) les polynômes orthogonaux p_n et p_{n+1} n'ont pas de zéro commun, de plus les zéros de p_n et de p_{n+1} sont *alternés* sur $\mathbb{R}$.

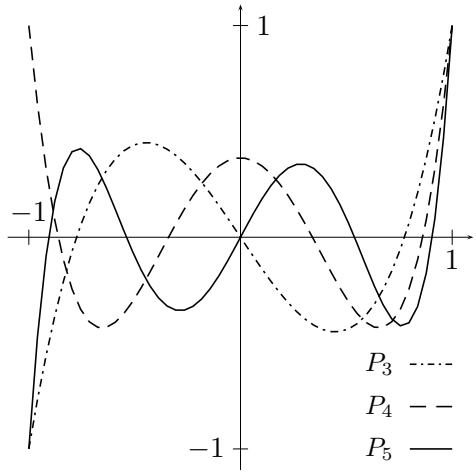


FIGURE III.4.1 – Graphes des polynômes de Legendre P_3, P_4, P_5

2.4 Approximation et formules de quadrature

Soit w un poids sur l'intervalle (α, β). Dans ce paragraphe et le suivant, *on supposera que* $w > 0$ *sur l'intervalle ouvert* $]\alpha, \beta[$.

Si f est une fonction continue sur (α, β), on pose $\|f\|_w := \Big[\int_\alpha^\beta f^2(t)w(t)\,\mathrm{d}t\Big]^{1/2}$; c'est un nombre réel positif ou $+\infty$. Soit $L_w\big((\alpha, \beta)\big)$ le sous-ensemble de $\mathcal{C}(\alpha, \beta)$ défini par $L_w\big((\alpha, \beta)\big) := \{f \mid \|f\|_w < +\infty\}$; c'est un espace vectoriel et la forme bilinéaire :

$$(f, g) \mapsto (\,f \mid g\,) := \int_\alpha^\beta f(t)g(t)w(t)\,\mathrm{d}t$$

définit un produit scalaire sur L_w. La vérification est laissée au lecteur, qui procédera comme dans le cas $w = 1$, en notant que $(\,f \mid f\,) = \int_\alpha^\beta f(t)^2 w(t)\,\mathrm{d}t = 0$ entraîne $f = 0$ (w est continue et strictement positive sur $]\alpha, \beta[$).

On remarque que $\mathbb{R}[T] \subset L_w\big((\alpha, \beta)\big)$, puisque tous les moments de w sont finis.

Proposition 10. Soient w un poids sur l'intervalle (α, β), $w > 0$, et (q_n) la suite de polynômes orthonormés associés. Soit f une fonction continue sur (α, β) telle que $\|f\|_w < +\infty$. Soit $n \in \mathbb{N}$.

(i) La fonction $\mathbb{R}_n[T] \to \mathbb{R}$ définie par $p \mapsto \|f - p\|_w$ admet un unique minimum en $r_n \in \mathbb{R}_n[T]$. On a $r_n = \sum_{k=0}^{n} \lambda_k q_k$, avec :

$$\lambda_k = (\,f \mid q_k\,) = \int_\alpha^\beta f(t)q_k(t)w(t)\,\mathrm{d}t.$$

(ii) La fonction $f - r_n$ est nulle ou change de signe au moins $n + 1$ fois sur l'intervalle $]\alpha, \beta[$.

Le polynôme r_n est appelé le polynôme de meilleure approximation quadratique (avec le poids w) de f à l'ordre n.

Démonstration. Le polynôme r_n est la projection orthogonale de $f \in L_w$ sur le sous-espace vectoriel de dimension finie $\mathbb{R}_n[T]$ et l'assertion (i) résulte du théorème 14 du module *Espaces préhilbertiens* du livre [L2].

Supposons $f \neq r_n$ et que la fonction $f - r_n$ change de signe au plus n fois sur l'intervalle $]\alpha, \beta[$. Notons $(\tau_1, \ldots, \tau_m)$ les points où $f - r_n$ change de signe (le lecteur vérifiera qu'il y en a en utilisant $(f - r_n \mid 1) = 0$). Soit $r(t) := (t - \tau_1) \cdots (t - \tau_m)$. On a $r \in \mathbb{R}_n[T]$, on peut donc écrire $r(t) = \sum_{k=0}^{n} \nu_k q_k$. On a $(f - r_n \mid q_h) = 0$, pour tout $h \in [\![1, n]\!]$, donc $(f - r_n \mid r) = 0$. La fonction continue $(f - r_n)r$ ne change pas de signe sur l'intervalle $]\alpha, \beta[$, comme $\displaystyle\int_\alpha^\beta (f(t) - r_n(t))r(t)w(t)\,\mathrm{d}t = 0$, et $w > 0$, elle est donc identiquement nulle sur cet intervalle et l'on a une contradiction. Cela prouve l'assertion (ii). ∎

Remarque. On peut écrire :

$$r_n(t) = \sum_{k=0}^{n} \lambda_k q_k(t) = \sum_{k=0}^{n} \int_\alpha^\beta f(\tau) q_k(t) q_k(\tau) w(\tau)\,\mathrm{d}\tau = \int_\alpha^\beta K_n(t,\tau) f(\tau)\,\mathrm{d}\tau,$$

avec $K_n(t,\tau) := \sum_{k=0}^{n} q_k(t) q_k(\tau)$. La projection sur le sous-espace $\mathbb{R}_n[T]$ est un exemple d'opérateur intégral de noyau $K_n(t,\tau)$.

Si $f = p \in \mathbb{R}_n[T]$, $r_n = p$ et l'on a $p(t) = \displaystyle\int_\alpha^\beta K_n(t,\tau) p(\tau) w(\tau)\,\mathrm{d}\tau$, ce qui justifie la terminologie « noyau reproduisant », parfois employée pour K.

On a en particulier $\displaystyle\int_\alpha^\beta K_n(t,\tau) w(\tau)\,\mathrm{d}\tau = 1$.

2.4.1 Propriétés de densité

Dans tout ce paragraphe, on désigne par (α, β) un intervalle *borné* (fermé, ouvert, ou semi-ouvert) et par w un poids sur (α, β) (on notera que w est continu sur (α, β), mais n'est pas nécessairement borné sur cet intervalle). L'hypothèse : *borné*, pour l'intervalle, est essentielle, les résultats que nous allons montrer pouvant être faux dans le cas contraire. Nous traiterons plus loin les cas des polynômes de Laguerre et d'Hermite, qui correspondent à des intervalles non bornés (*cf.* 3.7).

On reprend les notations du paragraphe précédent : $L_w\big((\alpha, \beta)\big)$, $(\ \mid\ )_w$, $\|\ \|_w$. On suppose toujours $w > 0$ sur $]\alpha, \beta[$.

Rappelons le résultat suivant (*cf.* le module *Suites et séries de fonctions* du livre [L2]).

Théorème 11 (Weierstrass). Soit $[\alpha, \beta]$ un segment. Les fonctions polynomiales forment un sous–espace vectoriel *dense* dans $\mathcal{C}([\alpha, \beta])$ pour la norme $\|\ \|_\infty$.

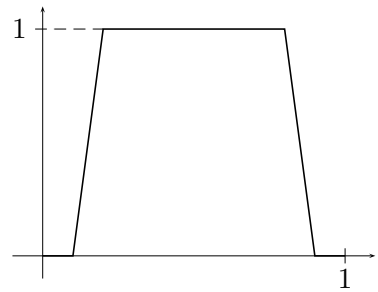
Soit $f \in \mathcal{C}([\alpha,\beta])$. On a $\|f\|_w^2 = \displaystyle\int_\alpha^\beta f^2(t)w(t)\,\mathrm{d}t \leqslant \|f\|_\infty^2 \int_\alpha^\beta w(t)\,\mathrm{d}t = \mu_0\|f\|_\infty^2$. Pour tout $\varepsilon > 0$, il existe un polynôme P, tel que $\|f-P\|_\infty \leqslant \sqrt{\frac{\varepsilon}{\mu_0}}\cdot$ On a alors $\|f-P\|_w \leqslant \varepsilon$. On a donc le résultat suivant.

Corollaire 12. Les fonctions polynomiales forment un sous–espace vectoriel *dense* dans $\mathcal{C}([\alpha,\beta])$ pour la norme $\|\ \|_w$.

Lemme 13. Soit $[\alpha,\beta]$ un segment. Soit $f \in L_w\big(]\alpha,\beta[\big)$. Pour tout $\varepsilon > 0$, il existe une fonction g, continue sur $[\alpha,\beta]$, $g \in L_w\big(]\alpha,\beta[\big)$, telle que $\|f-g\|_w \leqslant \varepsilon$.

Démonstration. Par affinité, on se ramène au cas où $]\alpha,\beta[\ := \ ]0,1[$.
Soit $n \in \mathbb{N}$, avec $n > 2$. Soit h_n la fonction continue et affine par morceaux sur $[0,1]$ définie par :

$$h_n(t) = \begin{cases} 1 & \text{si } t \in \left[\frac{1}{n}, 1-\frac{1}{n}\right] \\ 0 & \text{si } t \in \left[0, \frac{1}{2n}\right] \cup \left[1-\frac{1}{2n}, 1\right] \\ 2nt-1 & \text{si } t \in \left[\frac{1}{2n}, \frac{1}{n}\right] \\ 2n-1-2nt & \text{si } t \in \left[1-\frac{1}{n}, 1-\frac{2}{n}\right]. \end{cases}$$



L'application $g_n := h_n f$, prolongée par 0 en 0 et 1, est alors continue sur le *segment* $[0,1]$ (elle l'est sur $]0,1[$ et elle est nulle sur $\left[0,\frac{1}{2n}\right] \cup \left[1-\frac{1}{2n},1\right]$). On a $0 \leqslant h_n \leqslant 1$, donc $0 \leqslant 1-h_n \leqslant 1$.
On a :

$$\begin{aligned}\int_0^1 (f(t)-g_n(t))^2 w(t)\,\mathrm{d}t &= \int_0^{1/2n} f^2(t)w(t)\,\mathrm{d}t + \int_{1/2n}^{1/n} f^2(t)(1-h_n(t))^2 w(t)\,\mathrm{d}t \\ &\quad + \int_{1-1/n}^{1-1/2n} f^2(t)(1-h_n(t))^2 w(t)\,\mathrm{d}t + \int_{1-1/2n}^{1} f^2(t)w(t)\,\mathrm{d}t \\ &\leqslant \int_0^{1/n} f^2(t)w(t)\,\mathrm{d}t + \int_{1-1/n}^{1} f^2(t)w(t)\,\mathrm{d}t.\end{aligned}$$

Soit $\varepsilon > 0$. D'après la définition de l'intégrabiité, il existe $n \in \mathbb{N}$, $n > 2$, tel que :

$$\int_0^{1/n} f^2(t)w(t)\,\mathrm{d}t + \int_{1-1/n}^{1} f^2(t)w(t)\,\mathrm{d}t < \varepsilon.$$

La fonction $g := g_n$ répond donc à la question. ∎

En utilisant ce lemme et le corollaire 12 de la page précédente, on obtient le résultat suivant.

Proposition 14. Soit $[\alpha, \beta]$ un *segment*. Les fonctions polynomiales forment un sous–espace vectoriel *dense* dans $L_w\big(]\alpha, \beta[\big)$.

En utilisant le théorème 86 de la page 537 du livre [L2], on en déduit le résultat suivant.

Théorème 15. Soit w un poids sur le *segment* $[\alpha, \beta]$, avec $w > 0$ sur $]\alpha, \beta[$.
Soit $(q_n)_{n\in\mathbb{N}}$ la suite des polynômes orthonormés associés. Alors, pour toute fonction $f \in L_w\big((\alpha, \beta)\big)$ la suite de polynômes $\sum\limits_{n=0}^{n} (\, f \mid q_k \,)_w \, q_k$ converge vers f au sens de la norme $\| \ \|_w$ et l'on a l'égalité de Parseval :

$$\|f\|_w^2 = \sum_{n=0}^{+\infty} (\, f \mid q_k \,)^2.$$

2.4.2 Méthode de quadrature de Gauß

Nous allons construire, dans ce paragraphe, des « approximations » d'une intégrale en utilisant les zéros des polynômes orthogonaux. Cette théorie est due à Gauß. Cette question est présentée, du point de vue de l'analyse numérique (calcul des intégrales ou quadrature), dans le module *Méthodes numériques* du livre [L2] auquel nous renvoyons le lecteur. Nous étudierons ici *ses aspects plus théoriques* : les formules de quadrature de Gauß permettent d'obtenir une intégrale $\displaystyle\int_\alpha^\beta g(t)\,\mathrm{d}t = \int_\alpha^\beta f(t)w(t)\,\mathrm{d}t$ par un passage à la limite, sur n, de combinaisons linéaires de valeurs de f en certains points. De façon un peu plus « sophistiquée », cela correspond à l'approximation de la mesure *positive* de densité w : $w(t)\,\mathrm{d}t$ par une suite de mesures *positives* à support fini (un point de vue essentiel pour les compléments en fin de module).
On *choisit* un poids w strictement positif sur $]\alpha, \beta[$ et l'on écrit :

$$\int_\alpha^\beta g(t)\,\mathrm{d}t = \int_\alpha^\beta \frac{g(t)}{w(t)} w(t)\,\mathrm{d}t.$$

On est donc ramené au calcul approché d'une intégrale de la forme : $\displaystyle\int_\alpha^\beta f(t)w(t)\,\mathrm{d}t$. On cherche une formule du type $\displaystyle\int_\alpha^\beta f(t)w(t)\,\mathrm{d}t \approx \sum_{k=1}^{n} f(t_k)\lambda_{k,n}$, où les t_k sont les zéros du $n^{\text{ème}}$ polynôme orthogonal associé à w (les « nœuds ») et les $\lambda_{k,n}$ sont des nombres « universels », ne dépendant pas de f (seulement de w). L'idée est *d'imposer* que la formule soit *exacte*, quand f est une fonction polynomiale de degré strictement inférieur à $2n$.
Soit (p_n) une suite de polynômes orthogonaux associés à un poids w. Les zéros du $n^{\text{ème}}$ polynôme orthogonal p_n ne dépendent pas du choix de la suite (i.e. de la normalisation). On rappelle que ces zéros sont simples. On les notera $t_{1,n}, \ldots, t_{n,n}$ ou, plus simplement, $t_1, \ldots, t_n$, s'il n'y a pas d'ambiguïté possible.

Proposition et définition 16 (Formule de quadrature de Gauß).
Soit $n \in \mathbb{N}^*$. Soient $t_1, \ldots, t_n$ les zéros du $n^{ème}$ polynôme orthogonal p_n associé au poids w. Il existe une suite finie unique de nombres réels $\lambda_{1,n}, \ldots, \lambda_{n,n}$ tels que $\displaystyle\int_\alpha^\beta p(t)w(t)\,\mathrm{d}t = \sum_{k=1}^n p(t_k)\lambda_{k,n}$ pour toute fonction polynomiale p de degré strictement inférieur à $2n$. Les nombres $\lambda_{k,n}$ sont *strictement positifs*. Ils sont appelés *nombres de Christoffel*.

Démonstration. On utilise les polynômes d'interpolation de Lagrange associés à la suite de zéros $t_1, \ldots, t_n$. On pose $L_{k,n}(t) = \dfrac{\prod\limits_{h\neq k}(t-t_h)}{\prod\limits_{h\neq k}(t_k-t_h)} = \dfrac{p_n(t)}{(t-t_k)p_n'(t_k)}$.

Soit p un polynôme de degré strictement inférieur à $2n$. Soit $r_n(t) := \sum\limits_{k=1}^n p(t_k)L_{k,n}(t)$. C'est un polynôme de degré au plus $n-1$, coïncidant avec p aux points $t_1, \ldots, t_n$: $r_n(t_h) = p(t_h)$, pour tout $h \in [\![1, n]\!]$.
Le polynôme $p - r_n$ s'annule donc aux points $t_1, \ldots, t_n$; il est par conséquent divisible par p_n. Ainsi, $s_n := \frac{p-r_n}{p_n}$ est un polynôme de degré strictement inférieur à n. On écrit : $p = r_n + s_n p_n$, et, par intégration, l'on obtient :

$$\int_\alpha^\beta p(t)w(t)\,\mathrm{d}t = \sum_{k=1}^n p(t_k)\int_\alpha^\beta L_{k,n}(t)w(t)\,\mathrm{d}t + \int_\alpha^\beta s_n(t)p_n(t)w(t)\,\mathrm{d}t.$$

On a $\displaystyle\int_\alpha^\beta s_n(t)p_n(t)w(t)\,\mathrm{d}t = 0$, puisque s_n est de degré au plus $n-1$, donc :

$$\int_\alpha^\beta p(t)w(t)\,\mathrm{d}t = \sum_{k=1}^n p(t_k)\lambda_{k,n}$$

en posant $\lambda_{k,n} := \displaystyle\int_\alpha^\beta L_{k,n}(t)w(t)\,\mathrm{d}t$.
On a en particulier :

$$\forall\, h \in [\![0, 2n-1]\!], \quad \mu_h = \int_\alpha^\beta t^h w(t)\,\mathrm{d}t = \sum_{k=1}^n t_k^h\,\lambda_{k,n}, \tag{8}$$

Les n premières équations ($h \in [\![0, n-1]\!]$) permettent de calculer les nombres $\lambda_{k,n}$ ($k \in [\![0, n-1]\!]$) en fonction des moments μ_h ($h \in [\![0, n-1]\!]$) et des zéros $t_1, \ldots, t_n$, car le déterminant : $\begin{vmatrix} 1 & 1 & \ldots & 1 \\ t_1 & t_2 & \ldots & t_n \\ \vdots & \vdots & \vdots & \vdots \\ t_1^{n-1} & t_2^{n-1} & \ldots & t_n^{n-1} \end{vmatrix}$ est un déterminant de Vandermonde, donc non nul (les zéros sont simples). Les nombres de Christoffel $\lambda_{k,n}$ sont donc déterminés de façon unique par le poids w.

Il reste à montrer que les nombres de Christoffel sont *strictement positifs*. Cela résultera de la proposition 18 ci-dessous. ■

Corollaire 17. Soient w un poids et $(p_n)_{n\in\mathbb{N}}$ une suite de polynômes orthogonaux pour w. On note $(r_n)_{n\in\mathbb{N}}$ les polynômes de seconde espèce associés. On note, pour tout $n \in \mathbb{N}^*$, $t_{1,n}, \ldots, t_{n,n}$ les zéros de p_n et $\lambda_{1,n}, \ldots, \lambda_{n,n}$ les nombres de Christoffel. On a alors l'égalité de fractions rationnelles :

$$\forall\, n \in \mathbb{N}^*, \quad \frac{r_n}{p_n} = \sum_{k=1}^{n} \frac{\lambda_{k,n}}{t - t_{k,n}}. \tag{9}$$

Démonstration. On montre la formule (9) en appliquant la formule de quadrature de Gauß au polynôme en τ (à t fixé) : $p(\tau) := \frac{p_n(\tau)-p_n(t)}{(\tau-t)}$.

On a $\displaystyle\int_\alpha^\beta p(\tau)w(\tau)\,\mathrm{d}\tau = \sum_{k=1}^{n} p(t_{k,n})\lambda_{k,n}$. Par ailleurs :

$$\int_\alpha^\beta p(\tau)w(\tau)\,\mathrm{d}\tau = \int_\alpha^\beta \frac{p_n(\tau) - p_n(t)}{\tau - t} w(\tau)\,\mathrm{d}\tau = r_n(t),$$

et $p(t_{k,n}) = \frac{p_n(t)}{t-t_{k,n}}$, donc :

$$p_n(t) = \sum_{k=1}^{n} \lambda_{k,n} p(t_{k,n}) = \sum_{k=1}^{n} \lambda_{k,n} \frac{p_n(t)}{t - t_{k,n}}.$$

Le résultat s'en déduit (on le montre d'abord pour les fonctions polynômes pour $t \neq t_{k,n}$). ■

Proposition 18. Désignons, pour $n \in \mathbb{N}^*$, par $t_{k,n}$ les zéros des polynômes orthogonaux associés à un poids w et par $\lambda_{k,n}$ les nombres de Christoffel correspondants.
(i) On a :

$$\begin{aligned}
\lambda_{k,n} &= \int_\alpha^\beta \frac{q_n(t)}{(t - t_{k,n})q_n'(t_{k,n})} w(t)\,\mathrm{d}t \\
&= -\frac{\xi_{n+1}}{\xi_n} \frac{1}{q_n'(t_{k,n})q_{n+1}(t_{k,n})} = \frac{1}{\sum\limits_{j=0}^{n-1} q_j^2(t_{k,n})} \\
&= \int_\alpha^\beta \Big[\frac{q_n(t)}{(t - t_{k,n})q_n'(t_{k,n})}\Big]^2 w(t)\,\mathrm{d}t.
\end{aligned}$$

(ii) Les nombres de Christoffel $\lambda_{k,n}$ sont strictement positifs.

Démonstration. Par définition

$$\lambda_{k,n} = \int_\alpha^\beta L_{k,n}(t)w(t)\,\mathrm{d}t = \int_\alpha^\beta \frac{q_n(t)}{(t - t_{k,n})q_n'(t_{k,n})} w(t)\,\mathrm{d}t.$$

On a $q_n(t_{k,n}) = 0$, donc, en appliquant la formule de Darboux-Christoffel (*cf.* proposition 6 de la page 543) :

$$K_n(t, t_{k,n}) := \sum_{k=0}^{n-1} q_k(t) q_k(t_{k,n}) = \frac{\xi_n}{\xi_{n+1}} \Big[-\frac{q_n(t) q_{n+1}(t_{k,n})}{t - t_{k,n}} \Big].$$

Ainsi $\dfrac{q_n(t)}{t - t_{k,n}} = -\dfrac{\xi_{n+1} K_n(t, t_{k,n})}{\xi_n q_{n+1}(t_{k,n})}$, d'où :

$$\lambda_{k,n} = -\frac{\xi_{n+1}}{\xi_n q_n'(t_{k,n}) q_{n+1}(t_{k,n})} \int_\alpha^\beta K_n(t, t_{k,n}) w(t) \, \mathrm{d}t.$$

Mais $\displaystyle\int_\alpha^\beta K_n(t, t_{k,n}) w(t) \, \mathrm{d}t = 1$ (*cf.* page 547), donc $\lambda_{k,n} = -\dfrac{\xi_{n+1}}{\xi_n q_n'(t_{k,n}) q_{n+1}(t_{k,n})}$.
En appliquant la formule de Darboux-Christoffel, avec $t = t_{k,n}$, on obtient :

$$K_n(t_{k,n}, t_{k,n}) = \sum_{j=0}^{n-1} q_j(t_{k,n})^2 = -\frac{\xi_n}{\xi_{n+1}} \big[q_n'(t_{k,n}) q_{n+1}(t_{k,n}) \big].$$

On a $K_n(t_{k,n}, t_{k,n}) \neq 0$, donc $\lambda_{k,n} = \dfrac{1}{\sum_{j=0}^{n-1} q_j^2(t_{k,n})}$. Par suite $\lambda_{k,n} > 0$.

Pour établir la dernière relation, on utilise la formule de quadrature de Gauß. On l'applique, pour n et $k \in [\![1, n]\!]$ fixés, au polynôme $p := L_{k,n}^2$, qui est de degré $2n - 2$. On a $p(t_{h,n}) = 0$, si $h \neq k$ et $p(t_{k,n}) = 1$, donc :

$$\int_\alpha^\beta \Big[\frac{q_n(t)}{(t - t_{k,n}) q_n'(t_{k,n})} \Big]^2 w(t) \, \mathrm{d}t = \int_\alpha^\beta p(t) w(t) \, \mathrm{d}t = \sum_{h=1}^n p(t_{h,n}) \lambda_{h,n} = \lambda_{k,n}. \qquad \blacksquare$$

La positivité des nombres de Christoffel jouera un rôle important dans la suite.

Si maintenant f est une fonction continue sur l'intervalle (α, β), telle que l'intégrale $\displaystyle\int_\alpha^\beta f(t) w(t) \, \mathrm{d}t$ (éventuellement généralisée) existe, on peut décider de prendre $\sum_{k=1}^n f(t_k) \lambda_{k,n}$ comme « valeur numérique approchée » de cette intégrale. C'est la méthode de quadrature de Gauß. Les aspects numériques de ce problème sont traités dans le module *Méthodes numériques* du livre [L2]. Nous allons donner ici un sens théorique à cette approximation, en passant à la limite quand n tend vers $+\infty$.

Proposition 19 (Méthode de quadrature de Gauß).
Soient $[\alpha, \beta]$ un segment et w un poids sur $]\alpha, \beta[$. Pour tout $f \in \mathcal{C}([\alpha, \beta])$, la suite de nombres réels $\sum_{k=1}^n f(t_k) \lambda_{k,n}$ converge vers $\displaystyle\int_\alpha^\beta f(t) w(t) \, \mathrm{d}t$.

Remarquons que le résultat de la proposition ci-dessus reste vrai pour les fonctions f, à valeurs *complexes*, continues sur $[\alpha, \beta]$ (en considérant les parties réelle et imaginaire de f).

Démonstration. On remarque que, pour tout $n \in \mathbb{N}^*$, $\mu_0 = \displaystyle\int_\alpha^\beta w(t)\,\mathrm{d}t = \sum_{k=1}^n \lambda_{k,n}$.

Si $p \in \mathbb{R}[T]$, on peut écrire :

$$\sum_{k=1}^n f(t_k)\lambda_{k,n} - \int_\alpha^\beta f(t)w(t)\,\mathrm{d}t = \sum_{k=1}^n \big(f(t_k) - p(t_k)\big)\lambda_{k,n}$$
$$+ \sum_{k=1}^n p(t_k)\lambda_{k,n} - \int_\alpha^\beta p(t)w(t)\,\mathrm{d}t + \int_\alpha^\beta \big(p(t) - f(t)\big)w(t)\,\mathrm{d}t.$$

Soit $\varepsilon > 0$. Il existe $p \in \mathbb{R}[T]$, tel que $\|f - p\|_\infty < \frac{\varepsilon}{2\mu_0}$ (*cf.* le théorème 11 de la page 547). On a alors :

$$\Big|\int_\alpha^\beta (p - f)(t)w(t)\,\mathrm{d}t\Big| \leqslant \|f - p\|_\infty \mu_0 < \varepsilon/2.$$

De même (*cf.* (8)) :

$$\Big|\sum_{k=1}^n (f - p)(t_k)\lambda_{k,n}\Big| \leqslant \|f - p\|_\infty \mu_0 < \varepsilon/2.$$

Soit N le degré de p. Si $n > N$, on a $\sum\limits_{k=1}^n p(t_k)\lambda_{k,n} - \displaystyle\int_\alpha^\beta p(t)w(t)\,\mathrm{d}t = 0$, donc

$$\left|\sum_{k=1}^n f(t_k)\lambda_{k,n} - \int_\alpha^\beta f(t)w(t)\,\mathrm{d}t\right| < 2\varepsilon/2 = \varepsilon.$$

■

Nous allons reformuler les considérations précédentes en utilisant la notion de *mesure*. Nous nous limiterons ici au *strict minimum* en ne considérant que des mesures à support fini et des mesures définies par un poids w sur un intervalle ouvert borné. Nous *définissons* dans ce module les mesures comme formes linéaires continues sur des espaces de Banach de fonctions continues. Cette définition est « équivalente » à celle de mesure borélienne introduite dans les modules d'intégration (section 3.2 de la page 321) et de probabilité (page 668) (*cf.* [MPA] ou [Ru]). Nous n'utiliserons pas ce résultat.

Définition 3. Soient $[\alpha, \beta]$ un segment et $\mathcal{C}([\alpha, \beta])$ le $\mathbb{R}$-espace vectoriel des fonctions continues à valeurs réelles sur $[\alpha, \beta]$. On le munit de la norme $\|f\|_\infty = \sup\limits_{t \in [\alpha,\beta]} |f(t)|$. Une *mesure* sur $[\alpha, \beta]$ est une *forme linéaire continue* μ sur l'espace de Banach $\mathcal{C}([\alpha, \beta])$. Cette mesure est dite *positive* si $\mu(f) \geqslant 0$ pour toute fonction f positive sur $[\alpha, \beta]$. Les *moments* de la mesure sont les nombres réels $\mu_n = \mu(t^n)$, pour tout $n \in \mathbb{N}$.

On peut aussi définir une notion de mesure sur $]\alpha, \beta[$ en remplaçant l'espace des fonctions continues par celui des fonctions continues *bornées*. Par exemple, une fonction continue $g \in \mathcal{C}(]\alpha, \beta[)$, absolument intégrable sur $]\alpha, \beta[$ (au sens de Riemann généralisé), permet de définir une mesure :

$\mu_g : f \mapsto \int_\alpha^\beta f(t)g(t)\,\mathrm{d}t$. Cette forme linéaire, sur l'espace des fonctions continues bornées sur $]\alpha,\beta[$, est continue car $|\mu_g(f)| \leqslant \|f\|_\infty \int_\alpha^\beta |g(t)|\,\mathrm{d}t$. Si g est positive, cette mesure est positive.
Toutes les mesures ne sont pas définissables de cette façon (*cf.* l'exercice III.4.6).
Supposons $]\alpha,\beta[= \mathbb{R}$. Soit $c \in \mathbb{R}$. Si f est bornée sur $\mathbb{R}$, on a $|f(c)| \leqslant \|f\|_\infty$, l'application $f \mapsto f(c)$ est donc une forme linéaire continue.
On définit aussi une notion de *mesure à support fini*.

Définition 4. La masse de Dirac en $c \in \mathbb{R}$ est la forme linéaire $\delta_c : \mathcal{C}(\mathbb{R}) \to \mathbb{R}$ définie par $\delta_c(f) := f(c)$. Par définition, une mesure à support fini sur $\mathbb{R}$ est une forme linéaire sur $\mathcal{C}(\mathbb{R})$ définie par une combinaison linéaire *réelle* de masses de Dirac : $\mu := \sum_{i\in[\![1,k]\!]} \lambda_i \delta_{c_i}$.
Cette mesure est positive si tous les λ_i ($i \in [\![1,k]\!]$) sont positifs. Si les λ_i ($i \in [\![1,k]\!]$) sont tous non nuls, le support de la mesure est par définition l'ensemble $\{c_1,\dots,c_k\} \subset \mathbb{R}$.

Le lecteur vérifiera que l'écriture $\mu = \sum_{i\in[\![1,k]\!]} \lambda_i \delta_{c_i}$ est unique si l'on impose aux λ_i d'être tous non nuls. Le support est donc bien défini.
Les fonctions monomiales $t \mapsto t^n$ étant toutes continues, les moments $\mu_n = \mu(t^n) \in \mathbb{R}$ ($n \in \mathbb{N}$) d'une mesure positive à support fini sont toujours définis. Si μ est est une mesure à support fini $\Sigma \subset [\alpha,\beta]$, elle définit une mesure sur $[\alpha,\beta]$ que l'on note toujours μ.
Soient $[\alpha,\beta]$ un segment et w un poids sur $[\alpha,\beta]$. Soient $n \in \mathbb{N}^*$, $t_{1,n},\dots,t_{n,n}$ les zéros du $n^{\text{ème}}$ polynôme orthogonal correspondant et $\lambda_{1,n},\dots,\lambda_{k,n}$ les nombres de Christoffel. On pose :
$\mu^{(n)} = \sum_{k=0}^{n} \lambda_{k,n}\delta_{t_{k,n}}$. La suite de mesures positives discrètes $\left(\mu^{(n)}\right)_{n\in\mathbb{N}^*}$ converge en un sens que nous allons préciser vers la mesure μ définie par le poids w.
On reformule, pour cela, la proposition 19 de la page 552.

Théorème 20. Soient $[\alpha,\beta]$ un segment et w un poids sur $]\alpha,\beta[$.
Pour tout $f \in \mathcal{C}([\alpha,\beta])$, la suite de nombres réels $\left(\mu^{(n)}(f)\right)_{n\in\mathbb{N}^*} := \left(\sum_{k=1}^{n} f(t_k)\lambda_{k,n}\right)$ converge vers $\mu(f) := \int_\alpha^\beta f(t)w(t)\,\mathrm{d}t$.

On dit que la suite de mesures $(\mu^{(n)})_{n\in\mathbb{N}^*}$ *converge simplement* vers la mesure μ.
À une mesure *positive* μ sur $\mathbb{R}$ on associe sa *fonction de répartition* F. Pour une mesure de probabilité (c'est-à-dire si $\mu_0 = 1$), la définition « élémentaire » donnée ici des fonctions de répartition pour les cas qui nous intéressent est équivalente à celle donnée dans le module IV.3 (*cf.* la définition 3 et la proposition 3 de la page 757). Nous n'utiliserons pas ce résultat.

(i) Dans le cas d'un poids w sur un intervalle (α,β) (interprété comme fonction sur $\mathbb{R}$ par prolongement par zéro en dehors de (α,β)), la fonction de répartition F est définie par :

$$F : t \in \mathbb{R} \mapsto \begin{cases} F(t) := 0, & \text{si} \quad t < \alpha \\ F(t) := \displaystyle\int_\alpha^t w(\tau)\,\mathrm{d}\tau, & \text{si} \quad \alpha \leqslant t \leqslant \beta \\ F(t) := \displaystyle\int_\alpha^\beta w(\tau)\,\mathrm{d}\tau, & \text{si} \quad \beta < t \end{cases}.$$

Elle est *continue croissante* sur $\mathbb{R}$ et, pour tout $t \in \mathbb{R}$:

$$F(t) \leqslant \int_{-\infty}^{+\infty} w(\tau)\,\mathrm{d}\tau = \int_{\alpha}^{\beta} w(\tau)\,\mathrm{d}\tau = \mu_0.$$

(ii) Dans le cas d'une masse de Dirac δ_c en $c \in \mathbb{R}$, on définit la fonction de répartition $F = F_c$ par $F(t) := 0$, pour $t < c$ et $F_c(t) := 1$, pour $t \geqslant c$. Si $c = 0$, F s'appelle la fonction de Heaviside .

(iii) Dans le cas d'une combinaison linéaire finie à coefficients strictement positifs de masses de Dirac $\mu := \sum\limits_{k=1}^{m} \lambda_k \delta_{c_k}$, on définit la fonction de répartition F par $F := \sum\limits_{k=1}^{m} \lambda_k F_{c_k}$.

La fonction de répartition F est, dans tous les cas, une fonction croissante majorée par μ_0 sur $\mathbb{R}$, continue à droite en tout point. Dans le cas d'une mesure positive à support fini, F est constante par morceaux avec des discontinuités aux points appartenant au support de la mesure.
La fonction de répartition F étant croissante, $F(t)$ admet, pour tout $c \in \mathbb{R}$, une limite quand t tend vers c par valeurs strictement inférieures. On note $F(c^-)$ cette limite.
On vérifie que, dans les cas ci-dessus, la connaissance d'une mesure est équivalente à celle de sa fonction de répartition.
Dans le cas d'un poids w, on a :

$$F'(t) = w(t) \quad \text{pour } t \in \,]\alpha, \beta[\qquad \text{et} \qquad F'(t) = 0 \quad \text{pour } t \in \,]-\infty, \alpha[\, \cup \,]\beta, +\infty[.$$

Dans le cas d'une combinaison linéaire finie à coefficients strictement positifs de masses de Dirac $\mu := \sum\limits_{k=1}^{m} \lambda_k \delta_{c_k}$, on a, pour tout $h \in [\![1, m]\!]$:

$$F(c_h) - F(c_h^-) = \sum_{k=1}^{m} \lambda_k \big(F_k(c_h) - F_k(c_h^-)\big) = \lambda_h \big(F_h(c_h) - F_h(c_h^-)\big) = \lambda_h$$

et, F étant continue sur $\mathbb{R} \setminus \{c_1, \ldots, c_k\}$, $F(c) - F(c^-) = 0$, pour tout $c \notin \{c_1, \ldots, c_k\}$.

3 Polynômes orthogonaux classiques

La notion « moderne » de polynômes orthogonaux que nous venons de présenter est « très générale ». Les polynômes orthogonaux sont définis à partir d'un poids w sur lequel nous avons fait assez peu d'hypothèses. Mais, dans les applications « usuelles » (en physique classique ou quantique, en probabilités, en combinatoire, dans les sciences de l'ingénieur...) on ne rencontre le plus souvent *qu'un très petit nombre* de familles de polynômes orthogonaux, ce sont les *trois* familles de polynômes orthogonaux dits *classiques* :

1. les *polynômes de Jacobi* (avec leurs cas particuliers : *Lagrange*, *Tchebychev*, *Gegenbauer*...);
2. les polynômes de *Laguerre généralisés*;
3. les polynômes *d'Hermite*.

Les polynômes classiques vérifient bien sûr une relation de récurrence (théorème 4 (i)) mais ils ont des propriétés supplémentaires que nous allons mettre en évidence. Ils sont en particulier tous *solutions de certaines équations différentielles linéaires du second ordre* de la forme $ay'' + by' - \lambda_n y = 0$ (où a et b sont des polynômes en t indépendants de n, de

degrés respectifs 2 et 1, et λ_n une constante), de plus la dérivée d'un polynôme classique est encore un polynôme classique.
Aux *trois familles* de polynômes classiques certains auteurs en ajoutent une quatrième, celle des polynômes dits de Bessel[7] : *cf.* l'exercice III.4.9. (Comme l'a dit le mathématicien français Pascal Maroni : longtemps au nombre de trois, les familles de polynômes orthogonaux classiques sont, comme les trois mousquetaires, au nombre de quatre...) De façon un peu surprenante les polynômes de Bessel ont quasiment toutes les propriétés des polynômes orthogonaux classiques, sauf celle d'être orthogonaux pour un poids ! (Ils sont orthogonaux pour une forme linéaire $\mathscr{M}$ sur $\mathbb{R}[T]$ non triviale, mais celle-ci *ne peut pas être strictement positive* : *cf.* l'exercice 17 de la page 632.)
Dans un premier temps, nous montrerons comment les équations différentielles des trois grandes familles de polynômes orthogonaux classiques *apparaissent naturellement à partir d'un problème de valeurs propres* (ou problème spectral) en utilisant la théorie de Sturm-Liouville pour les équations différentielles linéaires du second ordre.
Nous donnerons ensuite un outil fondamental pour l'étude des familles de polynômes orthogonaux classiques : la *formule d'Olinde Rodrigues*.
Nous poursuivrons par une étude détaillée de chaque famille complétant les résultats précédents. Finalement les résultats seront rassemblés sous forme de table en 3.8.
Pour les cas de *Legendre*, de *Laguerre généralisé* et *d'Hermite*, nous détaillerons au passage *des applications à des problèmes importants issus de la physique* (classique et quantique). Ces applications sont principalement destinées aux étudiants intéressés par les mathématiques appliquées, la physique ou les sciences de l'ingénieur et peuvent être omises par les autres en première lecture. Nous terminerons par une introduction aux séries génératrices avec des applications à des problèmes d'approximation en normes quadratiques.

Dans cette partie, nous utiliserons les fonctions eulériennes Γ et B et les relations :

$$\Gamma(s+1) = s\Gamma(s) \qquad \text{et} \qquad B(u,v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}$$

(*cf.* les exercices 16 et 17 de la page 688 du livre [L2], ainsi que l'exercice 18 de la page 240 du présent ouvrage).

Nous utiliserons aussi la formule $\displaystyle\int_{-1}^{1} (1-t)^u (1+t)^v \,\mathrm{d}t = 2^{u+v+1} B(u,v)$, obtenue en faisant le changement de variable $s = \frac{t+1}{2}$.
Nous utiliserons le *symbole de Pochhammer* $(a)_n$ (introduit par le mathématicien prussien Leo A. Pochhammer et popularisé par le mathématicien français P. E. Appell à la fin du XIXe siècle). Pour tout $a \in \mathbb{C}$ et tout $n \in \mathbb{N}$, on note[8] :

- $(a)_n := a(a-1)\cdots(a-n+1)$ si $n \in \mathbb{N}^*$;
- $(a)_0 := 1$ si $n = 0$.

[7] Découverts par Burchnal et Chaundy en 1931.
[8] Dans certains livres sur les fonctions spéciales on trouve une notation différente :

$$(a)_n := a(a+1)\cdots(a+n-1).$$

Nous écrirons parfois, par convention, $a(a-1)\dots(a-h+1)$ au lieu de $(a)_h$. Lorsque $h=0$ le produit devra être interprété comme 1.
Si $a \in \mathbb{N}^*$, $(a)_n$ est le nombre d'arrangements : $(a)_n = A_a^n$.
Rappelons la définition des *coefficients binomiaux généralisés* (*cf.* le module *Arithmétique* du livre [L1]). Pour tout $a \in \mathbb{C}$ et tout $n \in \mathbb{N}$, on note : $\binom{a}{n} := \frac{(a)_n}{n!}$. On a :

- $\binom{a}{n} := \frac{a(a-1)\dots(a-n+1)}{n!}$ si $n \in \mathbb{N}^*$;
- $\binom{a}{n} := 1$ si $n = 0$.

En utilisant la fonction Γ, on obtient, si $a+n+1>0$:

$$(a+n)_n = \frac{\Gamma(a+n+1)}{\Gamma(a+1)} \qquad \text{et} \qquad \binom{a+n}{n} = \frac{\Gamma(a+n+1)}{n!\Gamma(a+1)}.$$

3.1 Équations différentielles et polynômes orthogonaux

3.1.1 Équations différentielles du second ordre sous forme de Sturm-Liouville

La théorie de Sturm-Liouville (découverte par Charles François Sturm vers 1830) permet de mettre en évidence les familles de polynômes orthogonaux classiques. Cette théorie a bien d'autres applications mais nous nous limiterons ici au minimum utile pour notre propos.

Problèmes spectraux en dimension finie

Dans le module *Espaces préhilbertiens* du livre [L2] nous avons rencontré (à propos de la *diagonalisation simultanée de deux formes quadratiques*) la situation suivante : $m \in \mathbb{N}^*$ est un entier, $A \in M_m(\mathbb{R})$ est une matrice *symétrique définie positive*, $B \in M_m(\mathbb{R})$ est une matrice *symétrique* et l'on cherche *un scalaire* λ *et un vecteur* $y \in \mathbb{R}^m$, $y \neq 0$, tels que[9] $(B-\lambda A)Y = 0$. Ce problème équivaut au problème de la recherche de valeurs propres et de vecteurs propres pour la matrice $A^{-1}B$. Nous allons revenir brièvement sur cette situation dont la théorie de Sturm-Liouville est une variante en dimension infinie. Pour plus de détails nous renvoyons le lecteur au paragraphe *Diagonalisation simultanée de deux formes quadratiques* du module *Espaces préhilbertiens* du livre [L2].
La matrice symétrique définie positive A définit une structure euclidienne Φ sur $\mathbb{R}^m$: $\left\|(x_1,\dots,x_m)\right\|_A^2 := {}^tXAX$, X étant le vecteur colonne associé à $(x_1,\dots,x_m)$.
On montre qu'il existe une matrice diagonale D et une matrice inversible P telles que $A = {}^tPP$ et $B = {}^tPDP$ et que l'on a $A^{-1}B = P^{-1}DP$. Les valeurs propres de $A^{-1}B$ (c'est-à-dire les termes diagonaux de D) sont les zéros $(\lambda_i)_{i\in[\![1,m]\!]}$ de $\det(B-\lambda A)$. Ceux-ci sont donc *réels*. S'ils sont *distincts*, toute famille correspondante de vecteurs propres $\big(i \in [\![1,m]\!]\big)$ est Φ*-orthogonale*.
Notons que la matrice $C := A^{-1}B$ n'est pas en général symétrique.
Soient $Y_1,\ Y_2 \in \mathbb{R}^m$ tels que $CY_1 = \lambda_1 Y_1$ et $CY_2 = \lambda_2 Y_2$ avec $\lambda_1 \neq \lambda_2$. Alors Y_1 et Y_2 sont Φ*-orthogonaux*, c'est-à-dire ${}^tY_1AY_2 = 0$. On a donc ${}^t(PY_1){}^t(PY_2) = {}^tY_1{}^tPPY_2 = 0$ et PY_1 et PY_2 sont orthogonaux pour la structure euclidienne canonique de $\mathbb{R}^m$.

[9] On rappelle que Y désigne le vecteur colonne associé à y.

Dans le cadre de la théorie de Sturm-Liouville, nous remplacerons A par un opérateur de multiplication par une fonction *strictement positive* w et B par un opérateur différentiel du second ordre (formellement) auto-adjoint ; C sera alors remplacée par un opérateur différentiel du second ordre D et nous étudierons les valeurs propres et les vecteurs propres de D. La matrice P sera remplacée par l'opérateur de multiplication par $\sqrt{w}$.

Opérateurs différentiels — Opérateurs de Sturm-Liouville — Problèmes spectraux

En relation avec les équations différentielles linéaires, nous utiliserons la notion *d'opérateur différentiel*. Ce point de vue a déjà été introduit dans le module *Équations différentielles* du livre [L1], mais on prendra garde à quelques changements de notation.

Proposition et définition 21. Soit J un intervalle ouvert non vide de $\mathbb{R}$. Étant données les fonctions $a_0, \ldots, a_m \in \mathcal{C}^\infty(J;\mathbb{R})$, on leur associe l'application $D : \mathcal{C}^\infty(J;\mathbb{R}) \to \mathcal{C}^\infty(J;\mathbb{R})$ définie par :

$$f \mapsto D(f) := a_m f^{(m)} + \cdots + a_0 f.$$

L'application D est linéaire et les $a_0, \ldots, a_m$ sont *déterminés de façon unique* par D. On dit que D est un *opérateur différentiel* sur $\mathcal{C}^\infty(J;\mathbb{R})$ et l'on note :

$$D = a_m \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^m + \cdots + a_0.$$

L'équation différentielle associée à D est :

$$Dy = a_m \frac{\mathrm{d}^m y}{\mathrm{d}t^m} + \cdots + a_0 y = a_m y^{(m)} + \cdots + a_0 y = 0.$$

Elle est linéaire homogène.
Si $a_m \neq 0$, on dit que D est d'ordre m.

Démonstration. Supposons qu'un endomorphisme linéaire de $\mathcal{C}^\infty(J;\mathbb{R})$ soit défini par $D : f \mapsto a_m f^{(m)} + \cdots + a_0 f$. Montrons que si $D = 0$, alors $a_0 = \cdots = a_m = 0$.
On a $D(1) = a_0$ et $D(1) = 0$, donc $a_0 = 0$. On a alors $D(t) = a_1$ et $D(t) = 0$, donc $a_1 = 0$. On montre $a_0 = a_1 = \cdots = a_m = 0$ par récurrence. ∎

Si $a_0, \ldots, a_m \in \mathbb{R}[T]$, D induit un opérateur sur $\mathbb{R}[T]$ (toujours noté D, par abus de langage).
On définit le *produit* de deux opérateurs différentiels D_1 et D_2 en utilisant la composition des applications : $D_1 D_2 := D_1 \circ D_2$. Ce produit est associatif.
On prendra garde au fait que les opérateurs différentiels *ne commutent en général pas*. Par exemple t et $\frac{\mathrm{d}}{\mathrm{d}t}$ sont des opérateurs différentiels d'ordres respectifs 0 et 1, et l'on a $\frac{\mathrm{d}}{\mathrm{d}t} t f(t) = t f'(t) + f(t)$ et $t \frac{\mathrm{d}}{\mathrm{d}t} f(t) = t f'(t)$, donc

$$\frac{\mathrm{d}}{\mathrm{d}t} t - t \frac{\mathrm{d}}{\mathrm{d}t} = 1.$$

Le produit de deux opérateurs différentiels est un opérateur différentiel. Pour le prouver, il suffit, en utilisant la linéarité et en raisonnant par récurrence, de montrer que, pour tout $c \in \mathcal{C}^\infty(J;\mathbb{R})$ et tous $p \in \mathbb{N}$, $\frac{\mathrm{d}}{\mathrm{d}t}c\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^p$ est un opérateur différentiel.
On a, pour tout $f \in \mathcal{C}^\infty(J;\mathbb{R})$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(c\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^p f\right) = c'\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^p f + c\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^{p+1} f \qquad \text{donc} \qquad \frac{\mathrm{d}}{\mathrm{d}t}c\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^p = c'\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^p + c\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^{p+1}.$$

Soient $D_1 = a_m\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^m + \cdots + a_0$ d'ordre m et $D_2 = b_n\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^n + \cdots + b_0$ d'ordre n, on vérifie que D_1D_2 est d'ordre $m+n$. Par suite, si $D_1D_2 = 1$, on a $m = n = 0$, $D_1 = a_0$, $D_2 = b_0$ et $a_0b_0 = 1$. Les opérateurs différentiels inversibles sont donc les opérateurs de multiplication par une fonction constante non nulle sur J.

On note traditionnellement de la même façon la *fonction* a et *l'opérateur de multiplication par* a. Par ailleurs, si D est un opérateur différentiel et f une fonction, on écrit souvent Df au lieu de $D(f)$. Cela peut créer des ambiguïtés. Ainsi la notation Daf pour la *fonction* $Da(f)$ (c'est-à-dire pour l'opérateur Da appliqué à f) peut conduire à des erreurs. On a $Da(f) = D(af)$, par contre $\big(D(a)\big)f$, c'est-à-dire le produit des *fonctions* $D(a)$ et f est, en général différent de la fonction $Da(f)$. Par exemple $\frac{\mathrm{d}}{\mathrm{d}t}f(g) = \frac{\mathrm{d}}{\mathrm{d}t}(fg) = f'g + fg'$, mais $\big(\frac{\mathrm{d}}{\mathrm{d}t}f)\big)\,g = f'g$.

Définition 5. Soit $D = a_m\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^m + \cdots + a_1\frac{\mathrm{d}}{\mathrm{d}t} + a_0$ un opérateur différentiel. On définit son *adjoint formel* D^* par :

$$D^* := \left(-\frac{\mathrm{d}}{\mathrm{d}t}\right)^m a_m + \cdots - \frac{\mathrm{d}}{\mathrm{d}t}a_1 + a_0.$$

Si $D = D^*$, on dit que D est *formellement auto-adjoint*.

Cette terminologie sera justifiée plus loin (*cf.* la proposition 23 de la page suivante). Elle est utilisée dans plusieurs théories, par exemple en théorie des distributions.[10]
Si $\lambda_1,\ \lambda_2 \in \mathbb{R}$, l'adjoint formel de $\lambda_1D_1 + \lambda_2D_2$ est $\lambda_1D_1^* + \lambda_2D_2^*$. Le passage à l'adjoint formel est linéaire.
L'opérateur de multiplication par a est formellement auto-ajoint. L'adjoint formel de $\frac{\mathrm{d}}{\mathrm{d}t}$ est $-\frac{\mathrm{d}}{\mathrm{d}t}$. L'adjoint formel de $a\frac{\mathrm{d}}{\mathrm{d}t}$ est $-\frac{\mathrm{d}}{\mathrm{d}t}a = -a\frac{\mathrm{d}}{\mathrm{d}t} - a'$. Par suite un opérateur d'ordre 1 n'est jamais auto-adjoint. Nous allons étudier le cas des opérateurs d'ordre 2.
Soit $D := a_2\big(\frac{\mathrm{d}}{\mathrm{d}t}\big)^2 + a_1\frac{\mathrm{d}}{\mathrm{d}t} + a_0$. Son adjoint formel est :

$$D^* = \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 a_2 - \frac{\mathrm{d}}{\mathrm{d}t}a_1 + a_0 = a_2\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 + (2a_2' - a_1)\frac{\mathrm{d}}{\mathrm{d}t} + (a_2'' - a_1' + a_0).$$

Par suite D est formellement auto-adjoint si, et seulement si, $a_2' = a_1$. Il s'écrit alors

$$D = \frac{\mathrm{d}}{\mathrm{d}t}a_2\frac{\mathrm{d}}{\mathrm{d}t} + a_0. \tag{10}$$

[10] *Cf.* L. Schwartz, *Méthodes mathématiques de la physique, Hermann, Paris.*

On vérifie facilement le résultat suivant :

Proposition 22 (Identité de Lagrange). Soit $D := a_2 \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 + a_1 \frac{\mathrm{d}}{\mathrm{d}t} + a_0$. Pour tous $f,\ g \in \mathcal{C}^\infty(J;\mathbb{R})$, on a *l'identité de Lagrange* :

$$gDf - fD^*g = \big((a_1 - a_2')fg + a_2(gf' - fg')\big)'.$$

Si $D = D^*$: $gDf - fDg = (a_2(gf' - fg'))'$.

Notons $P(f,g) := (a_1 - a_2')fg + a_2(gf' - fg')$.
L'identité de Lagrange s'écrit $gDf - fDg = \frac{\mathrm{d}}{\mathrm{d}t}P(f,g)$. Par intégration (quand elle est possible...), on obtient la *formule de Green*.

Proposition 23 (Formule de Green).
Soient $J :=]\alpha,\beta[$ et $D := a_2 \frac{\mathrm{d}}{\mathrm{d}t}^2 + a_1 \frac{\mathrm{d}}{\mathrm{d}t} + a_0$. Soient $f,\ g \in \mathcal{C}^\infty(J;\mathbb{R})$ telles que les intégrales $\displaystyle\int_J gDf$ et $\displaystyle\int_J fD^*g$ soient convergentes au sens de Riemann généralisé sur J. Alors :

(i) $[P(f,g)]_{\alpha'}^{\beta'}$ admet une limite, notée $[P(f,g)]_\alpha^\beta$, quand α' tend vers α^+ et β' tend vers β^- ;

(ii) on a la formule de Green :

$$\int_J gDf - \int_J fD^*g = [P(f,g)]_\alpha^\beta\,; \tag{11}$$

(iii) si D est formellement auto-adjoint et si, de plus, f et g (respectivement f' et g') tendent vers 0 quand t tend vers α ou β, on a

$$\int_J gDf = \int_J fDg.$$

Ainsi un opérateur différentiel d'ordre 2 *formellement auto-adjoint* peut être interprété comme un opérateur *auto-adjoint* sur des espaces convenables. Nous ne développerons pas ce point de vue et nous nous limiterons à l'étude de certains cas particuliers.

Définition 6. Un opérateur différentiel du second ordre D sur l'intervalle ouvert J est dit *sous la forme de Sturm-Liouville* si, et seulement s'il existe des fonctions $p, q \in \mathcal{C}^\infty(J,\mathbb{R})$ telles que :

$$D = \frac{\mathrm{d}}{\mathrm{d}t} p \frac{\mathrm{d}}{\mathrm{d}t} + q \tag{12}$$

Un opérateur est dit de Sturm-Liouville si on peut l'écrire sous la forme de Sturm-Liouville.

Des résultats établis ci-dessus (*cf.* (10)) on déduit le résultat suivant.

Proposition 24. Un opérateur $D := a_2 \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 + a_1 \frac{\mathrm{d}}{\mathrm{d}t} + a_0$ peut s'écrire sous forme de Sturm-Liouville $D = \frac{\mathrm{d}}{\mathrm{d}t} p \frac{\mathrm{d}}{\mathrm{d}t} + q$ si, et seulement s'il est formellement auto-adjoint. On a alors $p = a_2$ et $q = a_0$.

Considérons maintenant le problème aux valeurs propres pour un opérateur différentiel D. Il s'agit d'étudier les couples (λ, y), avec $\lambda \in \mathbb{R}$ et $y \in \mathcal{C}^\infty(J, \mathbb{R})$ (y non identiquement nul), vérifiant « certaines conditions au bord », tels que $Dy - \lambda y = 0$.
Compte tenu de l'assertion (iii) de la proposition 23 les conditions au bord que nous utiliserons seront de la forme $\lim\limits_{t \to \alpha^+} y(t) = \lim\limits_{t \to \beta^-} y(t) = 0$ *ou* $\lim\limits_{t \to \alpha^+} y'(t) = \lim\limits_{t \to \beta^-} y'(t) = 0$.
Dans ces conditions, si D est un opérateur de Sturm-Liouville et si l'on a deux couples (λ_1, y_1) et (λ_2, y_2), avec $\lambda_1 \neq \lambda_2$, tels que les intégrales $\displaystyle\int_J y_2 D y_1$ et $\displaystyle\int_J y_1 D y_2$ soient convergentes et que :

$$Dy_1 - \lambda y_1 = 0 \quad \text{et} \quad Dy_2 - \lambda y_2 = 0,$$

on a $(\, y_1 \mid y_2 \,) = \displaystyle\int_J y_1 y_2 \, \mathrm{d}t = 0$. Ainsi y_1 et y_2 sont orthogonales. En effet (d'après la proposition 23) :

$$0 = (\, Dy_1 \mid y_2 \,) - (\, y_1 \mid Dy_2 \,) = (\lambda_1 - \lambda_2)(\, y_1 \mid y_2 \,).$$

Exemple. On considère, sur l'intervalle réel $[0, \pi]$, l'équation différentielle

$$u'' := \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} u = \lambda u. \tag{13}$$

L'opérateur $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}$ est un opérateur de Sturm-Liouville (avec $p := 1$ et $q := 0$).
Nous allons résoudre le problème des valeurs propres pour $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}$ avec les conditions aux limites $u'(0) = u'(\pi) = 0$.
Il s'agit de trouver les couples (λ, u) vérifiant (13), tels que u vérifie les conditions aux limites.
À $\lambda \in \mathbb{R}$ fixé, la solution générale de l'équation différentielle (13) est :

1. si $\lambda < 0$ (en posant $\omega := \sqrt{-\lambda}$) : $u(\theta) = c_1 \cos \omega\theta + c_2 \sin \omega\theta$,
2. si $\lambda > 0$ (en posant $k = \sqrt{\lambda}$) : $u(\theta) = c_1 e^{k\theta} + c_2 e^{-k\theta}$,
3. si $\lambda = 0$: $u(\theta) = c_1 + c_2 \theta$.

On vérifie facilement que dans les cas 2 et 3 (c'est-à-dire si $\lambda \geqslant 0$), il n'existe pas de solution u vérifiant les conditions aux limites imposées. Supposons donc $\lambda < 0$.
Soit $u(\theta) = c_1 \cos \omega\theta + c_2 \sin \omega\theta$. On a $u'(\theta) = -c_1 \omega \sin \omega\theta + c_2 \omega \cos \omega\theta$, les conditions aux limites s'écrivent donc :

$$-c_1 \sin(\omega\pi) + c_2 \cos(\omega\pi) = 0 \quad \text{et} \quad c_2 = 0,$$

ce qui équivaut à $c_1 \sin(\omega\pi) = 0$. Les solutions non triviales (c'est-à-dire $c_1 \neq 0$) sont donc $\big((\lambda := -n^2, u := c_1 \cos n\theta)\big)_{n \in \mathbb{N}}$.

D'après la proposition 23, pour tous $m, n \in \mathbb{N}$, $m \neq n$, les fonctions $\cos n\theta$ et $\cos m\theta$ sont *orthogonales* sur $[0, \pi]$: $\displaystyle\int_0^\pi \cos m\theta \cos n\theta \, \mathrm{d}\theta = 0$. On le vérifie aussi directement en utilisant $\cos m\theta \cos n\theta = \frac{1}{2}\big(\cos(m+n)\theta + \cos(m-n)\theta\big)$.

Si f est une fonction 2π-périodique, l'intégrale $\displaystyle\int_\alpha^{\alpha+2\pi} f(\theta)\,\mathrm{d}\theta$ ne dépend pas de $\alpha \in \mathbb{R}$.

On parle d'intégrale sur une période et on la note $\displaystyle\int_{S^1} f(\theta)\,\mathrm{d}\theta$.

Ainsi $\displaystyle\int_{S^1} \cos m\theta \cos n\theta \,\mathrm{d}\theta = 0$ pour tous $m, n \in \mathbb{N}$ tels que $m \neq n$.

Le problème de Sturm-Liouville

Définition 7. Soient p, q, w des fonctions réelles sur un intervalle ouvert $J =]\alpha, \beta[$ de $\mathbb{R}$. On suppose p, q et w de classe $\mathcal{C}^\infty$ sur J et $w > 0$.
L'équation différentielle linéaire du second ordre :

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(p\frac{\mathrm{d}}{\mathrm{d}t}y\right) + qy - \lambda wy = 0 \tag{14}$$

est dite *sous la forme de Sturm-Liouville*.
La fonction w s'appelle le *poids*. Un couple (λ, y), où $\lambda \in \mathbb{R}$ et où y est une fonction de classe $\mathcal{C}^\infty$, non identiquement nulle, vérifiant (14) est, par définition, une *solution du problème de Sturm-Liouville* correspondant.

Plus généralement, on peut étendre la définition au cas où q et w sont continues sur J, p est de classe $\mathcal{C}^1$ sur J, $w \geqslant 0$ et considérer les solutions y de classe $\mathcal{C}^\infty$.

Exemples. (i) L'équation $\big((1-t^2)y'\big)' + n(n+1)y = 0$ est sous forme de Sturm-Liouville ($p(t) = 1 - t^2$, $q(t) = 0$, $\lambda = -n(n+1), w(t) = 1$). En développant, on trouve $(1-t^2)y'' - 2ty' + n(n+1)y = 0$. C'est l'équation différentielle de Legendre, dont, nous le verrons, le $n^{\text{ème}}$ polynôme de Legendre P_n est une solution.
(ii) L'équation $(ty')' + ty = \nu^2 \frac{1}{t} y$ est sous forme de Sturm-Liouville ($p(t) = t$, $q(t) = t$, $\lambda = \nu^2$, $w(t) = 1/t$). En développant et en multipliant par $t = 1/w(t)$, on trouve $t^2y'' + ty' + (t^2 - \nu^2)y = 0$ qui est une équation différentielle dite de Bessel.

Posons $S := \dfrac{\mathrm{d}}{\mathrm{d}t}p\dfrac{\mathrm{d}}{\mathrm{d}t} + q$. Si les intégrales ont un sens (c'est-à-dire convergent au sens de Riemann généralisé sur J), on note, pour tous f, $g \in \mathcal{C}^\infty$:

$$(\,f \mid g\,) := \int_\alpha^\beta f(t)g(t)\,\mathrm{d}t \qquad \text{et} \qquad (\,f \mid g\,)_w := \int_\alpha^\beta f(t)g(t)w(t)\,\mathrm{d}t.$$

Soient y_1, $y_2 \in \mathcal{C}^\infty$ telles que les intégrales $(\,y_1 \mid Sy_2\,)$ et $(\,Sy_1 \mid y_2\,)$ convergent. Nous avons vu (*cf.* la proposition 23 de la page 560) que si y_1 et y_2 « s'annulent aux extrémités

de J », on a $(\,Sy_1 \mid y_2\,) = (\,Sy_2 \mid y_1\,)$. Si, de plus, il existe $\lambda_1,\ \lambda_2 \in \mathbb{R}^*$ tels que (λ_1, y_1) et (λ_2, y_2) soient deux solutions du problème de Sturm-Liouville, avec $\lambda_1 \neq \lambda_2$, alors l'intégrale $(\,y_1 \mid y_2\,)_w = \displaystyle\int_J y_1(t)y_2(t)w(t)\,\mathrm{d}t$ est convergente et l'on a :

$$(\,Sy_1 \mid y_2\,) = (\,\lambda_1 w y_1 \mid y_2\,) = \lambda_1(\,y_1 \mid y_2\,)_w$$
$$(\,y_1 \mid Sy_2\,) = (\,y_1 \mid \lambda_2 w y_2\,) = \lambda_2(\,y_1 \mid y_2\,)_w.$$

D'où $\lambda_1(\,y_1 \mid y_2\,)_w = \lambda_2(\,y_1 \mid y_2\,)_w$. Si $\lambda_1 \neq \lambda_2$, $(\,y_1 \mid y_2\,)_w = 0$: y_1 et y_2 sont orthogonaux pour le produit scalaire avec le poids w.
On peut associer à y_1 et y_2 les fonctions $\psi_1 := \sqrt{w}y_1$ et $\psi_2 := \sqrt{w}y_2$.
Alors $(\,\psi_1 \mid \psi_2\,) = (\,y_1 \mid y_2\,)_w = 0$: les fonctions ψ_1 et ψ_2 sont orthogonales pour $(\,.\mid .\,)$.
Nous verrons plus loin, dans la partie 3, des exemples intéressants de ce mécanisme (passage des polynômes d'Hermite aux fonctions d'Hermite et des polynômes de Laguerre aux fonctions de Laguerre).

Exemple. On reprend l'exemple de la page 561. On peut « transporter » les équations différentielles d'ordre deux $\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^2 u + n^2 u = 0$ $(n \in \mathbb{N})$ sur $]-1,1[$ par le changement de variable $\theta \in]0,\pi[\mapsto t = \cos\theta \in]-1,1[$.
On obtient $\sqrt{1-t^2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{1-t^2}\frac{\mathrm{d}y}{\mathrm{d}t}\right) + n^2 y = 0$, et, en divisant par $\sqrt{1-t^2}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{1-t^2}\frac{\mathrm{d}y}{\mathrm{d}t}\right) + n^2\frac{1}{\sqrt{1-t^2}}y = 0. \tag{15}$$

C'est un exemple d'équation différentielle linéaire du deuxième ordre sous forme *de Sturm-Liouville*; $w : t \mapsto w(t) = \frac{1}{\sqrt{1-t^2}}$ est le *poids*. Ce poids vient ici de la comparaison des *longueurs*, mesurées, d'une part sur le cercle paramétré par l'angle θ et, d'autre part, sur le segment : $\mathrm{d}\theta = -w(t)\,\mathrm{d}t$.
En développant (15) et en multipliant par $\sqrt{1-t^2} = \frac{1}{w(t)}$, on trouve :

$$(1-t^2)y'' - ty' + n^2 y = 0, \tag{16}$$

qui est une équation différentielle linéaire du second ordre sous forme « ordinaire », appelée équation différentielle de *Tchebychev*.
La résolution des équations du type (13), avec $\lambda = -n^2, n \in \mathbb{N}$, permet de trouver des solutions pour les équations (16). Pour tout $n \in \mathbb{N}$, $\cos n\theta$ est un polynôme de degré n en $\cos\theta$: $\cos n\theta = T_n(\cos\theta)$; T_n est, par définition, le $n^{\text{ème}}$ *polynôme de Tchebychev* (*cf.* le module *Polynômes et fractions rationnelles* du livre [L1]). Ainsi le polynôme T_n est solution de l'équation différentielle (16), que nous appellerons équation différentielle des polynômes de Tchebychev. On peut dire les choses de façon un peu différente : l'opérateur différentiel :

$$D : y \mapsto Dy := (1-t^2)y'' - ty'$$

(que l'on peut voir, au choix, comme opérateur linéaire sur $\mathbb{R}[T]$ ou sur $\mathcal{C}^\infty\big([-1,1]\big)$), admet, pour tout $n \in \mathbb{N}$, la valeur propre $-n^2$ et un vecteur propre correspondant est le $n^{\text{ème}}$ polynôme de Tchebychev T_n.

La description que nous venons de donner permet de voir facilement que les polynômes de Tchebychev $(T_n)_{n\in\mathbb{N}}$ sont une famille de polynômes orthogonaux pour le poids $w(t) = \frac{1}{\sqrt{1-t^2}}$ sur l'intervalle $]1,-1[$. En effet, en revenant au cercle, nous avons vu (*cf.* l'exemple 3.1.1) que, si $n \neq m$, les fonctions $\cos n\theta$ et $\cos m\theta$ sont orthogonales pour $(f \mid g) = \int_{S^1} f(\theta)g(\theta)\,\mathrm{d}\theta$: $\int_{S^1} \cos n\theta \cos m\theta\,\mathrm{d}\theta = 0$.

Par parité des $\cos k\theta$: $\int_{S^1} \cos n\theta \cos m\theta\,\mathrm{d}\theta = 2\int_0^\pi \cos n\theta \cos m\theta\,\mathrm{d}\theta$. En passant en variable t, l'on obtient, si $n \neq m$:

$$0 = -\int_{-\pi}^{\pi} \cos n\theta \cos m\theta\,\mathrm{d}\theta = \int_{-1}^{1} T_n(t)T_m(t)\frac{\mathrm{d}t}{\sqrt{1-t^2}}.$$

Il est bon, pour la suite, d'interpréter cet exemple, en relation avec le problème physique des vibrations (on peut penser à un cercle métallique que l'on frappe avec un marteau). Ces problèmes de vibration sont modélisés par l'équation des ondes. Dans ce cas, c'est l'équation des ondes unidimensionnelle sur un cercle : il y a deux variables $\theta \in S^1 = \mathbb{R}/2\pi\mathbb{Z}$ et le temps $\tau \in \mathbb{R}$ et une fonction inconnue $U(\theta,\tau)$ (U est le « déplacement transversal »). L'équation est une équation aux dérivées partielles (dite hyperbolique) $\dfrac{\partial^2 U}{\partial\theta^2} - \dfrac{1}{c^2}\dfrac{\partial^2 U}{\partial\tau^2} = 0$, avec $U = U(\theta,\tau)$ (*cf.* la partie 4.1 du module *Équations aux dérivées partielles*).

Pour le cas classique des ondes sur la droite, d'Alembert a résolu le problème en 1747 (en faisant le changement de variables $u = \theta - c\tau, v = \theta + c\tau$), la solution est donnée par (*cf.* (82), page 409) :

$$U(\theta,\tau) = f(\theta - c\tau) + g(\theta + c\tau),$$

où f et g sont deux fonctions « arbitraires » de classe $\mathcal{C}^2$. On superpose une onde progressive de vitesse de propagation c et une onde régressive de même vitesse voyageant en sens inverse. Ici on s'intéresse à des solutions particulières, les solutions stationnaires ($f = g$). On cherche de telles solutions par la *méthode de séparation des variables* : on pose $U(\theta,\tau) := \Theta(\theta)V(\tau)$. On obtient (en supposant U non nulle) : $\dfrac{\frac{\mathrm{d}^2\Theta}{\mathrm{d}\theta^2}}{\Theta} = \dfrac{1}{c^2}\dfrac{\frac{\mathrm{d}^2 V}{\mathrm{d}\tau^2}}{V}$.

Le membre de gauche ne dépend que de θ, celui de droite ne dépend que de τ, ils sont donc égaux à une même constante $\lambda \in \mathbb{R}$. On obtient deux équations différentielles ordinaires découplées :

$$\frac{\mathrm{d}^2\Theta}{\mathrm{d}\theta^2} - \lambda\Theta = 0 \quad \text{et} \quad \frac{\mathrm{d}^2 V}{\mathrm{d}\tau^2} - c^2\lambda V = 0.$$

Les solutions $U = \Theta V$ correspondantes sont donc $U(\theta,\tau) = C\cos\omega(\theta-\theta_0)\cos\omega(\tau-\tau_0)$ (avec $\lambda = -\omega^2$). Si l'on veut des ondes stationnaires sur le cercle, il faut imposer la condition de périodicité sur θ. On est donc ramené à résoudre le problème étudié plus haut (*cf.* (13), page 561) : on doit avoir $\pm\omega \in \mathbb{N}$, $n^2 := \omega^2$; n correspond au $n^{\text{ème}}$ harmonique. Supposons $\theta_0 = 0, \tau_0 = 0, C = 1$, on a : $U(\theta,\tau) = \cos n\theta \cos cn\tau$. Les points fixes de la vibration, les *nœuds*, sont les $2n$ points $\theta = \pm\pi$, modulo $2\pi/n$. Par passage à la coordonnée t (i.e. par projection horizontale sur la verticale), ils correspondent aux n zéros $t_{1,n},\ldots,t_{n,n}$ du $n^{\text{ème}}$ polynôme de Tchebychev T_n : dans la vibration du cercle S^1, les n paires de points d'ordonnées respectives $t_{1,n},\ldots,t_{n,n}$ restent fixes (on dit que c'est une vibration *zonale*). Les zones deviennent de plus en plus étroites quand on se rapproche du « haut » ou du « bas » du cercle : les zéros des polynômes de Tchebychev « se condensent » sur 1 et -1.

Mise sous forme de Sturm-Liouville

Partons maintenant d'un problème aux valeurs propres pour une équation différentielle linéaire du second ordre, *sous forme classique* (a_0, a_1, a_2 sont des fonctions de classe $\mathcal{C}^\infty$ sur $J :=]\alpha, \beta[$) :

$$Dy = a_2(t)y'' + a_1(t)y' + a_0(t)y = \lambda y \tag{17}$$

et essayons de le mettre sous forme d'un problème de Sturm-Liouville équivalent (on suppose l'équation régulière sur $]\alpha, \beta[$, c'est-à-dire $a_2(t) \neq 0$ sur cet intervalle).
Pour cela, on multiplie l'équation aux vecteurs propres par une fonction w (un poids *inconnu*), en imposant $w > 0$ et w de classe $\mathcal{C}^\infty$ sur $]\alpha, \beta[$, et l'on cherche à identifier (p, q sont *inconnus*) :

$$\begin{aligned} w(t)\big(a_2(t)y'' + a_1(t)y' + a_0(t)y\big) &= \frac{\mathrm{d}}{\mathrm{d}t}\Big[p(t)\frac{\mathrm{d}}{\mathrm{d}t}y(t)\Big] + q(t)y(t) \\ &= p(t)y''(t) + p'(t)y'(t) + q(t)y(t). \end{aligned}$$

On doit avoir $wa_2 = p$ (donc p doit être de classe $\mathcal{C}^\infty$ et ne pas s'annuler sur $]\alpha, \beta[$), $wa_1 = p'$ et $wa_0 = q$, et par suite $\dfrac{p'}{p} = \dfrac{a_1}{a_2}$.
On peut donc prendre $p(t) := \exp\left(\int_{t_0}^t \frac{a_1(\tau)}{a_2(\tau)}\,\mathrm{d}\tau\right)$, où $t_0 \in]\alpha, \beta[$, puis :

$$w(t) := a_2(t)^{-1} \exp\left(\int_{t_0}^t \frac{a_1(\tau)}{a_2(\tau)}\,\mathrm{d}\tau\right).$$

On vérifie que la fonction w ainsi obtenue ne s'annule pas sur $]\alpha, \beta[$.
On peut donc remplacer l'étude du problème aux valeurs propres (17) par celle du problème de Sturm-Liouville $\dfrac{\mathrm{d}}{\mathrm{d}t}\Big[p\dfrac{\mathrm{d}}{\mathrm{d}t}y\Big] + qy = \lambda wy$, avec :

$$w(t) := a_2(t)^{-1}\mathrm{e}^{\int_{t_0}^t \frac{a_1(\tau)}{a_2(\tau)}\,\mathrm{d}\tau}, \quad p(t) := \mathrm{e}^{\int_{t_0}^t \frac{a_1(\tau)}{a_2(\tau)}\,\mathrm{d}\tau} \quad \text{et} \quad q(t) = w(t)a_0(t). \tag{18}$$

On peut maintenant préciser l'analogie avec la dimension finie évoquée dans la partie *Problèmes spectraux en dimension finie* page 557 : l'analogue de la matrice C est l'opérateur différentiel D, l'analogue de la matrice symétrique définie positive A est l'opérateur de multiplication par w (qui est inversible et formellement auto-adjoint) et l'analogue de la matrice symétrique positive B est l'opérateur $\frac{\mathrm{d}}{\mathrm{d}t}\Big[p\frac{\mathrm{d}}{\mathrm{d}t}\Big] + q$ (qui est formellement auto-adjoint).

Exercice 5.
Mettre l'équation différentielle $t^3y'' - ty' + 2y = 0$ sur $[0, 1]$ sous forme de Sturm-Liouville.
Solution. On a $w(t) = t^{-3}\,\mathrm{e}^{-\int t^{-2}\mathrm{d}t} = t^{-3}\,\mathrm{e}^{1/t}$, d'où la forme de Sturm-Liouville :

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\mathrm{e}^{1/t}\frac{\mathrm{d}}{\mathrm{d}t}y\right] + 2t^{-3}\,\mathrm{e}^{1/t}y = 0.$$

3.1.2 Équations différentielles et polynômes orthogonaux classiques

On rencontre *tous* les polynômes orthogonaux classiques dans l'étude de problèmes de valeurs propres pour des opérateurs différentiels linéaires du second ordre à coefficients polynomiaux. De tels problèmes spectraux apparaissent systématiquement dans un très grand nombre d'applications mathématiques et physiques, comme les problèmes de vibrations, d'ondes ou de fonctions d'onde en mécanique quantique (solutions stationnaires de l'équation de Schrödinger). Nous en avons vu un exemple plus haut, avec le cas des polynômes de Tchebychev. Il n'est donc pas surprenant que les polynômes orthogonaux classiques soient exactement les outils qui permettent de résoudre les plus « simples » de ces problèmes (vibrations de cercles, comme plus haut, de sphères, d'hypersphères, oscillateur harmonique et atome d'hydrogène en mécanique quantique...).

Si $Q, L \in \mathbb{R}[T]$, on considère l'opérateur différentiel linéaire d'ordre deux, sur $\mathbb{R}[T]$:

$$D : f \mapsto Df := Q(t)f'' + L(t)f',$$

On s'intéresse au problème des vecteurs propres et des valeurs propres de cet opérateur :

$$Df = Q(t)f'' + L(t)f' = \lambda f,$$

et l'on cherche des conditions, sur Q et L, pour que D admette une suite de vecteurs propres $(p_n)_{n\in\mathbb{N}}$, de degrés respectifs $\deg p_n = n$. (On est exactement dans cette situation dans le cas de Tchebychev étudié plus haut : $Q(t) = 1 - t^2$, $L(t) = -t$, $\lambda = -n^2$ et $p_n = T_n$; *cf.* (16).)
Dans ce cas $Dp_1 = Lp_1' = \lambda_1 p_1$, donc $\deg L \leqslant 1$, $Dp_n = Qp_n'' + Lp_n' = \lambda_n p_n$, pour $n > 1$, donc $\deg Q + n - 2 \leqslant n$ et $\deg Q \leqslant 2$.
On suppose, à partir de maintenant, que $\deg L \leqslant 1$ et $\deg Q \leqslant 2$ (toute équation différentielle de la forme $Qy'' + Ly' - \lambda y = 0$ est alors dite *hypergéométrique*). On note $Q(t) = \gamma t^2 + \gamma' t + \gamma''$ et $L(t) = \eta t + \eta'$.
Soit $n \in \mathbb{N}$. *S'il existe* un réel λ_n et une solution polynomiale p_n, de degré n, de l'équation différentielle $Qy'' + Ly' = \lambda_n y$, on a *nécessairement* :

$$\lambda_n = n\big((n-1)\gamma + \eta\big). \tag{19}$$

L'opérateur $D := Q\dfrac{\mathrm{d}^2}{\mathrm{d}t^2} + L\dfrac{\mathrm{d}}{\mathrm{d}t}$ étant fixé, on choisit un intervalle $J :=]\alpha, \beta[$ sur lequel Q ne s'annule pas et dont les extrémités finies éventuelles sont zéros du polynôme Q. Le signe de Q est constant sur cet intervalle et, quitte à multiplier D par -1, on peut supposer $Q > 0$, ce que nous ferons dans la suite.
On met ensuite l'équation $Df = Q(t)f'' + L(t)f' = \lambda f$ sous forme de Sturm-Liouville sur l'intervalle J. Le poids w doit vérifier la relation $(Qw)' = Lw$, que l'on appelle ici *l'équation résolvante*. On pose $w(t) = \frac{1}{Q(t)} e^{\int_{t_0}^{t} \frac{L(\tau)}{Q(\tau)} d\tau}$ (avec $t_0 \in J$). On a $w > 0$ sur J.
L'équation différentielle $Df = \lambda f$ est équivalente, sur J, à l'équation différentielle de Sturm-Liouville :

$$\big(wQf'\big)' = \lambda w f. \tag{20}$$

Théorème 25.
Soit l'équation différentielle $Dy := Qy'' + Ly' = \lambda y$, où $Q, L \in \mathbb{R}[T]$. On suppose $\deg Q \leqslant 2$ et $\deg L \leqslant 1$, et l'on note $Q(t) := \gamma t^2 + \gamma' t + \gamma''$ et $L(t) := \eta t + \eta'$. Soit $J :=]\alpha, \beta[$ un intervalle (borné ou non) sur lequel Q ne s'annule pas et dont les extrémités éventuelles à distance finie sont zéros du polynôme Q, on suppose $Q > 0$ sur J. Soit w une solution, sur J, de l'équation résolvante $(Qw)' = Lw$. On a $w > 0$ sur J.
On suppose vérifiées les conditions suivantes

(1) La fonction w a tous ses moments finis sur l'intervalle J.

(2) La fonction Qw tend vers zéro aux bornes *finies* de I.

(3) L'équation $\gamma t + \eta = 0$ n'admet pas de solution dans $\mathbb{N}$.

Alors, l'opérateur D sur $\mathbb{R}[T]$ admet une suite $(p_n)_{n\in\mathbb{N}}$ de vecteurs propres, avec $\deg p_n = n$. Les valeurs propres correspondantes sont $\lambda_n := n\big((n-1)\gamma + \eta\big)$. Les $(p_n)_{n\in\mathbb{N}}$ forment une suite de polynômes orthogonaux pour le poids w sur J.

Démonstration. On munit $\mathbb{R}[T]$ du produit scalaire $(\,p \mid q\,)_w := \displaystyle\int_J p(t)q(t)w(t)\,\mathrm{d}t$ (l'intégrale existe d'après l'hypothèse faite sur les moments). On a, en utilisant une intégration par parties :

$$\begin{aligned}\forall p, q \in \mathbb{R}[T],\ \ (\,Dp \mid q\,)_w = (\,w^{-1}\big(wQp'\big)' \mid q\,)_w &= \int_\alpha^\beta \big(w(t)Q(t)p'(t)\big)'q(t)\,\mathrm{d}t \\ &= [w(t)Q(t)p'(t)q(t)]_\alpha^\beta - \int_\alpha^\beta w(t)Q(t)p'(t)q'(t)\,\mathrm{d}t.\end{aligned}$$

En une borne infinie de I, la fonction $t \mapsto w(t)Q(t)p'(t)q(t)$ admet une limite, puisque :

$$[w(t)Q(t)p'(t)q(t)]_{t_0}^\beta = \int_{t_0}^\beta Dp(t)q(t)w(t)\,\mathrm{d}t + \int_{t_0}^\beta w(t)Q(t)p'(t)q'(t)\,\mathrm{d}t,$$

et que les deux intégrales existent. La convergence de $\displaystyle\int_\alpha^\beta w(t)Q(t)p'(t)q(t)\,\mathrm{d}t$ implique que cette limite est nulle. En une borne finie, il en est de même puisque nous avons supposé que la limite de $w(t)Q(t)$ est nulle. Ainsi, $[w(t)Q(t)p'(t)q(t)]_\alpha^\beta = 0$ et :

$$(\,Dp \mid q\,)_w = -(\,Q \mid p'q'\,)_w = (\,p \mid Dq\,)_w. \tag{21}$$

L'opérateur D sur l'espace préhilbertien réel $\big(\mathbb{R}[T], (\ \mid\ )_w\big)$ est donc *auto-adjoint*. Les valeurs propres *possibles* $\lambda_n := n(n\gamma - \gamma + \eta)$, $n \in \mathbb{N}$, sont *distinctes* pour des $n \in \mathbb{N}$ distincts : si $n \neq m$, et si $\lambda_n = \lambda_m$, on a $(n-m)[\gamma(n+m-1)+\eta] = 0$ ce qui contredit nos hypothèses (*cf.* ci-dessus, la condition (3)).
En restreignant D aux sous-espaces $\mathbb{R}_n[T]$ de $\mathbb{R}[T]$ ($n \in \mathbb{N}$), on obtient des opérateurs D_n auto-adjoints (pour la restriction de $(\ \mid\ )_w$).

L'opérateur symétrique D_n admet une base de vecteurs propres. Soit p un vecteur propre de degré $k \leqslant n$, il est associé à la valeur propre $\lambda_k = k\big((k-1)\gamma + \eta\big)$ (*cf.* (19)). Le sous-espace propre associé à cette valeur propre est nécessairement de dimension un. Sinon, il existerait, dans ce sous-espace, deux polynômes unitaires distincts p_1 et p_2, $p_1 - p_2$ serait de degré $h < k$ et l'on aurait $\lambda_h = \lambda_k$, ce qui est impossible. Le résultat s'en déduit. ■

Quitte à faire une transformation affine convenable sur t ($t \mapsto ct + d$) pour transformer l'intervalle J de la proposition précédente, et à multiplier, si nécessaire, D par un réel strictement positif, on peut (le lecteur le vérifiera facilement) supposer que l'on est dans l'un des *trois* cas suivants :

1. $Q(t) = -t^2 + 1$, $J =]-1, 1[$; on obtiendra les polynômes de Jacobi.
2. $Q(t) = t$, $J =]0, +\infty[$; on obtiendra les polynômes de Laguerre (généralisés).
3. $Q(t) = 1$, $J = \mathbb{R}$; on obtiendra les polynômes d'Hermite.

- On élimine les cas $Q(t) = 1 - t^2$, $J =]-\infty, -1[$ et $Q(t) = t^2 - 1$, $J =]1, +\infty[$, car ils conduisent à des poids de masse totale infinie (le lecteur pourra s'en assurer à titre d'exercice).
- Il reste à examiner le cas où le polynôme $Q(t)$ admet un *zéro double*. Ce cas se ramène à :
$$Q(t) = t^2, \quad J =]0, +\infty[, \quad L(t) = 2t + a \quad \text{et} \quad w(t) = t^{a-2}\,\mathrm{e}^{-2/t} \quad (a \in \mathbb{R}).$$
Il faut donc l'éliminer, car $\mu_k = +\infty$, si $k > -a$. Toutefois, il y a un phénomène *a priori* étonnant : si l'on « passe outre » la divergence des moments de μ_0 et si l'on applique « formellement » le mécanisme que nous allons mettre en place pour les polynômes orthogonaux classiques, on voit apparaître des familles intéressantes de polynômes ; par exemple, pour $L(t) = 2t + 2$, la famille des polynômes dits *de Bessel*. Ces derniers vérifient, tout comme les polynômes orthogonaux classiques, des équations différentielles $Qy'' + Ly' = \lambda_n y$, des relations de récurrence d'ordre deux et une formule d'Olinde Rodrigues. Ils ont aussi de nombreuses applications (en combinatoire, en théorie du signal, en théorie des ondes…). Ils font l'objet de l'exercice III.4.9.

Montrons maintenant que dans les trois cas qui nous restent, les conditions du théorème 25 de la page précédente sont satisfaites (nous modifierons l'intervalle ouvert J en le remplaçant par $I := [0, +\infty[$ dans le second cas).

Cas 1 : Polynômes de Jacobi

On a $Q(t) = -t^2 + 1$ ($\gamma = -1, \gamma' = 0, \gamma'' = 1$), $J =]-1, 1[$, $L(t) = \eta t + \eta'$.

Alors $\dfrac{L(t)}{Q(t)} = \dfrac{1}{2}\left(\dfrac{\eta' + \eta}{1 - t} + \dfrac{\eta' - \eta}{1 + t}\right)$. On définit deux réels a et b, par :

$$1 + a = -\frac{\eta' + \eta}{2} \quad \text{et} \quad 1 + b = \frac{\eta' - \eta}{2}.$$

On a :

$$L(t) = \eta t + \eta' = -(a + b + 2)t + b - a \quad \text{et} \quad \frac{L(t)}{Q(t)} = \frac{b+1}{t+1} - \frac{a+1}{1-t}.$$

Par suite $\lambda_n = n(n\gamma - \gamma + \eta) = -n(n + 1 + a + b)$. On a :

$$\int \frac{L(t)}{Q(t)}\,\mathrm{d}t = \int \left(\frac{b+1}{t+1} - \frac{a+1}{1-t}\right)\mathrm{d}t = \ln(1+t)^{1+b} + \ln(1-t)^{1+a} + C.$$

On choisit la constante réelle C nulle. On a donc $\mathrm{e}^{\int \frac{L(t)}{Q(t)}\,\mathrm{d}t} = (1-t)^{1+a}(1+t)^{1+b}$ et :

$$w(t) = \frac{1}{Q(t)}\mathrm{e}^{\int \frac{L(t)}{Q(t)}\,\mathrm{d}t} = (1-t)^a(1+t)^b; \tag{22}$$

w est intégrable sur $]-1,1[$ si, et seulement si, $a > -1$ et $b > -1$. On est, par définition, dans le cas *de Jacobi*.

Vérifions les conditions du théorème 25 de la page 567 :

(i) La fonction w est clairement strictement positive et continue sur $J :=]-1,1[$. La masse totale de w est $\mu_0 = \displaystyle\int_{-1}^{1} w(t)\,\mathrm{d}t = \int_{-1}^{1}(1-t)^a(1+t)^b\,\mathrm{d}t$ (l'intégrale est convergente, puisque $a > -1$ et $b > -1$).

Plus généralement les intégrales $\displaystyle\int_{-1}^{1} t^n w(t)\,\mathrm{d}t = \int_{-1}^{1}(1-t)^a(1+t)^b t^n\,\mathrm{d}t$, qui définissent les moments μ_n sont convergentes pour tout $n \in \mathbb{N}$. Ainsi la fonction w est bien un poids sur $I := [-1,1]$.

(ii) On a $Q(t)w(t) = (1-t)^{1+a}(1+t)^{1+b}$. Cette fonction tend vers zéro quand la variable t tend vers ± 1, en restant dans l'intervalle $]-1,1[$.

(iii) On a $\gamma x + \eta = -x - a - b - 2$, donc $\gamma x + \eta = 0$ si, et seulement si, $x = -a-b-2 \in \mathbb{N}$, mais cela est impossible, puisque $-a-b-2 < 0$.

Ainsi la fonction w est bien un poids sur $J :=]-1,1[$. C'est, par définition, le poids définissant les polynômes de Jacobi $P_n^{(a,b)}$. (La notation traditionnelle est $P_n^{(\alpha,\beta)}$, nous l'avons un peu modifiée pour éviter toute confusion avec notre notation pour les bornes d'intervalles).

Les polynômes de Jacobi $P_n^{(a,b)}$ sont normalisés par la condition :

$$P_n^{(a,b)}(1) = \frac{\Gamma(n+1+a)}{n!\Gamma(1+a)} = \frac{(a+n)\cdots(a+1)}{n!}\cdot$$

L'équation différentielle de Jacobi est :

$$(1-t^2)y'' + [b-a-(a+b+2)t]y' + n(n+1+a+b)y = 0. \tag{23}$$

Pour $a = b = 0$, on obtient l'équation différentielle de Legendre :

$$(1-t^2)y'' - 2ty' + n(n+1)y = 0. \tag{24}$$

Cas 2 : polynômes de Laguerre généralisés

On a $Q(t) = t$ ($\gamma = -1, \gamma' = 0, \gamma'' = 1$), $J =]0,+\infty[$. On prend $I = [\alpha,\beta[= [0,+\infty[$. Si ω est un nombre réel strictement positif, en changeant D en $\omega^2 D$ et t en $\omega t = u$, on obtient :

$$u\frac{\mathrm{d}^2 y}{\mathrm{d}u^2} + \omega(\eta t + \eta')\frac{\mathrm{d}y}{\mathrm{d}u} = \omega^2\lambda y.$$

Si $\eta \neq 0$, on peut choisir $\omega = \frac{1}{|\eta|}\cdot$ On peut donc se limiter au cas où $\eta = 0$ ou $\eta = \pm 1$. Le lecteur vérifiera que les cas $\eta = 1$ et $\eta = 0$ ne peuvent pas conduire à un poids : l'intégrale

de w correspondante diverge. Il reste donc le cas $L(t) = -t+\eta'$. Nous poserons $\eta' := a+1$ (pour nous rapprocher de la notation traditionnelle, qui est $\eta' = \alpha + 1$). On a :

$$L(t) = -t + 1 + a \quad \text{et} \quad \frac{L(t)}{Q(t)} = -1 + \frac{1+a}{t}\cdot$$

On a $\gamma = \eta = -1$, donc $\lambda_n = n(n\gamma - \gamma + \eta) = -n^2$.
Avec un choix convenable de constante d'intégration, on trouve :

$$w(t) = \frac{1}{Q(t)} \mathrm{e}^{\int \frac{L(t)}{Q(t)}\,\mathrm{d}t} = t^a \mathrm{e}^{-t}; \tag{25}$$

w est intégrable sur $[0, +\infty[$ si, et seulement si, $a > -1$. On est, par définition, dans le cas *de Laguerre généralisé*.

Vérifions les conditions du théorème 25 de la page 567.
(i) La fonction w est clairement strictement positive et continue sur $J :=]0, +\infty[$. La masse totale de w est $\mu_0 = \displaystyle\int_0^{+\infty} t^a e^{-t}\,\mathrm{d}t = \Gamma(a+1)$.
Plus généralement les intégrales $\displaystyle\int_0^{+\infty} t^n w(t)\,\mathrm{d}t = \int_0^{+\infty} t^{n+a}\mathrm{e}^{-t}\,\mathrm{d}t$, qui définissent les moments μ_n sont convergentes pour tout $n \in \mathbb{N}$.
(ii) On a $Q(t)w(t) = t^{1+a}\mathrm{e}^{-t}$. Cette fonction tend vers 0 quand la variable t tend vers 0, en restant dans l'intervalle $[0, +\infty[$.
(iii) On a $\gamma x + \eta = -1$, et $-1 \notin \mathbb{N}$.
Ainsi la fonction $w(t) = t^a\mathrm{e}^{-t}$ est bien un poids sur $I := [0, +\infty[$. C'est le poids définissant les *polynômes de Laguerre généralisés* L_n^a. Les *polynômes de Laguerre* proprement dits L_n correspondent au cas $a = 0$, c'est-à-dire à $w(t) = \mathrm{e}^{-t}$.
L'équation différentielle de Laguerre généralisée est :

$$ty'' + (1 + a - t)y' + ny = 0. \tag{26}$$

Cas 3 : polynômes d'Hermite

On a $Q(t) = 1$ ($\gamma = 0, \gamma' = 0, \gamma'' = 1$), $J = I = \mathbb{R}$.
Considérons $L(t) = \eta t+\eta'$. On a $L/Q = L$. Le cas $\eta' \neq 0$ ne peut pas conduire à un poids (l'intégrale correspondant à la masse totale de w diverge). Donc $\eta' = 0$. Par un argument du même type que ci-dessus (utilisation d'un réel ω), on peut se se ramener au cas $\eta = -2$. Ainsi,

$$L(t) = -2t \quad \text{et} \quad \frac{L(t)}{Q(t)} = -2t.$$

On a $\gamma = 0,\ \eta = -2$, donc $\lambda_n = n(n\gamma - \gamma + \eta) = -2n$.
Avec un choix convenable de constante d'intégration on trouve :

$$w(t) = \frac{1}{Q(t)} \mathrm{e}^{\int \frac{L(t)}{Q(t)}\,\mathrm{d}t} = \mathrm{e}^{-t^2}. \tag{27}$$

Vérifions les conditions du théorème 25 de la page 567 :
(i) La fonction w est clairement strictement positive et continue sur $I = \mathbb{R}$. La masse totale de w est $\mu_0 = \displaystyle\int_{-\infty}^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = \sqrt{\pi}$.
(ii) Cette condition est ici vide.
(iii) On a $\gamma x + \eta = -2$ et $-2 \notin \mathbb{N}$.
Ainsi la fonction w est bien un poids sur $I = \mathbb{R}$. C'est le poids définissant les *polynômes d'Hermite* H_n.
L'équation différentielle d'Hermite est :

$$y'' - 2ty' + 2ny = 0. \tag{28}$$

On peut l'écrire : $Dy = -2ny$, avec $D = \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^2 - 2x\frac{\mathrm{d}}{\mathrm{d}x}$. En d'autres termes, le polynôme d'Hermite H_n est une fonction propre de l'opérateur différentiel D associée à la valeur propre $-2n$.

Exercice 6.

Montrer directement, sans utiliser la théorie précédente, que, pour tout $n \in \mathbb{N}$, l'équation différentielle :

$$(1-t^2)y'' + [b - a - (a+b+2)t]y' + n(n+1+a+b)y = 0$$

admet une solution polynomiale non triviale. Cette solution est de degré n, et elle est unique à multiplication par un réel non nul près.

Solution. Soit $n \in \mathbb{N}$ fixé. Cherchons, pour $\lambda \in \mathbb{R}$, une solution de l'équation différentielle :

$$(1-t^2)y'' + [b - a - (a+b+2)t]y' + \lambda y = 0.$$

sous la forme d'une série *formelle* $\sum\limits_{h\in\mathbb{N}} a_h(t-1)^h$ ($a_h \in \mathbb{R}\ \forall\, h \in \mathbb{N}$). On obtient, par substitution, les conditions *nécessaires* :

$$-(t+1)\sum_{h\geqslant 2} h(h-1)a_h(t-1)^{h-1} - \big[2(a+1) + (a+b+2)(t-1)\big]\sum_{h\geqslant 1} a_h(t-1)^{h-1}$$
$$+\lambda\sum_{h\geqslant 0} a_h(t-1)^h = 0,$$

puis, par identification des coefficients des puissances successives de $(t-1)$, les conditions nécessaires : $\big[\lambda - h(h+1+a+b)\big]a_h = 2(h+1)(h+a+1)a_{h+1}$, pour tout $h \in \mathbb{N}$. Ainsi la série formelle $\sum\limits_{h\in\mathbb{N}} a_h(t-1)^h$ est un *polynôme non nul* si, et seulement s'il existe un entier $n \in \mathbb{N}$, tel que $\lambda = n(n+1+a+b)$. S'il en est ainsi $\lambda - h(h+1+a+b) \neq 0$, pour $h \in [\![1, n-1]\!]$, et l'on peut déterminer les a_h par récurrence à partir de a_0, en utilisant :

$$\frac{a_{h+1}}{a_h} = \frac{(n-h)(n+h+a+b+1)}{2(h+1)(h+a+1)}. \tag{29}$$

La solution polynomiale est non nulle si, et seulement si, $a_0 \neq 0$, et dans ce cas $a_n \neq 0$. Les solutions polynomiales forment une droite réelle.

Posons $Dy = (1-t^2)y'' + [b-a-(a+b+2)t]y'$. Nous avons vu (*cf.* 21) que, pour tous $p, q \in \mathbb{R}[T]$, $(\,Dp \mid q\,)_w = (\,p \mid Dq\,)_w$. Si $n, m \in \mathbb{N}$, avec $n \neq m$:

$$0 = (\,Dp_n \mid p_m\,)_w - (\,p_n \mid Dp_m\,)_w = (\lambda_n - \lambda_m)(\,p_n \mid p_m\,),$$

et, puisque $\lambda_n \neq \lambda_m$, on a $(\,p_n \mid p_m\,) = 0$.

On peut faire la même chose dans les cas d'Hermite et de Laguerre (*cf.* l'exercice III.4.12).

Exercice 7.

Soit p_n un polynôme orthogonal de degré n pour le poids de Jacobi de paramètres (a, b), tel que $p_n(1) = 1$. Montrer que le monôme de degré n de p_n a pour coefficient :

$$2^{-n}\frac{\Gamma(2n+a+b+1)\Gamma(a+1)}{\Gamma(n+a+b+1)\Gamma(n+a+1)}. \tag{30}$$

En déduire le coefficient du monôme de plus haut degré du polynôme de Jacobi $P_n^{(a,b)}$.

Solution. On note $p_n(t) := a_n(t-1)^n + \cdots + a_1(t-1) + a_0$ ($a_0 := 1$). En utilisant la relation 29, on obtient :

$$\begin{aligned} a_n = \frac{a_n}{a_0} = \frac{a_n}{a_{n-1}} \cdots \frac{a_1}{a_0} &= \frac{(2n+a+b)(2n+a+b-1)\cdots(n+a+b+1)n!}{2^n n!(n+a)\cdots(a+1)} \\ &= 2^{-n}\frac{\Gamma(2n+a+b+1)\Gamma(a+1)}{\Gamma(n+a+b+1)\Gamma(n+a+1)}. \end{aligned}$$

Le coefficient du monôme de degré n (en t) de p_n est a_n.

En utilisant $P_n^{(a,b)}(1) = \binom{a+n}{n} = \dfrac{\Gamma(a+n+1)}{n!\Gamma(a+1)}$, on en déduit que le coefficient du monôme de plus haut degré de $P_n^{(a,b)}$ est :

$$\begin{aligned} 2^{-n}\frac{\Gamma(2n+a+b+1)\Gamma(a+1)}{\Gamma(n+a+b+1)\Gamma(n+a+1)}\frac{\Gamma(a+n+1)}{n!\Gamma(a+1)} &= \frac{\Gamma(2n+a+b+1)}{n!\Gamma(n+a+b+1)} \\ &= 2^{-n}\binom{2n+a+b}{n}. \end{aligned}$$

L'énoncé suivant résume les résultats obtenus dans les cas (i), (ii) et (iii).

Théorème 26. (i) Soient $a, b \in \mathbb{R}$, avec $a, b > -1$.
Si $w(t) = (1-t)^a(1+t)^b$ sur $]-1, 1[$, pour tout $n \in \mathbb{N}$, le polynôme de Jacobi $P_n^{(a,b)}$ vérifie l'équation différentielle *de Jacobi* d'indice n :

$$(1-t^2)y'' + [b - a - (a+b+2)t]y' + n(n+1+a+b)y = 0.$$

(ii) Soit $a > -1$. Si $w(t) = t^a \mathrm{e}^{-t}$ sur $[0, +\infty[$, pour tout $n \in \mathbb{N}$, le polynôme de Laguerre généralisé L_n^a vérifie l'équation différentielle *de Laguerre généralisée* d'indice n :

$$ty'' + (1+a-t)y' + ny = 0.$$

(iii) Si $w(t) = \mathrm{e}^{-t^2}$ sur $\mathbb{R}$, pour tout $n \in \mathbb{N}$, le polynôme d'Hermite H_n vérifie l'équation différentielle *d'Hermite* d'indice n :

$$y'' - 2ty' + 2ny = 0.$$

Le polynôme dérivé d'un polynôme orthogonal classique, est encore un polynôme orthogonal classique.

Proposition 27. Soit p_n un polynôme orthogonal classique, correspondant au cas (Q, L, w). Alors le polynôme dérivé p_n' est aussi un polynôme orthogonal classique, correspondant au cas $(Q, L_1 := L + Q', w_1 := Qw)$. Si λ_n est la valeur propre correspondant à p_n, celle correspondant à p_n' est $\lambda_n - L'$.

Démonstration. On dérive l'équation différentielle $Qy'' + Ly' = \lambda_n y$, dont p_n est solution. On obtient $Qy''' + (Q' + L)y'' = (\lambda_n - L')y'$, d'où le résultat ($Q'$ est de degré 1 et L' est une constante). ∎

Notons que le cas des polynômes d'Hermite est tout à fait exceptionnel. On a $Q = 1$, donc $w_1 = Qw = w$: en dérivant, on reste donc dans la même famille. Nous reviendrons plus loin sur ce point, dans le paragraphe consacré aux polynômes d'Hermite (*cf.* ci dessous, la proposition 47 (ii), page 592). Cela est lié au mécanisme des *opérateurs de création annihilation*.
Signalons qu'il est utile, dans diverses questions de mathématique, physique, informatique… de considérer des opérateurs linéaires « fonctionnels » plus généraux que les opérateurs différentiels, par exemple en « discrétisant » $\mathrm{d}/\mathrm{d}t$ et en le remplaçant par $\Delta f(t) = f(t+1) - f(t)$ (opérateurs aux différences) ou $\delta_q f(t) = \frac{f(qt) - f(t)}{(q-1)t}$ (opérateurs aux q-différences, dits aussi quantiques). Il est donc naturel d'étendre les considérations ci-dessus à de tels opérateurs, pour faire apparaître des variantes discrètes des polynômes orthogonaux classiques. On peut remplacer, par exemple, le problème différentiel aux valeurs propres : $Q\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 y + L\frac{\mathrm{d}}{\mathrm{d}t}y = \lambda y$ par le problème aux différences analogue : $Q\Delta^2 y + L\Delta y = \lambda y$, et développer des variations *discrètes* de la théorie des polynômes orthogonaux classiques. On obtient les polynômes de Hahn, Tchebychev discrets, Markov, Poisson-Charlier, Meixner, Krawtchouk… ; certains de ces polynômes sont associés à des distributions de probabilités classiques : distribution de Poisson, distribution binomiale… (*cf.* l'exercice III.4.10).

Ce sujet de recherches est encore très actif[11], avec de multiples retombées : relations avec la combinatoire, les probabilités, la théorie des représentations, la mécanique quantique, la compression de l'information, la discrétisation des équations de la physique mathématique....

3.2 Formule d'Olinde Rodrigues

Benjamin Olinde Rodrigues (1794-1851) était un mathématicien, banquier et utopiste saint-simonien français. Il a démontré la formule qui porte son nom dans sa thèse en 1816 pour le cas des polynômes de Legendre.

La *formule d'Olinde Rodrigues* est une formule simple, permettant de calculer directement une suite de polynômes orthogonaux $(p_n)_{n\in\mathbb{N}}$, *dans le cas des polynômes orthogonaux classiques* (et dans ce cas seulement) en fonction du poids w et du polynôme Q :

$$p_n(t) := \frac{1}{w}\frac{\mathrm{d}^n}{\mathrm{d}t^n}(Q^n w).$$

Cette formule a de nombreuses applications, le calcul explicite des formules de récurrence par exemple. Elle permet aussi de retrouver les équations différentielles. Elle est parfois utilisée comme *définition* des familles classiques de polynômes orthogonaux.

La formule d'Olinde Rodrigues repose sur une application répétée de la formule d'intégration par parties dans le cas où tous les termes convergent et où les termes tout intégrés apparaissant successivement *sont tous nuls*.

Avec ces hypothèses, on a (pour des fonctions f, g de classe $\mathcal{C}^\infty$), la formule d'intégration par parties généralisée :

$$\int_\alpha^\beta f^{(n)}(t)g(t)\,\mathrm{d}t = -\int_\alpha^\beta f^{(n-1)}(t)g'(t)\,\mathrm{d}t = \cdots = (-1)^n\int_\alpha^\beta f(t)g^{(n)}(t)\,\mathrm{d}t. \quad (31)$$

Les termes tout intégrés sont, au signe près :

$$\Big[f^{(n-1)}(t)g(t)\Big]_\alpha^\beta, \ \dots, \ \Big[f(t)g^{(n-1)}(t)\Big]_\alpha^\beta.$$

Dans le cas où cette formule peut s'appliquer à une fonction f et à $g(t) = 1, t, \dots, t^{n-1}$, on obtient $\displaystyle\int_\alpha^\beta f^{(n)}(t)t^k\,\mathrm{d}t = (-1)^n\int_\alpha^\beta f(t)(t^k)^{(n)}\,\mathrm{d}t = 0$, pour tout $k \in [\![0, n-1]\!]$.

On a donc $0 = (\,f^{(n)} \mid t^k\,) = (\,w^{-1}f^{(n)} \mid t^k\,)_w$ pour tout $k \in [\![0, n-1]\!]$.

Cela suggère le procédé suivant de construction d'une suite de polynômes orthogonaux. Supposons que, pour tout $n \in \mathbb{N}$, on ait une fonction v_n, telle que l'on puisse appliquer la formule d'intégration par parties généralisée à $f = v_n$ et à tout $g \in \mathbb{R}_{n-1}[T]$ (si $[\alpha, \beta]$ est un segment, la condition $\frac{\mathrm{d}^m}{\mathrm{d}t^m}v_n(\alpha) = \frac{\mathrm{d}^m}{\mathrm{d}t^m}v_n(\beta) = 0$ pour tout $m \in [\![1, n-1]\!]$ suffit). On aura alors $(\,w^{-1}v_n^{(n)} \mid g\,)_w = 0$ pour tout $g \in \mathbb{R}_{n-1}[T]$. Si, de plus, $w^{-1}v_n^{(n)}$ est un polynôme de degré n, on pourra définir p_n par $p_n := w^{-1}v_n^{(n)}$.

Le poids w est une solution de l'équation résolvante $(wQ)' = wL$. Un polynôme orthogonal p_1, de degré un, vérifie l'équation différentielle $Qp_1'' + Lp_1' = \lambda_1 p_1$, soit $Lp_1' = \lambda_1 p_1$.

[11] Le lecteur intéressé pourra consulter la monographie de Mourad E. H. Ismail, *Classical and Quantum Orhogonal Polynomials in One Variable,* Cambrodge University Press.

Ainsi L et p_1 sont proportionnels, on peut choisir $L = p_1$ et écrire l'équation résolvante $wp_1 = (wQ)'$, ou, de façon équivalente :

$$p_1 = w^{-1}\frac{\mathrm{d}}{\mathrm{d}t}(Qw). \tag{32}$$

On peut prendre $v_1 = Qw$. C'est la première étape de la construction. Pour la poursuivre, nous allons d'abord nous placer, comme Olinde Rodrigues, *dans le cas de Legendre*.

3.2.1 Cas des polynômes de Legendre

On a $w = 1$, on cherche donc $v_n \in \mathbb{R}[T]$, de degré $2n$. Il faut pouvoir justifier la formule d'intégration par parties généralisée. On travaille sur un segment $I = [-1, 1]$ et toutes les fonctions sont continues, les intégrales existent donc. Par ailleurs, les termes tout intégrés sont, au signe près $[v_n^{(n-1)}(t)g(t)]_{-1}^1, \ldots, [v_n(t)g^{n-1}(t)]_{-1}^1$.

La réponse *possible* la plus simple est donc $v_n := (t^2-1)^n = \big(-Q(t)\big)^n$. Nous allons voir qu'elle convient. Cela résultera du lemme suivant, qui est facile.

Lemme 28. Soient $n \in \mathbb{N}$ et $m \in [\![0, n]\!]$. Il existe un polynôme $r_{n,m} \in \mathbb{R}[T]$ tel que $\frac{\mathrm{d}^m}{\mathrm{d}t^m}\left((t^2-1)^n\right) = (t^2-1)^{n-m}r_{n,m}$. Si $m \leqslant n-1$, le polynôme $\frac{\mathrm{d}^m}{\mathrm{d}t^m}\left((t^2-1)^n\right)$ s'annule en -1 et 1.

Proposition 29.

(i) Les polynômes $p_n := \dfrac{\mathrm{d}^n}{\mathrm{d}t^n}\left((t^2-1)^n\right)$, $n \in \mathbb{N}$, forment une suite de polynômes orthogonaux pour le poids $w = 1$ sur le segment $[-1, 1]$.

(ii) On a la *formule d'Olinde Rodrigues pour les polynômes de Legendre* :

$$P_n(t) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left((t^2-1)^n\right).$$

Démonstration.

(i) D'après le lemme ci-dessus, on peut appliquer la formule d'intégration par parties généralisée à $v_n = (t^2-1)^n$ et à tout $g \in \mathbb{R}_{n-1}[T]$. On a donc $(v_n^{(n)} \mid g) = (v_n \mid g^{(n)}) = 0$ pour tout $g \in \mathbb{R}_{n-1}[T]$. Ainsi $(p_n := v_n^{(n)})_{n\in\mathbb{N}}$ est une suite de polynômes orthogonaux pour le poids $w = 1$ sur le segment $[-1, 1]$.

(ii) On utilise la formule de Leibniz pour calculer :

$$p_n = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\left((t^2-1)^n\right) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\left((t-1)^n(t+1)^n\right),$$

puis l'on fait $t = 1$. Pour $0 \leqslant p < n$, le polynôme $\frac{\mathrm{d}^p}{\mathrm{d}t^p}\left((t-1)^n\right)$ s'annule en $t = 1$. Par suite tous les termes s'annulent en $t = 1$, sauf $\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left((t-1)^n\right)(t+1)^n = n!(t+1)^n$. On obtient donc $p_n(1) = 2^n n!$. Comme, par définition, $P_n(1) = 1$, l'on en déduit $p_n = 2^n n! P_n$. ∎

3.2.2 Cas des polynômes de Jacobi

Passons au cas général pour les polynômes de Jacobi. Compte tenu de ce qui précède, il est naturel de poser $v_n = Q^n w = (1-t)^{a+n}(1+t)^{b+n}$ et de vérifier les propriétés attendues. C'est l'objet du lemme suivant.

Lemme 30. Soient $n \in \mathbb{N}^*$ et $m \in [\![0, n]\!]$.

(i) Il existe un polynôme $r_{n,m}^{a,b} \in \mathbb{R}[T]$, *de degré* m, tel que les trois propriétés suivantes soient vérifiées :

(α) on a $\frac{\mathrm{d}^m}{\mathrm{d}t^m} v_n = \frac{\mathrm{d}^m}{\mathrm{d}t^m}[Q^n w] = r_{n,m}^{a,b} Q^{n-m} w$;

(β) la fonction $p_n := w^{-1} \frac{\mathrm{d}^n}{\mathrm{d}t^n}[Q^n w]$ est un *polynôme de degré* n, on a :

$$p_n = (-1)^n n! \sum_{h=0}^{n} \binom{n+a}{h} \binom{n+b}{n-h} (t-1)^{n-h}(t+1)^h;$$

(γ) on a $p_0(1) = 1$ et, pour tout $n \in \mathbb{N}^*$:

$$p_n(1) = (-2)^n (a+n)\dots(a+1) \quad \text{et} \quad p_n(-1) = 2^n (b+n)\dots(b+1).$$

(ii) fonction $t \in]-1, 1[\mapsto \frac{\mathrm{d}^m}{\mathrm{d}t^m}[Q(t)^n w(t)]$ tend vers 0 en -1 et 1.

Démonstration. On montre (α) en utilisant la formule de Leibniz.
On a, pour tout $m \in [\![0, n]\!]$:

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m}\left[(1-t)^{a+n}(1+t)^{b+n}\right] = \sum_{h=0}^{m} \binom{m}{h} \frac{\mathrm{d}^h}{\mathrm{d}t^h}\left[(1-t)^{a+n}\right] \frac{\mathrm{d}^{m-h}}{\mathrm{d}t^{m-h}}\left[(1+t)^{b+n}\right]. \quad (33)$$

Pour tout $h \in [\![0, m]\!]$:

$$\frac{\mathrm{d}^h}{\mathrm{d}t^h}\left[(1-t)^{a+n}\right] = (-1)^h h! \binom{n+a}{h} (1-t)^{a+n-h}$$

et :

$$\frac{\mathrm{d}^{m-h}}{\mathrm{d}t^{m-h}}\left[(1+t)^{b+n}\right] = (m-h)! \binom{n+b}{m-h} (1+t)^{b+n+h-m},$$

donc :

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m}\left[(1-t)^{a+n}(1+t)^{b+n}\right] = m! \sum_{h=0}^{m} (-1)^h \binom{n+a}{h} \binom{n+b}{m-h} (1-t)^{a+n-h}(1+t)^{b+n+h-m}.$$

On définit, pour tout $n \in \mathbb{N}^*$ et $m \in [\![0, n]\!]$ le polynôme :

$$r_{n,m}^{a,b} := m! \sum_{h=0}^{m} (-1)^h \binom{n+a}{h} \binom{n+b}{m-h} (1-t)^{m-h}(1+t)^h.$$

De :

$$\begin{aligned}(1-t)^{m-h}(1+t)^h Q^{n-m} w &= (1-t)^{m-h}(1+t)^h (1-t)^{n-m+a}(1+t)^{n-m+b} \\ &= (1-t)^{a+n-h}(1+t)^{b+n+h-m},\end{aligned}$$

l'on déduit :

$$\frac{\mathrm{d}^m}{\mathrm{d}t^m}\left[(1-t)^{a+n}(1+t)^{b+n}\right] = r_{n,m}^{a,b} Q^{n-m} w.$$

Le coefficient du terme de plus haut degré de $r_{n,m}^{a,b}$ est :

$$(-1)^m m!\, C \qquad \text{avec} \qquad C := \sum_{h=0}^{m} \binom{n+a}{h}\binom{n+b}{m-h}.$$

Puisque $a > -1$ et $b > -1$, on a $\binom{n+a}{h} > 0$ et $\binom{n+b}{m-h} > 0$, donc $C > 0$.

Par suite $(-1)^m m!\, C \neq 0$ et le polynôme $r_{n,m}^{a,b}$ est de degré m.

En prenant $m = n$, l'on en déduit (β) $(p_n = r_{n,n})$.

On a :

$$p_n = (-1)^n n! \sum_{h=0}^{n} \binom{n+a}{h}\binom{n+b}{n-h}(t-1)^{n-h}(t+1)^h, \tag{34}$$

donc $p_n(1) = (-2)^n n! \binom{a+n}{n}$ et $p_n(-1) = 2^n n! \binom{b+n}{n}$.

Prouvons (ii).

Pour tout $m \in [\![0, n-1]\!]$, la limite de $Q^{n-m}(t)w(t) = (1-t)^{a+n-m}(1+t)^{b+n-m}$, quand $t \in]-1, 1[$ tend vers -1 ou 1, est nulle. Par suite, d'après (α), il en est de même pour celle de $\frac{\mathrm{d}^m}{\mathrm{d}t^m} Q(t)^n w(t)$, si $m \in [\![0, n-1]\!]$. Dans la suite, on prolonge ces fonctions par continuité en -1 et 1. ∎

Proposition 31.

(i) Soit $w(t) = (1-t)^a(1+t)^b$, avec $a, b \in \mathbb{R}$, $a,\ b > -1$. Les polynômes :

$$\begin{aligned} p_n(t) &:= w^{-1}(t)\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left[(1-t^2)^n w(t)\right] \\ &= (1-t)^{-a}(1+t)^{-b}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left[(1-t)^{a+n}(1+t)^{b+n}\right] \end{aligned}$$

forment une suite de polynômes orthogonaux pour le poids w sur le segment $[-1, 1]$.

(ii) On a la *formule d'Olinde Rodrigues pour les polynômes de Jacobi* :

$$P_n^{(a,b)} = \frac{(-1)^n}{2^n n!}(1-t)^{-a}(1+t)^{-b}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left[(1-t)^{a+n}(1+t)^{b+n}\right].$$

(iii) On a la *formule explicite pour les polynômes de Jacobi* :

$$P_n^{(a,b)} = \frac{1}{2^n}\sum_{h=0}^{n}\binom{n+a}{h}\binom{n+b}{n-h}(t-1)^{n-h}(t+1)^h.$$

(iv) Le coefficient du terme de plus haut degré de $P_n^{(a,b)}$ est $\frac{1}{2^n}\binom{2n+a+b}{n}$.

Démonstration.

(i) D'après l'assertion (ii) du lemme 30, on peut appliquer la formule d'intégration par parties généralisée à $v_n = (t^2-1)^n$ et à tout $g \in \mathbb{R}_{n-1}[T]$.

On a donc $(\, w^{-1}v_n^{(n)} \mid g\,)_w = (\, v_n^{(n)} \mid g\,) = (\, v_n \mid g^{(n)}\,) = 0$ pour tout $g \in \mathbb{R}_{n-1}[T]$.

D'après l'assertion (β) du lemme 30, les $p_n := w^{-1}v_n^{(n)}$ $(n \in \mathbb{N})$ sont des polynômes, ils forment donc une suite de polynômes orthogonaux pour le poids w sur le segment $[-1,1]$.

(ii) Nous avons vu que $p_n(1) = (-2)^n n!\binom{a+n}{n}$ et l'on a normalisé les polynômes de Jacobi par $P_n^{(a,b)}(1) = \binom{a+n}{n}$. On a donc, pour tout $n \in \mathbb{N}$, $P_n^{(a,b)} = \frac{(-1)^n}{2^n n!}p_n$.

(iii) D'après (34), on a :

$$p_n = (-1)^n n! \sum_{h=0}^{n} \binom{n+a}{h}\binom{n+b}{n-h}(t-1)^{n-h}(t+1)^h,$$

donc :

$$P_n^{(a,b)} = \frac{1}{2^n}\sum_{h=0}^{n} \binom{n+a}{h}\binom{n+b}{n-h}(t-1)^{n-h}(t+1)^h.$$

(iv) Le coefficient h_n du terme de plus haut degré de $P_n^{(a,b)}$ est $\frac{1}{2^n}\sum\limits_{h=0}^{n}\binom{n+a}{h}\binom{n+b}{n-h}$.

D'après l'exercice III.4.13, on a $\sum\limits_{h=0}^{n}\binom{n+a}{h}\binom{n+b}{n-h} = \binom{2n+a+b}{n}$.

Par suite $h_n = \frac{1}{2^n}\binom{2n+a+b}{n}$. On retrouve le résultat de l'exercice 7 de la page 572. ■

Remarque. En utilisant la formule explicite pour les polynômes de Jacobi, on obtient : $P_n^{(a,b)}(-t) = (-1)^n P_n^{(b,a)}(t)$.

3.2.3 Cas des polynômes de Laguerre (généralisés)

Pour les polynômes de Laguerre généralisés, de paramètre $a \in \mathbb{R}$, $a > -1$, on a $Q(t) = t$ et $w(t) = t^a \mathrm{e}^{-t}$. On pose donc :

$$v_n(t) := t^{n+a}\mathrm{e}^{-t} \quad \text{et} \quad p_n(t) := t^{-a}\mathrm{e}^t v_n(t) = t^{-a}e^t \frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(t^{n+a}e^{-t}\right).$$

Lemme 32. Soient $n \in \mathbb{N}^*$ et $m \in [\![0,n]\!]$.

(i) Il existe un polynôme $r_{n,m}^a \in \mathbb{R}[T]$, *de degré* m, tel que l'on ait les trois assertions suivantes :

(α) on a $\frac{\mathrm{d}^m}{\mathrm{d}t^m}v_n = \frac{\mathrm{d}^m}{\mathrm{d}t^m}\left(t^{n+a}\mathrm{e}^{-t}\right) = r_{n,m}^a t^{n-m+a}\mathrm{e}^{-t}$;

(β) la fonction $p_n : t \mapsto t^{-a}\mathrm{e}^t \frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(t^{n+a}\mathrm{e}^{-t}\right)$ est une *fonction polynôme de degré* n ;

(γ) le coefficient du monôme de plus haut degré $c_n t^n$ de p_n est $c_n = (-1)^n$.

(ii) Puisque $a > -1$, si $0 \leqslant m \leqslant n-1$, pour tout polynôme $g \in \mathbb{R}[T]$, la limite de la fonction $t \mapsto g(t)\frac{\mathrm{d}^m}{\mathrm{d}t^m}\left[t^{n+a}\mathrm{e}^{-t}\right]$, quand $t \in \,]0,+\infty[$ tend vers 0 ou $+\infty$, est *nulle*.

Démonstration.

(i) En utilisant la formule de Leibniz, on vérifie que le polynôme :

$$r_{n,m}^a := m! \sum_{h=0}^{n} \frac{(-1)^h}{h!} \binom{n+a}{m-h} t^h$$

répond à la question. On a, en particulier :

$$p_n = r_{n,n}^a := n! \sum_{h=0}^{n} \frac{(-1)^h}{h!} \binom{n+a}{n-h} t^h$$

(ii) Pour tout $m \in [\![0, n-1]\!]$, la limite de $g(t)t^{n-m+a}\mathrm{e}^{-t}$ quand $t \in]0, +\infty[$ tend vers 0 ou $+\infty$, est *nulle*. ■

En raisonnant comme dans le cas des polynômes de Jacobi, on en déduit le résultat suivant.

Proposition 33. Soit $w(t) = t^a\mathrm{e}^{-t}$ sur $[0, +\infty[$. Les polynômes :

$$p_n(t) := w^{-1}(t)\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(t^n w(t)\right) = t^{-a}\mathrm{e}^t \frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(t^{n+a}e^{-t}\right)$$

forment une suite de polynômes orthogonaux pour le poids w sur l'intervalle $[0, +\infty[$.

Les polynômes de Laguerre généralisés $L_n^{(a)}$ sont, par définition, associés au poids $w(t) = t^a\mathrm{e}^{-t}$ et normalisés par le coefficient de leur terme de plus haut degré, *que nous supposerons égal à* $(-1)^n$. Les polynômes unitaires correspondants sont donc les $(-1)^n L_n^{(a)}$.

De nombreux auteurs utilisent une autre normalisation pour les polynômes de Laguerre et les polynômes de Laguerre généralisés en choisissant le coefficient de leur terme de plus haut degré égal à $(-1)^n/n!$. On notera $\widetilde{L}_n := L_n/n!$ et $\widetilde{L}_n^{(a)}/n!$ les polynômes correspondants. On vérifiera plus loin que la famille $(\widetilde{L}_n)$ est orthonormée.

Corollaire 34 (Formule d'Olinde Rodrigues pour les polynômes de Laguerre généralisés).

(i) Soit $a \in \mathbb{R}$, avec $a > -1$. Le $n^{\text{ème}}$ polynôme de Laguerre généralisé est :

$$L_n^{(a)} = t^{-a}\mathrm{e}^t \frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(t^{n+a}e^{-t}\right).$$

Le $n^{\text{ème}}$ polynôme de Laguerre est $L_n = \mathrm{e}^t \frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(t^n\mathrm{e}^{-t}\right)$.

(ii) La formule explicite des polynômes de Laguerre généralisés est :

$$L_n^{(a)} = n! \sum_{h=0}^{n} \frac{(-1)^h}{h!} \binom{n+a}{n-h} t^h.$$

En particulier :

$$L_n = n! \sum_{h=0}^{n} \frac{(-1)^h}{h!} \binom{n}{h} t^h = (n!)^2 \sum_{h=0}^{n} \frac{(-1)^h}{(h!)^2(n-h)!} t^h$$

$$= (-1)^n \Big(t^n - \frac{n^2}{1!}t^{n-1} + \frac{n^2(n-1)^2}{2!}t^{n-2} + \cdots + (-1)^n n!\Big).$$

3.2.4 Cas des polynômes d'Hermite

Terminons avec le cas d'Hermite. On a $w(t) = \mathrm{e}^{-t^2}$ sur $\mathbb{R}$ et $Q(t) = 1$.
On pose donc $v_n(t) = \mathrm{e}^{-t^2}$ et $p_n(t) = e^{t^2} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \mathrm{e}^{-t^2}$.

En procédant comme dans les cas précédents on obtient le résultat suivant.

Proposition 35. Soit $w(t) = \mathrm{e}^{-t^2}$ sur $\mathbb{R}$. Les polynômes :

$$p_n(t) := w^{-1}(t) \frac{\mathrm{d}^n}{\mathrm{d}t^n}[w(t)] = \mathrm{e}^{t^2} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \mathrm{e}^{-t^2}$$

forment une suite de polynômes orthogonaux pour le poids w sur $\mathbb{R}$.

Les polynômes d'Hermite H_n sont, par définition, normalisés par le coefficient 2^n de leur monôme de plus haut degré.

Corollaire 36 (Formule d'Olinde Rodrigues pour les polynômes d'Hermite). Le $n^{ème}$ polynôme d'Hermite est $H_n(t) = (-1)^n \mathrm{e}^{t^2} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \mathrm{e}^{-t^2}$.

On rassemble les divers résultats en l'énoncé suivant.

Théorème 37 (Formule d'Olinde Rodrigues). Soit (Q, w) définissant une famille de polynômes orthogonaux *classiques*. Alors, $p_n := \frac{1}{w} \frac{\mathrm{d}^n}{\mathrm{d}t^n}(Q^n w)$, $n \in \mathbb{N}$, est une suite de polynômes orthogonaux pour le poids w.

Polynômes de Bessel.

On peut appliquer formellement la formule d'Olinde Rodrigues au cas que nous avons écarté plus haut (en constatant que certains moments de w divergent) : $Q(t) := t^2$, $L(t) := 2t + a$, $w(t) := t^{a-2} e^{-\frac{2}{t}}$. On obtient ainsi des suites de polynômes $p_n := t^{2-a} e^{\frac{2}{t}} \frac{\mathrm{d}^n}{\mathrm{d}t^n}(t^{2n-2+a} e^{-\frac{2}{t}})$.

Pour $a = 2$, on obtient, à normalisation près, les *polynômes* de Bessel $(y_n)_{n \in \mathbb{N}}$ (ils sont normalisés par $y_n(0) = 1$). Ces polynômes sont étudiés dans l'exercice III.4.9.

On montre que $y_n(t) = \sum_{k=0}^{n} \frac{(n+k)!}{(n-k)!k!} \left(\frac{t}{2}\right)^k$. Les 4 premiers polynômes sont :

$$y_0(t) = 1, \quad y_1(t) = 1 + t, \quad y_2(t) = 1 + 3t + 3t^2, \quad y_3(t) = 1 + 6t + 15t^2 + 15t^3.$$

Ces polynômes vérifient l'équation différentielle $t^2 y'' + (2t + 2)y' - n(n + 1)y = 0$ et la relation de récurrence $y_{n+1} - (2n + 1)ty_n - y_{n-1} = 0$. Ils ne sont pas orthogonaux pour un poids au sens classique, mais, comme polynômes en $z \in \mathbb{C}$, ils sont *orthogonaux sur le cercle unité* (pour la mesure $\mathrm{d}\theta$). Ces polynômes sont liés aux *fonctions de Bessel*. Ils ont de nombreuses applications (les coefficients de leurs termes dominants : 1, 1, 3, 15, 105, 945... comptent, par exemple, des arbres binaires stricts – i.e. à 0 ou 2 fils par nœud – à n feuilles.

3.3 Polynômes de Jacobi et cas particuliers (Legendre et Tchebychev)

3.3.1 Polynômes de Jacobi

Nous avons vu que les polynômes de Jacobi $P_n^{(a,b)}$ forment une suite de polynômes orthogonaux associés au poids $w(t) := (1-t)^a(1+t)^b$ ($a > -1$, $b > -1$) sur l'intervalle $]-1,1[$. Si $a = b = \lambda - 1/2$, on obtient (moyennant un changement de normalisation) les polynômes de Gegenbauer C_n^λ (que l'on appelle aussi ultrasphériques), que nous n'étudierons pas en général (*cf.* l'exercice III.4.22). Trois cas particuliers de ces polynômes sont particulièrement intéressants : les polynômes de Legendre P_n ($a = b = 0$, et donc $w = 1$) que nous connaissons déjà bien, les polynômes de Tchebychev T_n ($a = b = -1/2$) et les polynômes de Tchebychev de seconde espèce U_n ($a = b = 1/2$).

Proposition 38. Soient $a, b > -1$. On a, pour tous $n, m \in \mathbb{N}$:

$$\int_{-1}^{1} P_n^{(a,b)} P_m^{(a,b)} (1-t)^a (1+t)^b \mathrm{d}t = \frac{2^{a+b+1}\Gamma(a+n+1)\Gamma(n+b+1)}{n!(a+b+2n+1)\Gamma(a+b+n+1)}\delta_{nm}.$$

Démonstration. On suppose $n \geqslant m$. On a :

$$P_n^{(a,b)} = \frac{(-1)^n}{2^n n!}(1-t)^{-a}(1+t)^{-b}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\big((1-t)^{a+n}(1+t)^{b+n}\big),$$

donc :

$$\begin{aligned}\int_{-1}^{1} P_n^{(a,b)} P_m^{(a,b)} (1-t)^a (1+t)^b \,\mathrm{d}t &= \frac{(-1)^n}{2^n n!}\int_{-1}^{1}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\big((1-t)^{a+n}(1+t)^{b+n}\big)P_m^{(a,b)}\,\mathrm{d}t \\ &= \frac{1}{2^n n!}\int_{-1}^{1}(1-t)^{a+n}(1+t)^{b+n}\frac{\mathrm{d}^n}{\mathrm{d}t^n}P_m^{(a,b)}\,\mathrm{d}t.\end{aligned}$$

Si $n > m$, l'intégrale est nulle, comme on le savait déjà. Supposons $m = n$. On a :

$$\begin{aligned}\|P_n^{a,b}\|_w^2 &= \int_{-1}^{1} P_n^{(a,b)} P_m^{(a,b)} (1-t)^a (1+t)^b \,\mathrm{d}t \\ &= h_n 2^{a+b+2n+1} 2^{a+b+n+1} B(a+n+1, b+n+1)\end{aligned}$$

où h_n est le coefficient du terme de plus haut degré de $P_n^{a,b}$.

On rappelle que $h_n = \frac{\Gamma(2n+a+b+1)}{2^n n!\Gamma(n+a+b+1)}$ (*cf.* l'exercice 7 de la page 572 ou l'assertion (iii) de la proposition 31 de la page 577), d'où :

$$\|P_n^{a,b}\|_w^2 = 2^{a+b+1}\frac{\Gamma(2n+a+b+1)}{n!\Gamma(n+a+b+1)}\frac{\Gamma(n+a+1)\Gamma(n+b+1)}{\Gamma(2n+a+b+2)}.$$

La formule s'en déduit. ∎

Exercice 8.
Calculer μ_0 pour le poids de Jacobi $w \mapsto (1-t)^a(1+t)^b$.

Solution. On a $\mu_0 = \int_{-1}^{1} w(t)\,\mathrm{d}t = \int_{-1}^{1}(1-t)^a(1+t)^b\,\mathrm{d}t$. Posons $1-t = 2u$, on obtient :

$$\mu_0 = 2^{a+b+1}\int_0^1 u^a(1-u)^b du = 2^{1+a+b}B(a+1,b+1) = 2^{a+b+1}\frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}.$$

Nous admettrons ici la formule explicite de récurrence pour les polynômes de Jacobi. On peut la prouver, par des calculs élémentaires, à partir du résultat ci-dessus et des valeurs de ces polynômes en 1 et -1 (*cf.* l'exercice III.4.21).

Proposition 39.
La formule de récurrence pour les polynômes de Jacobi de paramètres (a,b) est :

$$\begin{aligned}2n(n+s)(2n+s-2)P_n^{(a,b)}(t)\\ &= (2n+s-1)\left[(2n+s)(2n+s-2)t + a^2 - b^2\right]P_{n-1}^{(a,b)}(t)\\ &\quad - 2(n+a-1)(n+b-1)(2n+s)P_{n-2}^{(a,b)}(t),\end{aligned}$$

où $s := a+b$.
On a $P_0^{(a,b)} = 1$ et $P_1^{(a,b)}(t) = \frac{1}{2}(a+b+2)t + \frac{1}{2}(a-b)$.

3.3.2 Polynômes de Legendre

Notons $p_n(t) := \frac{\mathrm{d}^n}{\mathrm{d}t^n}(t^2-1)^n$. D'après la formule d'Olinde Rodrigues : $P_n(t) = \frac{1}{2^n n!}p_n(t)$.
On a $P_n(-t) = (-1)^n P_n(t)$. On en déduit que le terme en t^{n-1} dans P_n est nul.
D'après la proposition 38 de la page précédente ($a = b = 0$ ou un calcul direct, que nous conseillons au lecteur...), on obtient le résultat suivant.

Proposition 40. On a, pour tous $n, m \in \mathbb{N}$, $\displaystyle\int_{-1}^{1} P_n(t)P_m(t)\mathrm{d}t = \frac{2}{2n+1}\delta_{nm}$.

D'après le corollaire 3 de la page 537 et l'exercice III.4.2, la relation de récurrence pour les polynômes de Legendre unitaires u_n est :

$$u_{n+1} = tu_n + \frac{\|u_n\|^2}{\|u_{n-1}\|^2}u_{n-1}.$$

Le terme de plus haut degré de $p_n(t) := \frac{\mathrm{d}^n}{\mathrm{d}t^n}(t^2-1)^n$ est $\frac{(2n)!}{n!}t^n$, on a donc $u_n = \frac{n!}{(2n)!}p_n$.
Par ailleurs $\|p_n\|^2 = \frac{2^{2n+1}}{2n+1}(n!)^2$, d'où : $\|u_n\|^2 = \frac{2^{2n+1}}{2n+1}\frac{(n!)^4}{((2n)!)^2}$.
On en déduit :

$$\frac{\|u_n\|^2}{\|u_{n-1}\|^2} = 4\frac{2n-1}{2n+1}\frac{n^4}{(2n(2n-1))^2} = \frac{n^2}{4n^2-1}.$$

On a donc :

$$u_{n+1} = tu_n + \frac{n^2}{4n^2-1}u_{n-1}.$$

On a $P_n(t) = \frac{1}{2^n n!}p_n(t) = \frac{(2n)!}{2^n (n!)^2}u_n(t)$. On en déduit le résultat suivant.

Proposition 41. La relation de récurrence pour les polynômes de *Legendre* est :

$$\forall n \geqslant 1, \quad (n+1)P_{n+1}(t) = (2n+1)tP_n(t) - nP_{n-1}(t).$$

On a $P_0(t) = 1$ et $P_1(t) = t$.

Signalons, sans démonstration (*cf.* l'exercice III.4.11), les relations :

$$P'_{n-1}(t) = tP'_n(t) - nP_n(t) \quad \text{et} \quad P'_{n+1}(t) = tP'_n(t) + (n+1)P_n(t).$$

En appliquant la formule de Leibniz à $p_n(t) := \frac{\mathrm{d}^n}{\mathrm{d}t^n}(t^2-1)^n$, on obtient :

$$\begin{aligned} p_n(t) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}(t-1)^n(t+1)^n &= \sum_{k=0}^{n}\binom{n}{k}\frac{\mathrm{d}^k}{\mathrm{d}t^k}(t-1)^n\frac{\mathrm{d}^{n-k}}{\mathrm{d}t^{n-k}}(t+1)^n \\ &= \sum_{k=0}^{n}\binom{n}{k}(n)_k(n)_{n-k}(t-1)^{n-k}(t+1)^k \\ &= n!\sum_{k=0}^{n}\binom{n}{k}^2(t-1)^{n-k}(t+1)^k \end{aligned}$$

et :

$$P_n(t) = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}^2(t-1)^{n-k}(t+1)^k. \tag{35}$$

Exercice 9.

Retrouver, à l'aide de la formule ci-dessus, la relation de récurrence de la proposition 41.

Solution. On vérifie facilement que $\binom{2n}{n} = \sum\limits_{k=0}^{n}\binom{n}{k}^2$ (*cf.* l'exercice type corrigé du module *Arithmétique* du livre [L1]).

On en déduit que, pour $n \in \mathbb{N}$, le coefficient de t^n dans $P_n(t)$ est $2^{-n}\binom{2n}{n}$.

Par ailleurs $P_n(1) = 1$, pour tout $n \in \mathbb{N}$. D'après la formule de récurrence (*cf.* théorème 4 de la page 538), on a $(n+1)P_{n+1}(t) = (\lambda_n t + \mu_n)P_n(t) + \nu_n P_{n-1}(t)$, pour $\lambda_n, \mu_n, \nu_n \in \mathbb{R}$ convenables. En identifiant les coefficients de t^{n+1} et ceux de t^n, et en faisant, d'autre part, $t = 1$, on trouve les relations : $2^{-(n+1)}(n+1)\binom{2(n+1)}{n+1} = 2^{-n}\binom{2n}{n}\lambda_n$, $\mu_n = 0$ et $n+1 = \lambda_n + \mu_n + \nu_n$. Donc $\lambda_n = 2n+1$ et $\nu_n = -n$.

Exercice 10.

Soient $n, m \in \mathbb{N}$.

1. Montrer que le polynôme $Q_{n,m} := \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^m P_n(t)$ est solution de l'équation différentielle :
$$(1-t^2)y'' - 2(m+1)ty' + \big(n(n+1) - m(m+1)\big)y = 0. \tag{36}$$
2. Montrer que l'équation différentielle (36) est une équation différentielle de polynômes de Jacobi.
3. Montrer que $\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^m P_n(t) = cP_{n-m}^{(m,m)}$ où $c \in \mathbb{R}$ et calculer c.
4. Montrer que la fonction : $t \mapsto P_n^m(t) := (-1)^m(1-t^2)^{m/2}Q_{n,m}$ est une solution sur $]-1,1[$ de l'équation différentielle :
$$(1-t^2)y'' - 2ty' + \Big(n(n+1) - \frac{m^2}{1-t^2}\Big)y = 0. \tag{37}$$
5. Pourquoi introduire la fonction P_n^m ?

Solution.

1. Le polynôme de Legendre P_n est solution de l'équation différentielle de Legendre $(1-t^2)y'' - 2ty' + n(n+1)y = 0$. En dérivant m fois cette équation, on obtient :
$(1-t^2)y^{(m+2)} - 2mty^{(m+1)} - m(m-1)y^{(m)} - 2(ty^{(m+1)} + my^{(m)}) + n(n+1)y^{(m)} = 0$,
soit $(1-t^2)y^{(m+2)} - 2(m+1)ty^{(m+1)} + \big(n(n+1) - m(m+1)\big)y^{(m)} = 0$.
Le polynôme $Q_{n,m}$ est donc solution de l'équation différentielle :
$$(1-t^2)y'' - 2(m+1)ty' + \big(n(n+1) - m(m+1)\big)y = 0.$$
2. L'équation différentielle des polynômes de Jacobi $\{P_k^{m,m}\}_{n\in\mathbb{N}}$ est
$$(1-t^2)y'' - 2(m+1)ty' + k(k+2m+1)y = 0.$$
Pour $k := n-m$, on obtient $(1-t^2)y'' - 2(m+1)ty' + (n-m)(n+m+1)y = 0$, c'est-à-dire (36).
3. En utilisant la proposition 27 de la page 573), on en déduit qu'il existe $c \in \mathbb{R}$, tel que $Q_{n,m} = cP_{n-m}^{(m,m)}$. On calcule c en comparant les termes de plus haut degré. On a $\left(\dfrac{\mathrm{d}}{\mathrm{d}t}\right)^m \left(\dfrac{1}{2^n}\dbinom{2n}{n}t^n\right) = c\,\dfrac{\Gamma(2n+1)}{2^{n-m}(n-m)!\Gamma(n+m+1)}t^{n-m}$, d'où :
$$c = \frac{(n+m)!}{2^m n!} = \frac{(n+m)_m}{2^m}.$$
4. Si l'on pose $y = (1-t^2)^{-m/2}u$, dans (36), on obtient :
$$(1-t^2)u'' - 2tu' + \Big(n(n+1) - \frac{m^2}{1-t^2}\Big)u = 0.$$
Le résultat s'en déduit.
5. La fonction $w : t \in]-1,1[\mapsto (1-t^2)^m$ est le poids pour les polynômes de Jacobi de paramètres $a := m$ et $b := m$, donc $(1-t^2)^{m/2} = \sqrt{w(t)}$.
Ainsi la fonction P_n^m est construite en utilisant le mécanisme décrit page 563.

Solution complète de l'équation différentielle de Legendre

L'équation différentielle de Legendre $(1-t^2)y''-2ty'+n(n+1)y=0$ admet pour solution sur $\mathbb{R}$ le $n^{ème}$ polynôme de Legendre P_n. Cette équation est régulière sur les intervalles ouverts $]-\infty,-1[$, $]-1,1[$, $]1,+\infty[$ et d'après la théorie générale (*cf.* le module *Équations différentielles* du livre [L2]) l'espace des solutions sur chacun de ces intervalles est de dimension 2. Il est donc naturel de chercher une deuxième solution, indépendante de P_n.

Proposition 42. (i) Pour tout $n\in\mathbb{N}$, la fonction $t\mapsto Q_n(t):=\frac{1}{2}\displaystyle\int_{-1}^{1}\frac{P_n(\tau)}{t-\tau}d\tau$ est de classe $\mathcal{C}^\infty$ sur $]-\infty,-1[$ et $]-1,+\infty]$ et, sur chacun de ces intervalles, elle est une solution de l'équation différentielle de Legendre d'indice n :

$$(1-t^2)y''-2ty'+n(n+1)y=0.$$

(ii) On a $Q_n=\frac{1}{2}\ln\left|\frac{t+1}{t-1}\right|P_n-W_n$, où $W_n\in\mathbb{R}_{n-1}[T]$.

(iii) Pour tout $n\in\mathbb{N}$, les fonctions P_n et Q_n sont linéairement indépendantes.

La preuve de ce résultat fait l'objet de l'exercice III.4.14.

3.3.3 Fonctions sphériques

Les polynômes de Legendre ont été, pour la première fois, introduits pour l'étude de l'équation de Laplace $\Delta V=0$. Outre le point de vue historique, la relation entre les polynômes de Legendre et les fonctions sphériques, que nous étudions ci-dessous, reste très importante dans de nombreuses applications en mathématiques et en physique. Elle est similaire à celle existant entre les polynômes de Tchebychev et les polynômes en $\cos\theta$ expliquée plus haut : on passe du cercle S^1 à la sphère S^2 (le passage à la sphère unité S_{k-1} de $\mathbb{R}^k$ conduirait selon la même méthode aux polynômes ultra-sphériques de Gegenbauer, qui sont des polynômes de Jacobi).

Nous nous plaçons dans $\mathbb{R}^3$, orienté et muni de sa structure euclidienne canonique et posons $\vec{\imath}=(1,0,0)$, $\vec{\jmath}=(0,1,0)$, $\vec{k}=(0,0,1)$. Nous utiliserons, pour un point $M=(x,y,z)\in\mathbb{R}^3$, $M\neq O=(0,0,0)$, des *coordonnées sphériques*. Nous adoptons la notation des physiciens : $OM=r$, $(\vec{\imath},\overrightarrow{OM})=\varphi$, $(\vec{k},\overrightarrow{OM})=\theta$, avec les orientations évidentes :

$$x=r\sin\theta\cos\varphi,\quad y=r\sin\theta\sin\varphi,\quad z=r\cos\theta. \tag{38}$$

On considère *l'équation au dérivées partielles de Laplace* $\Delta V=0$, où V est une fonction inconnue sur $\mathbb{R}^3$ et $\Delta=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}$ l'opérateur laplacien :

$$\Delta V=\frac{\partial^2 V}{\partial x^2}+\frac{\partial^2 V}{\partial y^2}+\frac{\partial^2 V}{\partial z^2}=0.$$

Une solution de cette équation est dite *fonction harmonique*.

On montre que le laplacien est *invariant par une isométrie* : si u est une isométrie, $\Delta F=\Delta(F\circ u)$ (*cf.* l'exercice III.4.15). Son expression est donc la même dans tous les repères orthonormés. Elle est différente dans les autres repères.

Si l'on s'intéresse à des problèmes *à symétrie sphérique* (i.e. invariants sous l'action du groupe $SO_3(\mathbb{R})$), il est naturel d'exprimer le laplacien, puis l'équation de Laplace, en coordonnées sphériques. Nous admettrons ici le résultat, dont la vérification est facile mais un peu longue, *cf.* l'exercice III.4.15. On obtient :

$$\Delta V = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial V}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 V}{\partial\varphi^2} = 0. \tag{39}$$

En supposant, bien sûr (ici et dans la suite), $\sin\theta \neq 0$. Nous gardons V pour noter la fonction initiale $V(r,\theta,\varphi)$, par abus de langage.
On définit le *laplacien sphérique* Δ_S par :

$$\Delta_S := \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2} \tag{40}$$

et l'on pose :

$$\Delta_r V := \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial V}{\partial r}\right) = \frac{\partial^2 V}{\partial r^2} + \frac{2}{r}\frac{\partial V}{\partial r}\cdot \tag{41}$$

On a $\Delta = \Delta_r + \frac{1}{r^2}\Delta_S$. Si V ne dépend pas de r, on a $r^2\Delta V = \Delta_S V$.

Point Méthode

La méthode traditionnelle pour chercher *des* solutions de (39) est *la séparation des variables* : on cherche des solutions sous la forme $V(r,\theta,\varphi) = R(r)\Theta(\theta)\Phi(\varphi)$. Elle est justifiée au cas par cas, quand on constate que l'on a « suffisamment de solutions » pour résoudre le problème initial. Le but est de *simplifier* la résolution, en remplaçant un *problème d'équations aux dérivées partielles* (EDP) par un *problème d'équations différentielles ordinaires* (EDO). On a déjà vu plus haut une méthode analogue pour l'étude des ondes stationnaires (*cf.* page 564).

On note $Y(\theta,\varphi) := \Theta(\theta)\Phi(\varphi)$.
En reportant $V(r,\theta,\varphi) = Y(\theta,\varphi)R(r)$ dans (39) et en divisant par V (que l'on suppose non nul), on trouve $r^2\frac{\Delta_r R(r)}{R(r)} + \frac{\Delta_S Y(\theta,\varphi)}{Y(\theta,\varphi)} = 0$, soit $r^2\frac{\Delta_r R(r)}{R(r)} = -\frac{\Delta_S Y(\theta,\varphi)}{Y(\theta,\varphi)}$.
Le premier terme de cette égalité ne dépend que de r et le second ne dépend que de (θ,φ), on a donc $r^2\frac{\Delta_r R(r)}{R(r)} = -\frac{\Delta_S Y(\theta,\varphi)}{Y(\theta,\varphi)} = C_0$, où C_0 est une constante.
Ainsi les fonctions R et Y sont solutions du système d'EDP, où parties radiale et angulaire apparaissent de façon séparée :

$$r^2R'' + 2rR' = C_0R \quad \text{et} \quad \Delta_S Y = -C_0Y. \tag{42}$$

En utilisant $Y = \Theta\Phi$, en reportant dans $\Delta_S Y + C_0Y = 0$, en divisant par Y et multipliant par $\sin^2\theta$, on trouve :

$$\frac{\sin\theta\frac{\mathrm{d}}{\mathrm{d}\theta}(\sin\theta\frac{\mathrm{d}\Theta}{\mathrm{d}\theta})}{\Theta(\theta)} + C_0\sin^2\theta + \frac{\Phi''(\varphi)}{\Phi(\varphi)} = 0.$$

Dans le membre de gauche de cette égalité, la somme des deux premiers termes ne dépend que de θ et le troisième ne dépend que de φ, donc :

$$\frac{\sin\theta\frac{\mathrm{d}}{\mathrm{d}\theta}(\sin\theta\frac{\mathrm{d}\Theta}{\mathrm{d}\theta})}{\Theta(\theta)} + C_0\sin^2\theta = -\frac{\Phi''(\varphi)}{\Phi(\varphi)} = C_1$$

où C_1 est une constante. On obtient ainsi le système d'EDO :

$$\sin\theta \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sin\theta \frac{\mathrm{d}\Theta}{\mathrm{d}\theta}\right) + (C_0 \sin^2\theta - C_1)\Theta(\theta) = 0 \qquad \text{et} \qquad \Phi'' + C_1\Phi = 0$$

Pour des raisons techniques, on *pose* : $C_0 = n(n+1)$ et $C_1 = -m^2$ (on accepte, pour le moment, m et n complexes).

Nous sommes donc ramenés à un système de trois équations différentielles linéaires du second ordre *découplées* (les variables sont respectivement r, θ, φ et les fonctions inconnues respectives ne dépendent que de la variable correspondante). En revanche, on prendra garde au fait que les constantes n et m, qui apparaissant indépendamment dans la première et la troisième équation, figurent toutes les deux dans la seconde.

$$r^2R'' + 2rR' - n(n+1)R = 0 \tag{43}$$

$$\Theta'' + (\operatorname{cotan}\theta)\,\Theta' + \Big(n(n+1) - \frac{m^2}{\sin^2\theta}\Big)\Theta = 0 \tag{44}$$

$$\Phi'' + m^2\Phi = 0 \tag{45}$$

Si $n, m \in \mathbb{N}$, les fonctions $Y_{n,m}(\theta,\varphi) := \Theta(\theta)\Phi(\varphi)$, où Θ et Φ sont respectivement des solutions de (44) et (45) sont dites *harmoniques sphériques*. Elles sont fonctions propres du laplacien sphérique Δ_S, avec la valeur propre $-n(n+1)$ (*cf.* (42)). Ces fonctions interviennent, par exemple, dans la solution du problème des sphères vibrantes, et nous les retrouverons dans l'étude de l'atome d'hydrogène.

L'équation (45) est facile à résoudre pour $m \neq 0$: $\Phi(\varphi) = c_1 \sin m\varphi + c_2 \cos m\varphi$ ($c_1, c_2 \in \mathbb{R}$), et la condition de 2π-périodicité sur les solutions impose $\pm m \in \mathbb{N}$ (nous prendrons $m \in \mathbb{N}$). Pour $m := 0$, les solutions indiquées sont aussi les solutions périodiques de l'équation (45). Le paramètre m représente, dans le problème physique des vibrations des sphères, les *modes* de vibration d'un cercle. L'équation (44), correspond, dans le même problème physique, aux vibrations dites zonales de la sphère (i.e. invariantes par rotation autour de l'axe vertical) ; pour l'étudier, on pose $t := \cos\theta$ (t est la coordonnée de la projection d'un point M, de la sphère unité, sur l'axe vertical ; avec les coordonnées initiales on a $t = z$). On obtient :

$$(1-t^2)y'' - 2ty' + \Big(n(n+1) - \frac{m^2}{1-t^2}\Big)y = 0. \tag{46}$$

On reconnaît, pour $m := 0$, l'équation différentielle des polynômes de Legendre, et, pour $m \in \mathbb{N}$ quelconque l'équation différentielle (37) de l'exercice 10, page 584. Ainsi, pour $m := 0$ et pour tout $n \in \mathbb{N}$, le polynôme de Legendre $P_n(t)$ fournit une solution $P_n(\cos\theta)$ de l'équation (44). On l'appelle $n^{\text{ème}}$ *fonction sphérique zonale*.

Pour $m \in \mathbb{N}$ quelconque, les fonctions : $P_n^m(t) = (-1)^m(1-t^2)^{m/2}\frac{\mathrm{d}^m}{\mathrm{d}t^m}P_n(t)$ de l'exercice 10 fournissent des solutions : $P_n^m(\cos\theta)$. On a $1 - t^2 = \sin^2\theta$ et $\frac{\mathrm{d}}{\mathrm{d}t} = -\frac{1}{\sin\theta}\frac{\mathrm{d}}{\mathrm{d}\theta}$, donc :
$P_n^m(\cos\theta) = \frac{\mathrm{d}^m}{\mathrm{d}\theta^m}P_n(\cos\theta)$.

En résolvant l'équation (43), ce qui est facile, on obtient ainsi *une* famille de solutions de notre problème initial, l'équation aux dérivées partielles $\Delta V = 0$:

$$V(r,\theta,\varphi) = \Big(c_3 r^n + \frac{c_4}{r^{n+1}}\Big)\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^m P_n(\cos\theta)(c_1 \sin m\varphi + c_2 \cos m\varphi).$$

On peut en fait montrer que l'équation (44) admet une solution bornée Θ si, et seulement si, $n \in \mathbb{N}$, mais c'est au-delà des techniques dont nous disposons dans ce livre. Ainsi la condition $n \in \mathbb{N}$ est en fait *nécessaire* (on cherche Θ continue et la sphère est compacte). On retrouvera cela dans l'étude

de l'atome d'hydrogène, $\pm m$ et n seront deux des trois nombres quantiques définissant les états (n sera changé en ℓ, pour des raisons de tradition).
Pour des fonctions continues f et g sur la sphère unité S^2 de $\mathbb{R}^3$, on introduit le produit scalaire $\frac{1}{4\pi}(\,f \mid g\,)_\sigma := \displaystyle\int_{S^2} fg d\sigma = \int_0^{2\pi}\int_0^{\pi} f(\theta,\varphi)g(\theta,\varphi)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi$ (le lecteur vérifiera que c'est bien un produit scalaire sur l'espace des fonctions continues). Ce produit scalaire généralise celui que nous avons introduit sur l'espace des fonctions continues sur le cercle unité S^1 de $\mathbb{R}^2$ pour l'étude des séries de Fourier : $(\,f \mid g\,) = \frac{1}{2\pi}\displaystyle\int_{S^1} fg\,\mathrm{d}s = \frac{1}{2\pi}\int_0^{2\pi} f(\varphi)g(\varphi)\,\mathrm{d}\varphi$.

Exercice 11.

Soient $m, n \in \mathbb{N}$, avec $n \neq m$. Montrer que les deux fonctions sphériques zonales $P_n(\cos\theta)$ et $P_m(\cos\theta)$ sont orthogonales pour $(\ \mid\ )_\sigma$.

Solution. On utilise la *bijection* entre le cylindre $C =]0, 2\pi[\times S^1$ et la sphère S^2 privée des pôles nord et sud définie par $u : (\theta,\varphi) \mapsto (t := \cos\theta, \varphi)$ (φ étant défini module 2π). Le déterminant jacobien de u est $\frac{D(t,\varphi)}{D(\theta,\varphi)} = -\sin\theta$. L'application u *conserve les aires* :

$$d\sigma = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\varphi = -\frac{D(t,\varphi)}{D(\theta,\varphi)}\,\mathrm{d}\theta \wedge \mathrm{d}\varphi = -\,\mathrm{d}t \wedge \mathrm{d}\varphi.$$

Ce résultat était déjà connu d'Archimède (-287, -212), qui aurait demandé qu'il soit gravé sur sa tombe. On a, pour $n \neq m$:

$$\begin{aligned}(\,P_n \mid P_m\,)_\sigma = \int_{S^2} P_n P_m d\sigma &= \int_0^{2\pi}\int_{-\pi}^{\pi} P_n(\cos\theta)P_m(\cos\theta)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi \\ &= \int_0^{2\pi}\int_{-1}^{1} P_n(t)P_m(t)\,\mathrm{d}t\,\mathrm{d}\varphi = \int_0^{2\pi} 0\ \mathrm{d}\varphi = 0.\end{aligned}$$

Soient $t_{k,n} \in\ ]-1,1[$ ($k \in [\![1,n]\!]$) les zéros du polynôme de Legendre P_n. Du point de vue physique, une fonction zonale $P_n(\cos\theta)$ correspond à une vibration de la sphère unité, isotrope par rotation autour de l'axe vertical et telle que les cercles horizontaux de côte $t_{k,n}$ (i.e. les ensembles de points de coordonnée $z = t_{k,n}$) restent *fixes*.
Le lecteur comparera avec la description des vibrations zonales du cercle unité par les zéros des polynômes de Tchebychev que nous avons donnée plus haut.
Si V est une fonction homogène de degré n sur $\mathbb{R}^3$, en coordonnées sphériques $\big(V(\lambda r,\theta,\varphi) = \lambda^n V(r,\theta,\varphi)\big)$, on a :

$$\Delta V = \frac{1}{r^2}\big[\Delta_S V + n(n+1)V\big]$$

(*cf.* exercice III.4.17). Si V est, de plus, harmonique, sa restriction à la sphère unité S^2 de $\mathbb{R}^3$ est donc un vecteur propre de Δ_S, avec la valeur propre $-n(n+1)$. La relation entre polynômes harmoniques et fonctions sphériques est l'objet de l'exercice III.4.17.

Exemple. On a $P_2(\cos\theta) = \frac{1}{2}(3\cos^2\theta - 1) = \frac{1}{2}(2z^2 - x^2 - y^2)$, en utilisant les coordonnées (x,y,z) et en se restreignant à la sphère S^2. Le polynôme $Q_2(x,y,z) := \frac{1}{2}(2z^2 - x^2 - y^2)$ est *homogène* de degré 2 et $\Delta Q = 0$, il est donc *harmonique*. Cela s'étend aux autres polynômes de Legendre : ils correspondent à des restrictions à la sphère unité de polynômes homogènes harmoniques à trois variables. Par exemple $P_3(\cos\theta)$ est la restriction à la sphère unité de $Q_3(x,y,z) = \frac{1}{2}(2z^3 - 3x^2z - 3y^2z)$, qui est homogène de degré 3 et harmonique.

3.3.4 Polynômes de Tchebychev

Les polynômes de Tchebychev $(T_n)_{n\in\mathbb{N}}$ forment, nous l'avons vu, une suite de polynômes orthogonaux associés au poids $w(t) = (1-t^2)^{-1/2}$ (c'est le cas de Jacobi, avec $a = -1/2$, $b = -1/2$). Ils ont été introduits par le mathématicien russe Pafnouti Tchebychev (1821-1894) pour l'étude de l'interpolation de Lagrange : le choix des zéros de T_n comme nœuds de l'interpolation fournit de très bons résultats numériques (*cf.* le module *Méthodes numériques* du livre [L2])

Les polynômes de Tchebychev (de première et deuxième espèce) ont été introduits dans le livre [L1], à partir de quelques formules de trigonométrie. Voici quelques rappels et compléments.

Définition 8. Les *polynômes de Tchebychev de première espèce* sont les polynômes définis par la relation de récurrence : $T_0 := 1$, $T_1 := t$ et, pour tout $n \geqslant 1$, $T_{n+1} := 2tT_n - T_{n-1}$.

Les *polynômes de Tchebychev de deuxième espèce* sont les polynômes définis par la relation de récurrence : $U_0 := 1$, $U_1 := 2t$ et, pour tout $n \geqslant 1$, $U_{n+1} := 2tU_n - U_{n-1}$.

On prendra garde au fait que cette terminologie traditionnelle n'est pas en accord avec celle introduite dans la proposition 5 de la page 540, pour les polynômes orthogonaux « généraux ». On a :

$$\begin{array}{lll} T_2(t) = 2t^2 - 1 & T_3(t) = 4t^3 - 3t & T_4(t) = 8t^4 - 8t^2 + 1 \\ U_2(t) = 4t^2 - 1 & U_3(t) = 8t^3 - 4t & U_4(t) = 16t^4 - 12t^2 + 1. \end{array}$$

Proposition 43. Soit $n \in \mathbb{N}$.

- Le polynôme T_n est l'unique polynôme tel que $T_n(\cos\theta) = \cos n\theta$, pour tout $\theta \in \mathbb{R}$.
- Le polynôme U_n est l'unique polynôme tel que $U_n(\cos\theta) = \dfrac{\sin(n+1)\theta}{\sin\theta}$ pour tout $\theta \in \mathbb{R} \setminus \pi\mathbb{Z}$.
- Si $n \geqslant 1$, $T_n' = nU_{n-1}$.

Pour $n \geqslant 1$, le coefficient du terme de plus haut degré de T_n est 2^{n-1}, celui du terme de plus haut degré de U_n est 2^n.

Nous avons vu (*cf.* page 564) que les polynômes $(T_n)_{n\in\mathbb{N}}$ sont orthogonaux sur $]-1,1[$ pour la fonction poids $w(t) := (1-t^2)^{-1/2}$, et sont respectivement solutions de l'équation différentielle $(1-t^2)y'' - ty' + n^2y = 0$. On vérifie que cette équation est un cas particulier de celle de Jacobi : $a := b := -1/2$.

Considérons maintenant le poids $w(t) := (1-t^2)^{1/2}$ sur $]-1,1[$.

On a, pour tous $n, m \in \mathbb{N}$, $\displaystyle\int_{-1}^{1} U_n(t)U_m(t)w(t)\,\mathrm{d}t = \int_0^{\pi} \sin\big((n+1)\theta\big)\sin\big((m+1)\theta\big)\,\mathrm{d}\theta$, donc $\displaystyle\int_{-1}^{1} U_n(t)U_m(t)w(t)\,\mathrm{d}t = 0$ si $n \neq m$. Ainsi les polynômes $(U_n)_{n\in\mathbb{N}}$ sont orthogonaux pour le poids $w(t) = (1-t^2)^{1/2}$. Ils vérifient donc respectivement les équations différentielles $(1-t^2)y'' - 3ty' + n(n+2)y = 0$ (obtenues à partir de celles de Jacobi, dans le cas où $a := b := 1/2$).

Exercice 12.
Retrouver les équations différentielles des polynômes de Tchebychev de deuxième espèce par un argument analogue à celui employé pour les polynômes de première espèce (*cf.* l'exemple de la page 563).
Solution. On a $\sqrt{1-t^2}U_n(t) = \sin(n+1)\theta$ et la fonction $u(\theta) = \sin(n+1)\theta$ vérifie l'équation différentielle $u'' + (n+1)^2 u = 0$.
Posons $u = v\sin\theta$. On a $(\sin\theta)v'' + 2(\cos\theta)v' + n(n+2)(\sin\theta)v = 0$. On a :

$$\begin{aligned}\sin\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^2 + 2\cos\theta\frac{\mathrm{d}}{\mathrm{d}\theta} &= \sin\theta\left(-\sin\theta\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 - 2\sin\theta\cos\theta\frac{\mathrm{d}}{\mathrm{d}t}\\ &= \sqrt{1-t^2}\left(\sqrt{1-t^2}\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 - 2t\sqrt{1-t^2}\frac{\mathrm{d}}{\mathrm{d}t}\\ &= \sqrt{1-t^2}\left((1-t^2)\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 - 3t\frac{\mathrm{d}}{\mathrm{d}t}\right).\end{aligned}$$

D'où le résultat.

La proposition suivante ne résulte pas de l'énoncé analogue pour les polynômes de Jacobi, car la normalisation n'est pas la même…

Proposition 44. On a, pour tous $m, n \in \mathbb{N}$:

(i) $$\int_{-1}^{1} T_n(t)T_m(t)(1-t^2)^{-1/2}\,\mathrm{d}t = \int_0^{\pi}\cos n\theta\cos m\theta\,\mathrm{d}\theta = \begin{cases}0 & \text{si} \quad m \neq n\\ \frac{\pi}{2} & \text{si} \quad m = n \in \mathbb{N}^*\\ \pi & \text{si} \quad m = n = 0\end{cases}.$$

(ii) $$\int_{-1}^{1} U_n(t)U_m(t)(1-t^2)^{1/2}\,\mathrm{d}t = \int_0^{\pi}\sin\big((n+1)\theta\big)\sin\big((m+1)\theta\big)\,\mathrm{d}\theta = \frac{\pi}{2}\delta_{nm}.$$

3.4 Polynômes d'Hermite

Nous avons vu que les polynômes d'Hermite $H_n(t)$ forment une suite de polynômes orthogonaux associés au poids $w(t) := \mathrm{e}^{-t^2}$ sur la droite $\mathbb{R}$. On a la formule d'Olinde Rodrigues : $H_n(t) = (-1)^n \mathrm{e}^{t^2}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathrm{e}^{-t^2}$ (*cf.* le corollaire 36 de la page 580).
Les premiers polynômes d'Hermite sont donc :

$$H_0(t) = 1, H_1(t) = 2t, H_2(t) = 4t^2 - 2, H_3(t) = 8t^3 - 12t, H_4(t) = 16t^4 - 48t^2 + 12.$$

Rappelons que le $n^{ème}$ polynôme d'Hermite vérifie l'équation différentielle :

$$y'' - 2ty' + 2ny = 0. \tag{47}$$

On en déduit l'expression explicite du $n^{ème}$ polynôme d'Hermite.

Proposition 45. On a, pour tout $n \in \mathbb{N}$:

$$H_n(t) = n! \sum_{k=0}^{[n/2]} \frac{(-1)^k}{k!} \frac{(2t)^{n-2k}}{(n-2k)!}.$$

Démonstration. Supposons $n \geqslant 2$. Posons $H_n = \sum_{k=0}^{n} c_{n-k} t^{n-k}$, avec $c_n := 2^n$, conformément à la normalisation choisie pour les polynômes d'Hermite. D'après (47), on a : $H_n'' = 2(tH_n' - nH_n)$. Par ailleurs :

$$tH_n' - nH_n = -\sum_{k=1}^{n} k c_{n-k} t^{n-k} \quad \text{et} \quad H_n'' = \sum_{k=2}^{n} (n+2-k)(n+1-k) c_{n+2-k} t^{n-k},$$

donc $H_n'' = 2(tH_n' - nH_n)$ si, et seulement si, :

$$c_{n-1} = 0 \quad \text{et} \quad \forall k \in [\![2, n]\!],\ c_{n-k} = -\frac{1}{2k}(n+2-k)(n+1-k) c_{n+2-k}.$$

On en déduit :

$$\forall k \in [\![0, [n/2]]\!],\ c_{n-2k-1} = 0 \quad \text{et} \quad c_{n-2k} = \frac{(-1)^k n!}{k!} \frac{2^{n-2k}}{(n-2k)!}.$$

Ainsi $H_n(t) = n! \sum_{k=0}^{[n/2]} \frac{(-1)^k}{k!} \frac{(2t)^{n-2k}}{(n-2k)!}$.

Remarquons que l'on retrouve les propriétés de parité des H_n. ∎

Proposition 46. On a $(H_n \mid H_m)_w = \int_{-\infty}^{+\infty} H_n(t) H_m(t) e^{-t^2}\, \mathrm{d}t = 2^n n! \sqrt{\pi} \delta_{nm}$.

Démonstration. Il suffit de prouver le résultat pour $m \geqslant n$. On a :

$$(H_n \mid H_m)_w = \int_{-\infty}^{+\infty} \mathrm{e}^{-t^2} H_n(t) H_m(t)\, \mathrm{d}t = (-1)^n \int_{-\infty}^{+\infty} \frac{\mathrm{d}^n}{\mathrm{d}t^n}(\mathrm{e}^{-t^2}) H_m(t)\, \mathrm{d}t.$$

On peut appliquer la formule (31) d'intégration par parties généralisée et l'on obtient :

$$(H_n \mid H_m)_w = \int_{-\infty}^{+\infty} \mathrm{e}^{-t^2} H_m^{(n)}(t)\, \mathrm{d}t.$$

Si $m > n$, $H_m^{(n)} = 0$ et l'on retrouve $(H_n \mid H_m)_w = 0$, et, si $n = m$, on obtient :

$$\|H_n\|^2 = 2^n n! \int_{-\infty}^{+\infty} \mathrm{e}^{-t^2}\, \mathrm{d}t = 2^n n! \sqrt{\pi}.$$ ∎

On associe aux polynômes d'Hermite les *fonctions d'Hermite* $\psi_n(t) := \mathrm{e}^{-t^2/2} H_n(t)$. En notant, quand les intégrales existent :

$$(f \mid g) := \int_{-\infty}^{+\infty} f(t) g(t)\, \mathrm{d}t \quad \text{et} \quad (f \mid g)_w := \int_{-\infty}^{+\infty} f(t) g(t) \mathrm{e}^{-t^2}\, \mathrm{d}t,$$

on a : $(\psi_n \mid \psi_m) = (H_n \mid H_m)_w$ (les intégrales sont évidemment convergentes). Il en résulte que, si $m \neq n$, les fonctions d'Hermite ψ_n et ψ_m sont orthogonales, au sens ordinaire du produit scalaire des fonctions continues, de carré intégrable, sur $\mathbb{R}$ (sans poids).

On a $\psi_n(t) = (-1)^n e^{t^2/2} \frac{\mathrm{d}^n}{\mathrm{d}t^n} e^{-t^2}$.
Les fonctions d'Hermite apparaissent dans plusieurs domaines des mathématiques et de la physique. En mécanique quantique, elles fournissent les solutions de l'équation de Schrödinger stationnaire pour *l'oscillateur harmonique*, comme nous le verrons plus loin.
En remplaçant y par $e^{t^2/2}\psi$ dans l'équation différentielle des polynômes d'Hermite : $y'' - 2ty' + 2ny = 0$, on obtient $\psi'' + (2n+1-t^2)\psi = 0$. C'est l'équation différentielle des fonctions d'Hermite. Notons qu'elle ne contient pas de terme en ψ' (et que le facteur $\mathrm{e}^{-t^2/2}$ est exactement celui qu'il fallait pour cela).
Certains auteurs (en particulier les probabilistes) utilisent le poids $\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-t^2/2}$ (la densité de la mesure gaussienne sur la droite réelle, *cf.* le module IV.3), au lieu du poids e^{-t^2} pour définir d'autres polynômes d'Hermite He_n. Le dictionnaire est $He_n(t) = 2^{-n/2} H_n(t/\sqrt{2})$.

Relations de récurrence

Proposition 47. On a, pour tout $n \in \mathbb{N}^*$:

(i) $H_{n+1}(t) - 2tH_n(t) + 2nH_{n-1}(t) = 0$,

(i) $H_n'(t) = 2nH_{n-1}(t)$.

Démonstration. (*i*) On utilise la formule d'Olinde Rodrigues. On a :

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{n+1}(e^{-t^2}) = \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{n}(-2te^{-t^2}) = -2t\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{n}(e^{-t^2}) - 2n\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{n-1}(e^{-t^2}).$$

En multipliant l'égalité par $(-1)^{n+1}e^{t^2}$, on obtient le résultat.
(*ii*) On a $H_n'(t) = (-1)^n e^{t^2}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{n+1}(e^{-t^2}) + (-1)^n 2te^{t^2}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{n}(e^{-t^2})$, c'est-à-dire : $H_n'(t) = -H_{n+1}(t) + 2tH_n(t)$. En utilisant (i), on obtient

$$H_n'(t) = -2tH_n(t) + 2nH_{n-1}(t) + 2tH_n(t) = 2nH_{n-1}(t).$$

Les équations différentielles des fonctions d'Hermite appartiennent à la famille :

$$\psi'' + (2\lambda + 1 - t^2)\psi = 0 \qquad (\lambda \in \mathbb{R}).$$

On rencontre cette famille dans de nombreuses applications, comme l'oscillateur harmonique (*cf.* ci-dessous 3.4.1). On a l'important résultat suivant sur les solutions de carré sommable (qui sera prouvé dans la correction de l'exercice type III.4.0 de la page 647).

Théorème 48. (i) Si ψ est une solution *de carré intégrable*, sur $\mathbb{R}$, de l'équation différentielle $\psi'' + (2\lambda+1-t^2)\psi = 0$ ($\lambda \in \mathbb{R}$), alors, on a nécessairement $\lambda := n \in \mathbb{N}$ et $\psi = c\psi_n$, où c est une constante et ψ_n la $n^{\text{ème}}$ fonction d'Hermite.
(ii) Soit $\gamma \in \mathbb{R}$, $\frac{1}{\sqrt{2}} < \gamma < 1$. Soit $y : t \mapsto y(t)$ une solution de l'équation différentielle $y'' - 2ty' + 2\lambda y = 0$ ($\lambda \in \mathbb{R}$).
Alors, ou bien il existe $C > 0$ et $t_0 \geqslant 0$ tels que $|y(t)| > Ce^{\gamma t^2}$, pour $t \geqslant t_0$, ou bien il existe $C' > 0$ et $t_0' \leqslant 0$ tels que $|y(t)| > C'e^{\gamma t^2}$, pour $t \leqslant t_0$, ou bien $\lambda = n \in \mathbb{N}$ et $y = cH_n$, où c est une constante et H_n le $n^{\text{ème}}$ polynôme d'Hermite.

Le lecteur vérifiera facilement que l'assertion (i) résulte de l'assertion (ii) (une fonction $t \mapsto \psi(t)$ qui tend vers l'infini quand t tend vers $+\infty$ (respectivement $-\infty$) ne peut pas être de carré sommable).

Opérateurs de création-annihilation

Les opérateurs *de création-annihilation* apparaissent en physique et chimie moderne (théorie quantique des champs), en ingénierie électrique… Ils créent une particule à partir du vide, respectivement détruisent une particule (i.e. la restituent au vide). Des opérateurs vérifiant un formalisme similaire apparaissent en mécanique quantique, théorique ou appliquée (états liés, lasers, nanotechnologies…) ; ce sont les opérateurs de montée-descente, ils ajoutent un quantum d'énergie, respectivement enlèvent un tel quantum. Les auteurs anglo-saxons parlent de *Creation-Annihilation, Raising-Lowering, Promotion-Demotion, Ladder*. Nous simplifierons ici la terminologie et parlerons aussi d'opérateurs de création-annihilation pour la montée-descente.
Le formalisme des opérateurs de création-annihilation est élémentaire, mais puissant et profond. Nous allons le mettre en place pour les cas des polynômes d'Hermite et des fonctions d'Hermite. Il permet, dans ce cadre, d'étudier l'oscillateur harmonique en mécanique quantique.
On désigne toujours par w le poids d'Hermite sur $\mathbb{R}$: $w(t) := \mathrm{e}^{-t^2}$.
On définit des opérateurs différentiels $\partial := \frac{1}{\sqrt{2}}\frac{\mathrm{d}}{\mathrm{d}t}$ et $\partial^* := -\frac{1}{\sqrt{2}}\frac{\mathrm{d}}{\mathrm{d}t} + \sqrt{2}\,t$ sur $\mathbb{R}[T]$ (et plus généralement sur l'espace des fonctions de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$) :

$$\partial p(t) = \frac{1}{\sqrt{2}}p'(t) \qquad \text{et} \qquad \partial^* p(t) = -\frac{1}{\sqrt{2}}p'(t) + \sqrt{2}\,tp(t).$$

On vérifie :

$$\partial^* = \mathrm{e}^{t^2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-t^2} \quad \text{et} \quad \partial^*\partial = -\frac{1}{2}D, \quad \text{avec } D = \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 - 2t\frac{\mathrm{d}}{\mathrm{d}t}\cdot \tag{48}$$

Proposition 49. Si l'espace vectoriel réel $\mathbb{R}[T]$ est muni du produit scalaire $(\,.\mid .\,)_w$, défini par le poids d'Hermite $w(t) = \mathrm{e}^{-t^2}$, et si ∂ et ∂^* sont considérés comme opérateurs sur $\mathbb{R}[T]$:
(i) les opérateurs ∂ et ∂^* sont *adjoints* ;
(ii) $\partial\partial^* - \partial^*\partial = [\partial, \partial^*] = \mathrm{Id}$;
(iii) L'opérateur $\mathcal{N}' := \partial^*\partial$ est *autoadjoint* (on dit aussi symétrique).

Remarquons que, d'après (ii), les opérateurs ∂ et ∂^* ne sont *pas* autoadjoints.

Démonstration. (i) On a, pour tous $p, q \in \mathbb{R}[T]$:

$$\begin{aligned}(\,\partial p \mid q\,)_w &= \frac{1}{\sqrt{2}}\int_{-\infty}^{+\infty} p'(t)q(t)\mathrm{e}^{-t^2}\mathrm{d}t \\ &= \frac{1}{\sqrt{2}}\big[p(t)q(t)\mathrm{e}^{-t^2}\big]_{-\infty}^{+\infty} - \frac{1}{\sqrt{2}}\int_{-\infty}^{+\infty} p(t)\big(q'(t) - 2tq(t)\big)\mathrm{e}^{-t^2}\mathrm{d}t,\end{aligned}$$

et donc $(\partial p \mid q)_w = (p \mid \partial^* q)_w$.
(ii) $\partial\partial^* - \partial^*\partial = \frac{\mathrm{d}}{\mathrm{d}t}\left(-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} + t\right) + \left(\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} - t\right)\frac{\mathrm{d}}{\mathrm{d}t} = \left[\frac{\mathrm{d}}{\mathrm{d}t}, t\right] = 1.$ ■

Proposition 50. Les polynômes d'Hermite $(H_n)_{n\in\mathbb{N}}$ vérifient les relations suivantes :

(i) $\partial H_n = \sqrt{2}nH_{n-1}$, pour tout $n \geqslant 1$, et $\partial^* H_n = \frac{1}{\sqrt{2}}H_{n+1}$, pour tout $n \in \mathbb{N}$;

(ii) $\mathcal{N}' H_n = nH_n$;

(iii) $H_n = 2^{n/2}(\partial^*)^n 1$.

Démonstration. Montrons (i).
On a $\partial H_n = \frac{1}{\sqrt{2}}H_n' = \sqrt{2}nH_{n-1}$ et :

$$\sqrt{2}\partial^* H_n = -H_n' + 2tH_n = -H_n' + 2nH_{n-1} + H_{n+1} = H_{n+1}$$

(*cf.* la proposition 47, (ii), puis (i)).
L'assertion (ii) est une conséquence immédiate de (i) :

$$\mathcal{N}' H_n = \partial^*\partial H_n = \sqrt{2}n\partial^* H_{n-1} = nH_n.$$

On peut aussi voir que H_n est un vecteur propre de $\mathcal{N}'$ associé à la valeur propre n en utilisant (48).

Pour terminer, on prouve (iii), par récurrence, en utilisant $H_n = \sqrt{2}\partial^* H_{n-1}$ $(n \geqslant 1)$ et $H_0 = 1$. Cette assertion nous dit que l'on obtient (à une constante multiplicative près) les polynômes d'Hermite, à partir de $H_0 = 1$, par applications successives de l'opérateur ∂^*. ■

Terminologie : les opérateurs ∂^* et ∂ sont respectivement appelés *opérateurs de création* et *opérateurs d'annihilation*, l'opérateur $\mathcal{N}' = \partial^*\partial$ est appelé *opérateur nombre de quanta* ou *opérateur nombre de particules*, la relation $\partial\partial^* - \partial^*\partial = 1$ est la *relation d'Heisenberg*. Ainsi l'opérateur *autoadjoint* $\mathcal{N}'$ admet pour *valeurs propres* les entiers $n \in \mathbb{N}$ et les polynômes d'Hermite sont des *vecteurs propres* correspondants (on retrouve ainsi la propriété d'orthogonalité des polynômes d'Hermite).
Du point de vue des physiciens le polynôme H_n correspond à un état à n particules ou, comme nous le verrons, à un état d'énergie $\hbar\omega\left(\frac{1}{2} + n\right)$. À partir *d'un seul* état (n fixé), on peut obtenir *tous les autres* à l'aide des opérateurs de création et d'annihilation $\big($*cf.* la proposition 50 ci-dessus, (i) et (ii)$\big)$ on peut en particulier partir de l'état d'énergie minimale (le « vide »), représenté ici par $H_0 = 1$, et obtenir tous les autres en itérant l'application de l'opérateur de création.
On peut développer la théorie des opérateurs de création-annihilation de façon purement *formelle*, à partir de deux opérateurs *adjoints* a et a^+ sur un espace préhilbertien *vérifiant* la relation de Heisenberg $[\mathrm{a}, \mathrm{a}^+] = \mathrm{a}\mathrm{a}^+ - \mathrm{a}^+\mathrm{a} = \mathrm{Id}$, en utilisant des calculs *élémentaires* d'algèbre de Lie (*cf.* l'exercice III.4.18).

Exercice 13.

Retrouver la formule d'Olinde Rodrigues pour les polynômes d'Hermite en utilisant l'opérateur de création ∂^*.

Solution. On a l'égalité d'opérateurs $\sqrt{2}\,\partial^* = -\frac{\mathrm{d}}{\mathrm{d}t} + 2t = -\mathrm{e}^{t^2}\frac{\mathrm{d}}{\mathrm{d}t}e^{-t^2}$.

Pour les applications à la physique, il est commode de transposer le formalisme que nous venons de décrire aux *fonctions d'Hermite*. Le produit scalaire est alors $(\,f \mid g\,) = \displaystyle\int_{-\infty}^{+\infty} f(t)g(t)\mathrm{d}t$; il n'y a plus de poids (on suppose que f et g sont de classe $\mathcal{C}^\infty$ et de carré intégrable sur $\mathbb{R}$ ainsi que toutes leurs dérivées). Le passage des polynômes d'Hermite aux fonctions d'Hermite s'effectue facilement à l'aide de la formule suivante que l'on prouve facilement.

Lemme 51. On a l'égalité d'opérateurs $t - \frac{\mathrm{d}}{\mathrm{d}t} = -\,\mathrm{e}^{t^2/2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-t^2/2}$ (sur $\mathbb{R}[T]$, ou plus généralement sur l'espace des fonctions de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$).

En utilisant la formule d'Olinde Rodrigues, on a $H_n = (-1)^n\mathrm{e}^{t^2}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n(\mathrm{e}^{-t^2})$. Ainsi, puisque $\mathrm{e}^{-t^2} = \mathrm{e}^{-t^2/2}\psi_0$:

$$\psi_n = \mathrm{e}^{-t^2/2}H_n = (-1)^n\mathrm{e}^{t^2/2}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n e^{-t^2/2}(\psi_0).$$

En utilisant $\mathrm{e}^{t^2/2}(\frac{\mathrm{d}}{\mathrm{d}t})^n\mathrm{e}^{-t^2/2} = \left(\mathrm{e}^{t^2/2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-t^2/2}\right)^n$, on obtient :

$$\psi_n = (-1)^n\left(\mathrm{e}^{t^2/2}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-t^2/2}\right)^n(\psi_0) = \left(t - \frac{\mathrm{d}}{\mathrm{d}t}\right)^n(\psi_0).$$

L'adjoint de l'opérateur $t - \frac{\mathrm{d}}{\mathrm{d}t}$ pour le produit scalaire ordinaire (sans poids) est $t + \frac{\mathrm{d}}{\mathrm{d}t}$.
On a $\left[t + \frac{\mathrm{d}}{\mathrm{d}t}, t - \frac{\mathrm{d}}{\mathrm{d}t}\right] = 2$. On pose donc $\mathrm{a} = \frac{1}{\sqrt{2}}\left(t + \frac{\mathrm{d}}{\mathrm{d}t}\right)$, $\mathrm{a}^+ = \frac{1}{\sqrt{2}}\left(t - \frac{\mathrm{d}}{\mathrm{d}t}\right)$ et $\mathcal{N} = \mathrm{a}^+\mathrm{a}$.
On prouve facilement le résultat suivant, en adaptant le cas des polynômes d'Hermite. On vérifie que les fonctions d'Hermite et toutes leurs dérivées sont de carré intégrable (elles sont chacune le produit de e^{-t^2} par un polynôme).

Théorème 52. L'espace vectoriel réel $\mathbb{R}[T]$ est muni du produit scalaire $(\,.\mid .\,)$ ordinaire, et l'on définit des opérateurs $\mathrm{a} := \frac{1}{\sqrt{2}}\left(t + \frac{\mathrm{d}}{\mathrm{d}t}\right)$ et $\mathrm{a}^+ := \frac{1}{\sqrt{2}}\left(t - \frac{\mathrm{d}}{\mathrm{d}t}\right)$ (sur l'espace des fonctions de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$ qui sont de carré intégrable pour le poids w, ainsi que toutes leurs dérivées), alors :

1. (i) les opérateurs a et a^+ sont *adjoints* ;
   (ii) $[\mathrm{a}, \mathrm{a}^+] = \mathrm{Id}$;
   (iii) (L'opérateur $\mathcal{N} := \mathrm{a}^+\mathrm{a}$ est *autoadjoint*.
2. Les fonctions d'Hermite $(\psi_n)_{n\in\mathbb{N}}$ vérifient les relations suivantes :
   (i) $\mathrm{a}\psi_n = \sqrt{2}n\psi_{n-1}$, $\mathrm{a}^+\psi_n = \frac{1}{\sqrt{2}}\psi_{n+1}$;
   (ii) $\mathcal{N}\psi_n = n\psi_n$;
   (iii) $\psi_n = 2^{n/2}(\mathrm{a}^+)^n\psi_0$;

3.4.1 Oscillateur harmonique

Historiquement l'oscillateur harmonique apparaît avec la *loi de Hooke*, sur les petites déformations élastiques des solides : « ut tensio sic vis » (Robert Hooke 1678), c'est-à-dire *l'allongement est proportionnel à la force exercée*.

L'oscillateur harmonique *classique* est un corps de masse m attaché à un ressort *linéaire*. On note $k > 0$ la constante du ressort. L'équation de Newton s'écrit $m\overrightarrow{\Gamma} = \overrightarrow{F}$, i.e. $m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -kx$, soit $x'' + \omega^2 x = 0$, avec $\omega = \sqrt{k/m}$. La force dérive du potentiel $V(x) = k\frac{x^2}{2}$ (à une constante additive près) et l'on a l'intégrale première de l'énergie :

$$E := \frac{m(\mathrm{d}x/\,\mathrm{d}t)^2}{2} + V(x) = \frac{1}{2m}p^2 + \frac{1}{2}kx^2,$$

avec $p = m\,\mathrm{d}x/\,\mathrm{d}t$ (l'impulsion ou quantité de mouvement). L'importance de ce système vient du fait qu'il approxime tout système « assez général » proche d'un état d'équilibre stable (i.e. d'un minimum du potentiel), *cf.* par exemple le cas du *pendule simple* dans le module *Équations différentielles* du livre [L2].

L'importance de l'analogue quantique de l'oscillateur harmonique a été mise en évidence par Einstein en 1905, pour la résolution du problème de la chute de chaleur spécifique des corps à basse température. Sa théorie est basée sur *la quantification des états d'énergie* de l'oscillateur, qui, du point de vue mathématique, n'est pas autre chose, nous allons le voir, que *la théorie des polynômes d'Hermite*. Pour des compléments sur les aspects physiques du problème le lecteur pourra se reporter par exemple à [PCP].

Aux grandeurs physiques, dites *observables classiques*, de la physique classique, comme la position x, l'impulsion p ou l'énergie E, on fait correspondre en mécanique quantique des *opérateurs* dits *observables quantiques*. Ces opérateurs sont des opérateurs linéaires sur des espaces de fonctions, les *fonctions d'ondes*, qui sont des *espaces de Hilbert complexes* (complexe est absolument *essentiel*). Les observables quantiques sont des opérateurs *autoadjoints*, les valeurs prises par les observables classiques deviennent les *valeurs propres* des opérateurs correspondants, qui, nous le savons (*cf.* le module *Espaces préhilbertiens* du livre [L2]) sont *réelles*.

On a des *règles de correspondance* qui sont une « recette » permettant de passer des *observables classiques* aux *observables quantiques*. Dans le cas simple qui nous concerne ici, les fonctions d'onde dépendent de (t, x) (t étant le temps et x la position), on fait correspondre $\Big($ h étant la *constante de Planck* et $\hbar := \dfrac{h}{2\pi}\Big)$:

- à la *position* x, l'opérateur X de *multiplication* par x ;
- à *l'impulsion* p, l'opérateur de *dérivation partielle* $P := -i\hbar\dfrac{\partial}{\partial x}$;
- à *l'énergie totale* E, ou hamiltonien H, l'opérateur dit hamiltonien $H := i\hbar\dfrac{\partial}{\partial t}$.

Il est fondamental de noter que les observables classiques, comme la position ou l'impulsion sont des réels et *commutent*. Il n'en est plus du tout de même pour les opérateurs correspondants : par exemple $[P, X] := PX - XP = -i\hbar\frac{\partial}{\partial x}x + x\left(-i\hbar\frac{\partial}{\partial x}\right) = -i\hbar$; c'est la relation canonique de commutation, appelée *de Heisenberg* par les mathématiciens. Elle traduit le *principe d'incertitude* de Heisenberg : une première mesure (de la position ou de la vitesse) *modifie* le système observé.

La fonction d'onde φ est à valeurs complexes. Si elle est normalisée par $\displaystyle\int_{\mathbb{R}} \big|\varphi(t,x)\big|^2\,\mathrm{d}x = 1$, on appelle *probabilité de présence* le carré de son module $\big|\varphi(t,x)\big|^2$. La définition s'étend en dimensions 2 et 3.

Dans un problème tridimensionnel (i.e. avec trois variables d'espace (x, y, z), en coordonnés *cartésiennes*), on a un *vecteur impulsion* : $\mathbf{p} := m\left(\frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, \frac{\mathrm{d}z}{\mathrm{d}t}\right)$. Au carré de sa norme, noté $\mathbf{p}^2$, on fait correspondre l'opérateur $-\hbar^2\Delta$, où Δ est le laplacien.
En mécanique classique l'énergie totale est la somme de l'énergie cinétique et de l'énergie potentielle, soit $E = \frac{\mathbf{p}^2}{2m} + V(x)$, on l'appelle aussi l'hamiltonien du système. Pour un système *conservatif* c'est une intégrale première (*cf.* le module *Équations différentielles* du livre [L2]). Les règles de correspondance associent à E, d'une part : $H := i\hbar\frac{\partial}{\partial t}$, et, d'autre part : $-\frac{\hbar^2}{2m}\Delta + V$. On en déduit :

$$H := i\hbar\frac{\partial}{\partial t} = -\frac{\hbar^2}{2m}\Delta + V \tag{49}$$

et l'équation de Schrödinger :

$$i\hbar\frac{\partial\varphi(t, x, y, z)}{\partial t} = H\varphi(t, x, y, z) \tag{50}$$

C'est une équation aux dérivées partielles.
On s'intéresse aux états *stationnaires* : il s'agit, par définition des états pour lesquels la probabilité de présence est *indépendante du temps*. On a alors $\|\varphi(t, x, y, z)\| = \|\varphi(t_0, x, y, z)\|$, donc, en posant $\varphi(t_0, x, y, z) = \psi(x, y, z)$, on a $\varphi(t, x, y, z) = u(t)\psi(x, y, z)$, avec $u(t)$ de module 1, pour tout $t \in \mathbb{R}$. On est ainsi amené à chercher les solutions stationnaires de l'équation de Schrödinger par la *méthode de séparation des variables*. En reportant $\varphi(t, x, y, z) = u(t)\psi(x, y, z)$ dans l'équation et en divisant par φ (en supposant que cette fonction ne s'annule pas), on trouve :

$$i\hbar\frac{\frac{\partial u(t)}{\partial t}}{u(t)} = \frac{H\psi(x, y, z)}{\psi(x, y, z)}.$$

Le premier membre est fonction de t seul, le second seulement de (x, y, z), cette quantité est donc *constante*, et notée $E \in \mathbb{C}$, et l'on obtient un système :

$$i\hbar\frac{\partial u(t)}{\partial t} = Eu(t) \quad \text{et} \quad H\psi(x, y, z) = E\psi(x, y, z)$$

Donc $u(t) = \mathrm{e}^{-\frac{iE}{\hbar}t}$. La constante E étant valeur propre de l'opérateur autoadjoint H est *réelle*, ce qui entraîne $\|u\| = 1$. Ainsi, pour toute solution φ de *l'équation de Schrödinger stationnaire* :

$$H\psi(x, y, z) = E\psi(x, y, z), \tag{51}$$

la fonction $\varphi(t, x, y, z) = \mathrm{e}^{-\frac{iE}{\hbar}t}\psi(x, y, z)$ est une solution stationnaire de l'équation de Schrödinger.
L'équation de Schrödinger des états stationnaires d'un oscillateur harmonique, de masse m, à une dimension s'écrit donc :

$$H\psi(x) = \left[\frac{-\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \frac{1}{2}kx^2\right]\psi(x) = E\psi(x). \tag{52}$$

On la réécrit $\frac{\hbar^2}{m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi + (2E - m\omega^2x^2)\psi = 0$. La comparaison avec l'équation différentielle des fonctions d'Hermite $y'' + (2n + 1 - t^2)y = 0$ suggère alors de poser :

$$t := \sqrt{\frac{m\omega}{\hbar}}x, \quad f(t) := \psi(x) \quad \text{et} \quad E := \hbar\omega\left(\lambda + \frac{1}{2}\right).$$

La fonction f est alors solution de l'équation différentielle :

$$\frac{\mathrm{d}^2 f}{\mathrm{d}t^2} + (2\lambda + 1 - t^2)f = 0. \tag{53}$$

On reconnait l'équation différentielle des fonctions d'Hermite (*cf.* la page 592).

Selon l'interprétation de Max Born[12] de la fonction d'onde, $\int_A \|f\|^2 dx$ est la probabilité de trouver la particule dans le domaine A, si l'on effectue une mesure. La condition physique de *localisation* de l'état se traduit ainsi par le fait que la fonction f doit décroître « suffisamment vite » à l'infini, ce qui implique en particulier qu'elle doit être de carré sommable. Pour qu'il en soit ainsi, il faut et il suffit que $\lambda \in \mathbb{N}$ (on note $\lambda = n$); f est alors la $n^{ème}$ fonction d'Hermite (on applique le théorème 48 de la page 592 ; *cf.* aussi l'exercice type de la page 647). Ainsi, les niveaux d'énergie pour l'oscillateur harmonique sont $E_n = \hbar\omega\left(n + \frac{1}{2}\right)$ $(n \in \mathbb{N})$. Les fonctions d'onde correspondantes sont les $\psi_n(x) = C_n \Psi_n(t) = \Psi_n\left(\frac{\sqrt{m\omega}}{\hbar}x\right)$, où Ψ_n est la $n^{ème}$ fonction d'Hermite. L'état fondamental correspond à la plus basse énergie, donc $E_0 = \dfrac{\hbar\omega}{2}\cdot$ La fonction d'onde correspondante est $C_0 \mathrm{e}^{-\frac{m\omega}{\hbar}x^2}$ (C_0 est une constante de normalisation, que l'on peut prendre égale à $\dfrac{m\omega}{\pi\hbar}$, pour que le carré de $|\psi_0|$ soit une densité de probabilité).

À partir de *l'hypothèse* que le spectre d'énergie des états liés est *discret*, les physiciens utilisent le formalisme des *opérateurs de création-annihilation* (*cf.* la page 593) pour montrer que ce spectre est nécessairement l'ensemble $\{n + \frac{1}{2} |\ n \in \mathbb{N}\}$, puis reconstituer tous les états à partir de l'état fondamental. Cette approche est intéressante pour comprendre une interprétation *physique* du mécanisme *mathématique*, que nous avons décrit plus haut. Nous allons la détailler.

L'opérateur hamiltonien est $H = \frac{P^2}{2m} + \frac{1}{2}kX^2$. Par analogie avec le problème classique, on pose $\omega := \sqrt{k/m}$. On introduit les opérateurs : $Q_1 := \sqrt{\frac{m\omega}{\hbar}}X$, $P_1 := \frac{P}{\sqrt{m\hbar\omega}}$, $H_1 := \frac{H}{\hbar\omega}$ et l'on réécrit l'équation de Schrödinger :

$$H_1\psi = \frac{1}{2}(P_1^2 + Q_1^2)\psi = \frac{E}{\hbar\omega}\cdot$$

On remarquera que $\hbar\omega$ est le *quantum d'énergie* pour la longueur d'onde correspondant à ω et que l'équation est, du point de vue physique, *sans dimension*. On vérifie que :

$$[Q_1, P_1] = Q_1P_1 - P_1Q_1 = iI$$

(I est l'opérateur identité), ce qui traduit la relation de Heisenberg.

On définit alors des opérateurs de création et d'annihilation

$$A^+ = \frac{1}{\sqrt{2}}(Q_1 - iP_1) \quad \text{et} \quad A = \frac{1}{\sqrt{2}}(Q_1 + iP_1)$$

et l'on pose $N = A^+A$. On a $N = \frac{1}{2}(P_1^2 + Q_1^2) + \frac{i}{2}[Q_1, P_1] = H_1 - \frac{1}{2}I$. La recherche des vecteurs propres de l'hamiltonien est donc ramenée au problème analogue pour l'opérateur « nombre de quanta » N. En *admettant* l'existence d'un état fondamental (que les physiciens notent $|0>$ et appellent « le vide »), on peut alors montrer par des manipulations purement algébriques (*cf.* l'exercice III.4.18), que les valeurs propres de N sont les entiers $n \in \mathbb{N}$ et reconstituer les états correspondants (i.e. les vecteurs propres) en appliquant n fois l'opérateur de création : $|n>= (A^+)^n|0>$. Il en résulte que les valeurs propres de l'hamiltonien H (i.e. le spectre d'énergie) sont les $E_n = \hbar\omega\left(n + \frac{1}{2}\right)$. Le niveau fondamental correspond à $E_0 = \frac{\hbar\omega}{2}$ et la signification physique des opérateurs de création-annihilation est alors claire : A^+ (respectivement A^-) fait passer de l'état $|n>$, d'énergie E_n à l'état $|n+1>$, d'énergie E_{n+1} (respectivement $|n-1>$,

[12] Physicien allemand 1882-1970.

d'énergie E_{n-1} ; il *crée* donc un quantum d'énergie $\hbar\omega$ (respectivement il *annihile* un quantum d'énergie). En théorie des champs (i.e. dans le cadre quantique *relativiste*) les opérateurs de création-annihilation créent ou détruisent des *particules* (dans ce cadre, il y a une équivalence entre la masse m et l'énergie E : $E = mc^2 \ldots$).
Le lecteur intéressé approfondira aisément le parallèle entre les opérateurs physiques et mathématiques, en traduisant les calculs faits plus hauts (en comparant, en particulier, les fonctions d'Hermite ψ_n et les états $|n>$).

3.5 Polynômes de Laguerre

Nous avons vu que les polynômes de Laguerre généralisés $L_n^{(a)}$ forment une suite de polynômes orthogonaux associés au poids $w(t) = t^a e^{-t}$ $(a > -1)$ sur la demi-droite $[0, +\infty[$. Nous les avons normalisés, en imposant que le coefficient du terme de plus haut degré de $L_n^{(a)}$ soit $(-1)^n$.
La formule d'Olinde Rodrigues donne alors $L_n^a(t) = t^{-a} e^t \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n (t^{n+a} e^{-t})$.

Proposition 53. On a :

$$(L_n^{(a)} \mid L_m^{(a)})_w = \int_0^{+\infty} L_n^{(a)}(t) L_m^{(a)}(t) t^a e^{-t}\,\mathrm{d}t = n!\Gamma(n+1+a)\delta_{nm}.$$

Démonstration. Supposons $n \geqslant m$, $(L_n^{(a)} \mid L_m^{(a)})_w = \displaystyle\int_0^{+\infty} (t^{n+a} e^{-t})^{(n)} L_m^{(a)}\,\mathrm{d}t$. La formule (31) d'intégration par parties généralisée peut être appliquée (les conditions en $+\infty$ sont évidemment réalisées et, en 0, on utilise $a+1>0$) et l'on obtient :

$$(L_n^{(a)} \mid L_m^{(a)})_w = (-1)^n \int_0^{+\infty} t^{n+a} e^{-t} \left(L_m^a\right)^{(n)}(t)\,\mathrm{d}t.$$

On retrouve, si $n > m$, $(L_n^{(a)} \mid L_m^{(a)})_w = 0$ et, si $n = m$,

$$\left\|L_n^a\right\|^2 = n! \int_0^{+\infty} t^{n+a} e^{-t}\,\mathrm{d}t = n!\,\Gamma(n+a+1).$$

■

Relations de récurrence

On applique les résultats du théorème 4 de la page 538 et de l'exercice III.4.2 de la page 651. On a, avec les notations correspondantes, $B_n = A_n\left(\frac{\gamma_n'}{\gamma_n} - \frac{\gamma_{n-1}'}{\gamma_{n-1}}\right)$ et, en utilisant l'assertion (ii) du corollaire 34 de la page 579, $\gamma_n = (-1)^n$ et $\gamma_n' = (-1)^{n-1} n(a+n)$. Donc $A_n = -1$ et $B_n = -n(n+a) + (n-1)(n-1+a) = -(2n+a-1)$. Par ailleurs, $C_n = \dfrac{\left\|L_{n-1}^{(a)}\right\|^2}{\left\|L_{n-2}^{(a)}\right\|^2} = (n-1)(n+a-1)$. On en déduit la formule de récurrence suivante.

Proposition 54.

$$L_n^{(a)} + \big(t - (2n + a - 1)\big) L_{n-1}^{(a)} + (n-1)(n+a-1) L_{n-2}^{(a)} = 0.$$

On a une relation simple entre les polynômes d'Hermite et les polynômes de Laguerre généralisés de paramètre $\pm 1/2$.

Proposition 55. On a :

$$H_{2n}(t) = (-1)^n 2^{2n} L_n^{(-1/2)}(t^2) \quad \text{et} \quad H_{2n+1}(t) = (-1)^n 2^{2n+1} t L_n^{(1/2)}(t^2).$$

Démonstration. Pour établir ce résultat, on compare les expressions explicites de la proposition 59 (ii), page 607, et de la formule (34). ∎

Opérateurs de création et d'annihilation

On rappelle l'équation différentielle des polynômes de Laguerre généralisés (*cf.* (26)) :

$$ty'' + (1 + a - t)y' + ny = 0.$$

On peut l'écrire $D_a y = -ny$, avec $D_a := t\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^2 + (1 + a - t)\frac{\mathrm{d}}{\mathrm{d}t}$. Ainsi, les polynômes de Laguerre $L_n^{(a)}$ sont des fonctions propres de l'opérateur D_a associées à la valeur propre $-n$.
On définit $\partial_a^* := t\frac{\mathrm{d}}{\mathrm{d}t} + a + 1 - t$, $\partial_a := -\frac{\mathrm{d}}{\mathrm{d}t}$ et $\mathcal{N}' := \partial_a^* \partial_a$. On vérifie :

$$\partial_a = t^{-a}\mathrm{e}^t \frac{\mathrm{d}}{\mathrm{d}t}\left(t^{1+a}\mathrm{e}^{-t}\right) \quad \text{et} \quad \mathcal{N}' = -D_a.$$

On peut alors reprendre le formalisme des opérateurs de création et d'annihilation (*cf.* la page 593). Le lecteur pourra le faire à titre d'exercice.

3.5.1 Atome d'hydrogène

Nous rappelons brièvement les bases physiques de l'étude de l'atome d'hydrogène. Pour plus de détails, le lecteur pourra se reporter par exemple à [PCP].

Les inverses des longueurs d'ondes des *raies spectrales* de l'atome d'hydrogène, sont proportionnelles à $\frac{1}{n^2} - \frac{1}{m^2}$, l'entier m numérotant les raies d'une « série » définie par n. Ce résultat a été découvert (ou plutôt deviné…) par Balmer (un professeur de mathématiques du lycée de jeunes-filles de Bâle), en 1885, à partir de données expérimentales pour les quatre raies visibles (dans ce cas, $n = 2$: c'est la *série de Balmer*). D'autres séries de raies ont été plus tard découvertes dans le spectre invisible : Lyman (dans l'ultraviolet, $n = 1$), Paschen (dans l'infrarouge, $n = 3$), Brackett ($n = 4$), Pfund ($n = 5$), Humphreys ($n = 6$) ; on a ensuite renoncé à donner des noms…
L'émission d'une raie spectrale de fréquence ν (proportionnelle à $1/\lambda$) correspond, selon le principe fondamental de la théorie des quanta de Max Planck, à une diminution d'énergie de l'atome de $\Delta E = h\nu$ (où h est la constante de Planck). Ainsi les niveaux d'énergie possibles pour l'atome d'hydrogène forment une suite discontinue. C'est l'un des points de départ essentiels de la mécanique quantique et c'est l'origine de l'article fondateur d'Erwin Schrödinger, où apparaît, pour la première fois, la célèbre équation qui porte son nom. Nous allons voir, en reprenant pour l'essentiel l'argumentation initiale de Schrödinger, que la *clé mathématique* de la question est la famille des

polynômes de Laguerre généralisés, via leurs équations différentielles. Cela nous permettra d'expliquer l'apparition des $\frac{1}{n^2}$ dans le spectre d'énergie et donc de justifier théoriquement la formule de Balmer et la découverte des séries de ses successeurs. L'équation différentielle de Laguerre (dont la source est un problème de fractions continues et de sommation de série divergente, relevant des mathématiques dites pures, *cf.* 4.2.3) a ainsi joué un rôle fondamental dans l'évolution de la physique du XXe siècle.

Pour mettre en évidence les idées centrales, nous ne détaillerons pas ci-dessous un certain nombre de calculs aussi élémentaires que fastidieux. Un état de l'atome H est « étiqueté » par *trois nombres quantiques* m, ℓ, n. Les deux premiers m, ℓ proviennent de la quantification fournie par les fonctions sphériques, en relation avec les polynômes de Lagrange P_ℓ et les fonctions associées P_ℓ^m, comme nous l'avons vu plus haut. Nous ne redétaillerons donc pas ce point, nous concentrant sur le troisième nombre quantique n, qui quantifie l'énergie E et explique les séries de Balmer. Ce nombre n, à ℓ fixé, apparaît dans l'équation radiale, sur laquelle nous allons nous concentrer. Nous allons voir que la quantification correspond exactement, dans ce cas, à l'apparition de familles de polynômes orthogonaux (c'est un exemple concret de relation entre polynômes orthogonaux classiques et problème du type Sturm-Liouville).

L'équation de Schrödinger des états stationnaires d'une particule de masse m' dans un champ central symétrique de potentiel $V(r)$ s'écrit $-\frac{\hbar^2}{2m'}\Delta\psi + V(r)\psi = E\psi$, où : $\hbar = \frac{h}{2\pi}$, ψ est la fonction d'onde (elle est à valeurs *complexes*) et E l'énergie[13]

Pour l'atome d'hydrogène H, qui est formé d'un proton et d'un électron (de masses respectives m_p et m_e), on se ramène au cas ci-dessus : m' est la masse réduite ($m' = \frac{m_e m_P}{m_e + m_p}$), comme dans le cas classique du problème à deux corps de Kepler (*cf.* le module *Équations différentielles* du livre [L2]) et $V(r) = -\frac{1}{4\pi\varepsilon_0}\frac{e^2}{r}$ (potentiel coulombien d'interaction électrostatique). Pour simplifier, nous remplacerons $4\pi\varepsilon_0$ par 1. Notre équation est donc $\big($*cf.* (49) et (51)$\big)$:

$$-\frac{\hbar^2}{2m'}\Delta\psi - \frac{e^2}{r}\psi = E\psi.$$

Comme dans le cas de l'EDP de Laplace étudié plus haut (*cf.* page 585), on passe en coordonnées sphériques et l'on cherche la fonction d'onde sous forme de produit : $\psi(r,\theta,\varphi) = R(r)\Theta(\theta)\Phi(\varphi)$ (et l'on note $Y := \Theta\Phi$). Un calcul semblable à celui déjà fait nous donne trois équations différentielles d'ordre 2, découplées (i.e. portant séparément sur chaque variable). Les deux équations angulaires sont les mêmes (*cf.* $\big(44\big)$ et (45)) : elles sont équivalentes à $\Delta_S Y = -\ell(\ell+1)Y$. Si $\Theta_{\ell,m}$ et Φ_m sont des solutions respectives de ces équations, on pose $Y_{\ell m} := \Theta_{\ell,m}\Phi_m$ (l'entier n utilisé plus haut dans la section 3.3.3 est noté ici ℓ, en conformité avec la convention des physiciens). La fonction $Y_{\ell m}$ est une *fonction propre* du laplacien sphérique Δ_S, associée à la *valeur propre* $\ell(\ell+1)$. L'équation radiale, i.e. l'équation sur r s'écrit $\big($après un calcul analogue à celui qui nous a conduit à l'équation (43)$\big)$:

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\big(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\big) - \frac{\ell(\ell+1)}{r^2}R + \frac{2m'}{\hbar^2}\big(E + \frac{e^2}{r}\big)R = 0 \tag{54}$$

On fait les changements de variable et de paramètres : $n^2 = -\frac{m'e^4}{2\hbar^2 E}$ $(E < 0)$ et $t := \frac{2m'e^2}{n\hbar^2}r$ ($r > 0$, donc $t > 0$), et l'on obtient (avec un abus de notation pour R qui reste inchangé) :

$$\frac{\mathrm{d}^2R}{\mathrm{d}t^2} + \frac{2}{t}\frac{\mathrm{d}R}{\mathrm{d}t} + \Big[-\frac{\ell(\ell+1)}{t^2} - \frac{1}{4} + \frac{n}{t}\Big]R = 0 \tag{55}$$

[13] Erwin Schrödinger 1926 ; *cf. Quantisation as a Problem of Proper Values (Part I)*, dans *Collected Papers on Wave Mechanics,* Blackie & Son Limited, 1928.

Nous cherchons des *états liés*, c'est-à-dire que la fonction d'onde cherchée ψ doit être régulière en 0 et décroître « très vite » à l'infini ($\|\psi\|^2$ représente la « probabilité de présence » de l'électron, elle doit donc être pratiquement nulle loin du noyau). Une étude rigoureuse du comportement des solutions en 0 et $+\infty$ permettrait de confirmer des « approximations de physiciens » et de voir que $t^{-l}R(t)$ tend vers une limite finie, non nulle, quand t tend vers 0 et que, si R est borné, il est « de l'ordre de » $e^{-t/2}$, quand t tend vers $+\infty$. Cela suggère de faire le changement de fonction inconnue $R(t) = t^\ell e^{-t/2}u(t)$: c'est analogue au passage des fonctions d'Hermite aux polynômes d'Hermite. On obtient, en posant $k = n - \ell - 1$, l'équation différentielle :

$$t\frac{\mathrm{d}^2u}{\mathrm{d}t^2} + (1 + 2\ell - t)\frac{\mathrm{d}u}{\mathrm{d}t} + ku = 0. \tag{56}$$

Si k est entier, on reconnaît une équation différentielle de Laguerre généralisée, de paramètre $a = \ell$. Supposons, pour l'instant, $k \in \mathbb{R}$ quelconque.

Nous sommes maintenant au cœur de notre problème. La solution u cherchée doit être non nulle à l'origine et y être développable en série. Nous sommes donc amenés à étudier les solutions de (56), sur $\mathbb{R}_+$, développables en série à l'origine, et en particulier, leur comportement quand t tend vers $+\infty$. On commence par chercher une solution sous forme d'une série entière formelle $\hat{u} \in \mathbb{R}[[T]]$: $\hat{u} = \sum\limits_{h\in\mathbb{N}} a_h t^h$. En reportant dans l'équation, et en identifiant les termes en x^h, on trouve la relation de récurrence $(h+1)(h+2\ell+2)a_{h+1} = (h-k)a_h$ pour les coefficients.

On suppose $a_0 \neq 0$ (sinon la solution est identiquement nulle). Si, quel que soit $h \in \mathbb{N}$, $h - k$ ne s'annule jamais (c'est-à-dire si k n'est pas un entier positif), on a, quand h tend vers $+\infty$: $a_{h+1} \sim \frac{1}{h}a_h$. On en déduit que la série $\hat{u}$ est *convergente*, de rayon de convergence infini. On note u sa somme. On peut choisir le signe de a_0 tel que a_h soit positif, pour h assez grand.

Soit $0 < \gamma < 1$. Il existe $h_0 \in \mathbb{N}^*$ tel que, pour $h > h_0$, l'on ait $a_h > 0$ et $a_{h+1} > \frac{\gamma}{h}a_h$. On a :

$$\sum_{h=h_0}^{+\infty} a_h t^h \geqslant a_{h_0}\left(t^{h_0} + \frac{\gamma}{h_0}t^{h_0+1} + \frac{\gamma^2}{h_0(h_0+1)}t^{h_0+2} + \cdots\right),$$

donc

$$\sum_{h=h_0}^{+\infty} a_h t^h \geqslant t\frac{a_{h_0}(h_0-1)!}{\gamma^{h_0-1}}\left(\frac{(\gamma t)^{h_0-1}}{(h_0-1)!} + \frac{(\gamma t)^{h_0}}{h_0!} + \frac{(\gamma t)^{h_0+1}}{(h_0+1)!} + \cdots\right).$$

Pour $t_0 > 0$ assez grand, on a $t_0\frac{a_{h_0}(h_0-1)!}{\gamma^{h_0-1}} > 1$, donc :

$$\forall t \geqslant t_0,\quad \sum_{h=h_0}^{+\infty} a_h t^h \geqslant \mathrm{e}^{\gamma t} - P(t),$$

où P es un polynôme. Quitte à augmenter t_0, on a $\mathrm{e}^{\gamma t} - P(t) > \frac{1}{2}\mathrm{e}^{\gamma t}$, donc,

$$\forall t \geqslant t_0,\quad \sum_{h=h_0}^{+\infty} a_h t^h \geqslant \frac{1}{2}\mathrm{e}^{\gamma t}.$$

Il en résulte que $R(t) = t^l\mathrm{e}^{-t/2}u(t) > \frac{1}{2}t^l e^{(\gamma-\frac{1}{2})t}$, et donc que $R(t)$ tend vers $+\infty$ quand t (ou de façon équivalente r) tend vers $+\infty$. Cela est évidemment incompatible avec la contrainte physique relative aux états liés (c'est-à-dire $\lim\limits_{t\to+\infty} R(t) = 0$). Donc, on a nécessairement, pour un état lié, $k \in \mathbb{N}$: c'est ici qu'apparaît la condition de « quantification radiale » pour les solutions (k est appelé nombre quantique radial). La solution u est alors un *polynôme*, et plus précisément un *polynôme de Laguerre généralisé* : $u = L_k^{(\ell)}$.

Pour ℓ et k fixés, on pose : $n := \ell + k + 1$. En remontant les calculs (et en remplaçant 1 par $4\pi\varepsilon_0^2$ dans le potentiel électrostatique simplifié), on obtient :

$$E_n = -\frac{m'e^4}{32\pi^2\varepsilon_0^4\hbar^2}\frac{1}{n^2},$$

ce qui explique que l'on a un spectre énergétique discret et la formule de Balmer. Le niveau le plus bas (ou *niveau fondamental*) est $E_1 = -\frac{m'e^4}{32\pi^2\varepsilon_0^4\hbar^2}$ (on a alors $\ell = k = 0$), la chute d'un électron vers ce niveau correspond à une raie de la série de Lyman.

On notera, en reprenant nos calculs, que les fonctions d'onde des états liés de H (indexées par les nombres quantiques (m, ℓ, n)) sont toutes *explicitement calculables* à l'aide des fonctions trigonométriques et exponentielles et des polynômes de Legendre et de Laguerre généralisés (de paramètre entier positif).

Nous avons ainsi montré que les deux modèles fondamentaux de la mécanique quantique : l'oscillateur harmonique et l'atome d'hydrogène sont intégrables par les fonctions élémentaires et les polynômes orthogonaux (d'Hermite, Legendre, Laguerre).

Remarque. On peut considérer l'oscillateur harmonique en dimension 3. Son équation de Schrödinger stationnaire est $H\psi(x) = \left[\frac{-\hbar^2}{2m}\Delta + \frac{1}{2}k\rho^2\right]\psi(x) = E\psi(x)$: on remplace dans (52) l'opérateur $\mathrm{d}^2/\,\mathrm{d}x^2$ par le laplacien tridimensionnel Δ et x^2 par $\rho^2 = x^2 + y^2 + z^2$). Cette équation se réécrit $\frac{\hbar^2}{m}\Delta\psi + (2E - m\omega^2\rho^2)\psi = 0$. En appliquant la méthode de séparation des variables (*cf.* le point méthode de la page 586), on obtient l'équation radiale que l'on transforme par un changement de variable élémentaire $\big($*cf.* (53)$\big)$ en l'équation :

$$\frac{\mathrm{d}^2 f}{\mathrm{d}t^2} - \frac{\ell'(\ell'+1)}{t^2} + (2\lambda + 1 - t^2)f = 0. \tag{57}$$

Considérons l'équation radiale pour l'atome d'hydrogène

$$\frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}R}{\mathrm{d}r}\right) + \left(-\frac{\ell(\ell+1)}{r^2} + a + \frac{b}{r}\right)R = 0$$

($a := \frac{2m'E}{\hbar^2}, b := \frac{2m'e^2}{\hbar^2}$). On pose $t = \sqrt{r}$ et $R(r) = t^{\alpha-2}f(t)$. On obtient alors l'équation différentielle :

$$\frac{\mathrm{d}^2 f}{\mathrm{d}t^2} + \frac{2\alpha - 1}{t}\frac{\mathrm{d}f}{\mathrm{d}t} + \Big(\frac{\alpha(\alpha-2) - 4\ell(\ell+1)}{t^2} + 4b + 4at^2\Big)f = 0.$$

On choisit $\alpha := 1/2$, ce qui élimine le terme en $\frac{\mathrm{d}f}{\mathrm{d}t}\cdot$ En posant $\ell' = 2\ell + \frac{1}{2}, b := \frac{\lambda}{2} + \frac{1}{4}$ et $a := -\frac{1}{4}$, on retrouve l'équation différentielle (57) ci-dessus.

On obtient ainsi un « dictionnaire » entre l'étude de l'atome d'hydrogène et celle de l'oscillateur harmonique. Il y a d'intéressantes relations entre les paramètres physiques des deux côtés. Du point de vue mathématique, c'est une traduction de la formule (55). Signalons qu'il existe en mécanique classique une « transmutation » entre potentiel de Newton (en k/r) et potentiel de Hooke (en $k'r^2$). Cette relation (que l'on trouve dans les *Principia* d'Isaac Newton), met en correspondance deux descriptions du mouvement d'un corps sur une trajectoire elliptique : vu du centre ou vu d'un foyer. Cette transmutation (trouvée vers 1684) a sans doute joué un rôle clé dans la découverte de la loi de gravitation par Newton. La transmutation a été plus tard étendue à des potentiels en k/r^n et $k'/r^{n'}$, où n, n' satisfont la relation *involutive* $(3-n)(3-n') = 4$, par divers auteurs, dont Goursat en 1889.

3.6 Séries génératrices

Définition 9. À une suite de polynômes orthogonaux $(p_n)_{n\in\mathbb{N}}$, on associe les *séries formelles* :

$$\sum_{n\in\mathbb{N}} p_n(t)x^n \quad \text{et} \quad \sum_{n\in\mathbb{N}} \frac{p_n(t)}{n!}x^n.$$

La première est appelée *série génératrice* des $(p_n)_{n\in\mathbb{N}}$, la seconde *série génératrice exponentielle* des $(p_n)_{n\in\mathbb{N}}$.

Le passage de l'une à l'autre série correspond à un changement de normalisation pour les polynômes orthogonaux. Il s'agit de séries formelles à coefficients dans l'anneau $A = \mathbb{R}[T]$ des polynômes à une variable. À ces séries formelles on associe des séries de fonctions polynômiales de deux variables dont on étudie la convergence, la somme...
Soient a et b deux réels strictement positifs. Si la série de fonctions polynomiales $\sum\limits_{n\in\mathbb{N}} p_n(t)x^n$ converge simplement sur $]-a,a[\times]-b,b[$ vers une fonction F, on dit que F est une *fonction génératrice* de la suite de polynômes orthogonaux $(p_n)_{n\in\mathbb{N}}$. On définit de façon analogue une *fonction génératrice exponentielle*.
Nous nous limiterons aux cas de Legendre, Laguerre et Hermite.

Série génératrice des polynômes de Legendre

Les polynômes de Legendre interviennent dans diverses questions de *théorie du potentiel*, nous en avons vu un exemple plus haut, avec l'étude de l'équation de Laplace. En voici un autre, voisin et plus simple, qui conduit naturellement à la série génératrice des polynômes de Legendre, à partir d'un problème élémentaire de physique.
On considère dans le plan « physique » une origine O et un point Q fixes. On suppose placée, en Q, une charge électrique, et l'on veut calculer son potentiel coulombien en un point P du plan, tel que $OP > OQ$. En choisissant des unités de distance et de charge électrique convenables, on peut supposer que $OP = 1$, que $0 \leqslant OQ = x < 1$ et que le potentiel coulombien est $\frac{1}{QP}$. On note θ l'angle $(\overrightarrow{OQ}, \overrightarrow{OP})$. On a, en posant $t = \cos\theta$, $QP = \sqrt{1-2xt+x^2}$. Le potentiel cherché est donc $\frac{1}{\sqrt{1-2xt+x^2}}$.
Posons $y = 2xt - x^2$. Supposons que $|y| < 1$. On sait que $(1-y)^{-1/2} = \sum\limits_{n\geqslant 0} \alpha_n y^n$, avec $\alpha_n = \frac{1}{2^{2n}}\binom{2n}{n}$.
On a (série binomiale convergente) :

$$\begin{aligned}(1-y)^{-1/2} &= 1 + \frac{1}{2}y + \frac{3}{8}y^2 + \frac{5}{16}y^3 + \cdots \\ &= 1 + \frac{1}{2}(2xt-x^2) + \frac{3}{8}(2xt-x^2)^2 + \frac{5}{16}(2xt-x^2)^3 + \cdots .\end{aligned}$$

En développant chaque terme, puis en réordonnant suivant les puissances de x (i.e. en considérant, pour chaque valeur de t, la série obtenue par composition d'une série en y

par une série en x, qui est de fait un polynôme), on obtient une série en x de la forme : $\sum_{n\in\mathbb{N}} r_n(t)x^n$, où les r_n sont des polynômes en t (le lecteur le vérifiera à titre d'exercice).
En calculant les premiers termes, on trouve :

$$\sum_{n\in\mathbb{N}} r_n(t)x^n = 1 + tx + \frac{1}{2}(3t^2-1)x^2 + \frac{1}{2}(5t^3-3)x^3 + \cdots .$$

On voit apparaître les premiers polynômes de Legendre : $r_0 = P_0$, $r_1 = P_1$, $r_2 = P_2$, $r_3 = P_3$. Nous allons montrer que $r_n = P_n$, pour tout $n \in \mathbb{N}$.
Si $|t| \leqslant 1$ et $|x| < 1/3$, on a $|y| \leqslant u := 2|x||t| + |x|^2 < 1$. La série à termes positifs $\sum_{n\geqslant 0} \alpha_n u^n$ est convergente. On peut donc appliquer les résultats sur la convergence des séries doubles à termes positifs (*cf.* le théorème de Fubini pour les séries absolument convergentes du module *Suites et séries de fonctions* du livre [L2]), par suite $\sum_{n\in\mathbb{N}} r_n(t)x^n$ converge et :

$$\frac{1}{\sqrt{1-2xt+x^2}} = \sum_{n=0}^{+\infty} r_n(t)x^n. \tag{58}$$

Notons que la convergence de $\sum_{n=0}^{+\infty} r_n(t)x^n$ est absolue et *uniforme* en $t \in [-1,1]$.
En faisant $t := 1$ dans (58), on trouve $r_n(1) = 1$.
Supposons $|t| \leqslant 1$, et $0 < x < 1/3$, $0 < \xi < 1/3$. On a (en utilisant la convergence absolue des deux séries) :

$$\frac{1}{\sqrt{1-2xt+x^2}}\frac{1}{\sqrt{1-2\xi t+\xi^2}} = \sum_{n=0}^{+\infty}\sum_{m=0}^{+\infty} r_n(t)r_m(t)x^n\xi^m. \tag{59}$$

On calcule :

$$\int_{-1}^{1} \frac{1}{\sqrt{1-2xt+x^2}}\frac{1}{\sqrt{1-2\xi t+\xi^2}}\,\mathrm{d}t = \frac{1}{\sqrt{x\xi}}\ln\frac{1+\sqrt{x\xi}}{1-\sqrt{x\xi}},$$

en remarquant qu'une primitive de la fonction $t \mapsto \frac{1}{\sqrt{(a-t)(b-t)}}$ sur $]-\infty, \min(a,b)[$ est $t \mapsto -2\ln\bigl(\sqrt{a-t}+\sqrt{b-t}\bigr)$.
En développant, on trouve :

$$\frac{1}{\sqrt{x\xi}}\ln\frac{1+\sqrt{x\xi}}{1-\sqrt{x\xi}} = \sum_{n=0}^{+\infty}\frac{2}{2n+1}x^n\xi^n.$$

À x et ξ fixés, avec $0 < x < 1/3$ et $0 < \xi < 1/3$, la convergence de la série de fonctions $t \mapsto r_n(t)r_m(t)x^n\xi^m$ est *uniforme* pour $t \in [-1,1]$, on peut donc intégrer terme à terme. En identifiant les coefficients des deux développements, on obtient (*cf.* l'exercice III.4.23 pour la justification) : $\int_{-1}^{1} r_n(t)r_m(t)\,\mathrm{d}t = 0$ si $n \neq m$ et $\int_{-1}^{1} r_n(t)r_m(t)\,\mathrm{d}t = \dfrac{2}{2n+1}$, si $n = m$. Ainsi $(r_n)_{n\in\mathbb{N}}$ est une suite de polynômes orthogonaux pour la mesure de Legendre. Comme ces polynômes vérifient $r_n(1) = 1$, ce sont les polynômes de Legendre.

Théorème 56. On a, pour tout $t \in [-1,1]$ et tout $x \in \mathbb{R}$, $|x| < 1/3$:

$$\frac{1}{\sqrt{1-2xt+x^2}} = \sum_{n=0}^{+\infty} P_n(t)x^n,$$

où les P_n sont les polynômes de Legendre.

La fonction $(t,x) \mapsto \frac{1}{\sqrt{1-2xt+x^2}}$ ($|t| \leqslant 1$, $|x| < 1/3$) est, par définition, la fonction génératrice des polynômes de Legendre.
On peut montrer que le résultat ci-dessus reste vrai pour $|x| < 1$ par un argument de prolongement analytique en utilisant $|P_n(t)| \leqslant 1$, pour tout $n \in \mathbb{N}$ et pour tout $t \in]-1,1[$.

Série génératrice exponentielle des polynômes de Laguerre

Rappelons la formule (*cf.* (34)) :

$$L_n(t) = e^t \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n (t^n e^{-t}) = n! \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{t^k}{k!}.$$

On considère la série *formelle* en x à coefficients polynomiaux en t :

$$\sum_{n\geqslant 0} \frac{L_n(t)}{n!} x^n = \sum_{n\geqslant 0} \left(\sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{t^k}{k!} \right) x^n.$$

C'est la série génératrice exponentielle des polynômes de Laguerre
La série double *formelle* associée est $\sum\limits_{k\geqslant 0} \sum\limits_{n\geqslant 0} \frac{(-1)^k t^k}{k!} \binom{n}{k} x^n$.

Si $x \in]-1,1[$, la série $\sum\limits_{n\geqslant k} \binom{n}{k} x^n$ converge et sa somme est $\frac{x^k}{(1-x)^{k+1}}$.

On a $\left|\frac{x^k}{(1-x)^{k+1}}\right| \leqslant \frac{1}{(1-|x|)^{k+1}}$. La série $\sum\limits_{k\geqslant 0} \frac{1}{k!} \frac{1}{(1-|x|)^k} |t|^k$ est convergente pour tout $t \in \mathbb{R}$, la série double $\sum\limits_{k\geqslant 0} \sum\limits_{n\geqslant k} \frac{(-1)^k t^k}{k!} \binom{n}{k} x^n$ est donc *absolument convergente* et l'on peut permuter les sommations (théorème de Fubini). Ainsi, la série génératrice des polynômes de Laguerre converge, pour tout $t \in \mathbb{R}$ et tout $x \in]-1,1[$, et l'on a :

$$\sum_{n=0}^{+\infty} \frac{L_n(t)}{n!} x^n = \sum_{k=0}^{+\infty} \frac{(-1)^k t^k}{k!} \sum_{n=k}^{+\infty} \binom{n}{k} x^n = \sum_{k=0}^{+\infty} \frac{(-1)^k}{k!} \frac{x^k}{(1-x)^{k+1}} t^k = \frac{e^{-\frac{x}{1-x}t}}{1-x}. \tag{60}$$

Théorème 57. On a, pour tout $t \in \mathbb{R}$ et tout $x \in]-1,1[$:

$$\psi(t,x) := \frac{e^{-\frac{x}{1-x}t}}{1-x} = \sum_{n=0}^{+\infty} \frac{L_n(t)}{n!} x^n.$$

La fonction ψ est, par définition, la *fonction génératrice exponentielle des polynômes de Laguerre*.

Proposition 58.
La fonction $\psi^a(t,x) := (1-x)^{-a}\psi(t,x)$ $(a > -1, x \in]-1,1[, t \in \mathbb{R})$ est une fonction génératrice exponentielle des polynômes de Laguerre généralisés L_n^a.

La preuve de ce résultat fait l'objet de l'exercice III.4.20

Série génératrice exponentielle des polynômes d'Hermite

La proposition suivante donne la *fonction génératrice exponentielle des polynômes d'Hermite* et une nouvelle preuve de la proposition 45 de la page 591.

Proposition 59. Soit $G(t,z) = e^{2tz-z^2}$. Pour tout $t \in \mathbb{R}$, $z \mapsto G(t,z)$ est une *fonction entière*, la série $\displaystyle\sum_{n\in\mathbb{N}} H_n(t)\frac{z^n}{n!}$ converge et l'on a :

(i) $\displaystyle G(t,z) = \sum_{n=0}^{+\infty} H_n(t)\frac{z^n}{n!}$;

(ii) $\displaystyle H_n(t) = \sum_{k=0}^{[n/2]} n!\,\frac{(-1)^k}{k!}\frac{(2t)^{n-2k}}{(n-2k)!}$.

La fonction G est, par définition, la fonction génératrice exponentielle des polynômes d'Hermite.

Démonstration. Soit $t \in \mathbb{R}$, la fonction $z \mapsto G(t,z) := e^{2tz}e^{-z^2}$ est le produit de deux fonctions entières, elle est donc entière. On remarque $e^{-t^2}G(t,z) = e^{-(t-z)^2}$, la fonction $z \mapsto e^{-t^2}G(t,z)$ étant entière, elle est égale à la somme de sa série de Taylor à l'origine. Pour tout réel x, $\frac{\partial^n}{\partial x^n}e^{-(t-x)^2} = e^{-(t-x)^2}H_n(t-x)$.

Ainsi $e^{-t^2}G(t,z) = \sum\limits_{n=0}^{+\infty} e^{-t^2}H_n(t)\frac{z^n}{n!}$, soit $G(t,z) = \sum\limits_{n=0}^{+\infty} H_n(t)\frac{z^n}{n!}$.

En faisant le produit des deux séries entières $\sum\limits_{n=0}^{+\infty}\frac{(2t)^n z^n}{n!}$ et $\sum\limits_{n=0}^{+\infty}(-1)^n\frac{z^{2n}}{n!}$, de rayon de convergence $+\infty$ et de sommes respectives e^{2tz} et e^{z^2}, on obtient :

$$G(t,z) = \sum_{n=0}^{+\infty}\Big(\sum_{k=0}^{[n/2]} n!\frac{(-1)^k}{k!}\frac{(2t)^{n-2k}}{(n-2k)!}\Big)z^n.$$

On en déduit $H_n(t) = \sum\limits_{k=0}^{[n/2]} n!\frac{(-1)^k}{k!}\frac{(2t)^{n-2k}}{(n-2k)!}$, pour tout $n \in \mathbb{N}$. ∎

3.7 Propriétés de densité pour les polynômes d'Hermite et de Laguerre

Nous avons vu (*cf.* le théorème 15 de la page 549) que, dans le cas d'un intervalle *borné* (α, β), il y a un résultat de densité : toute fonction f continue sur $]\alpha, \beta[$, de carré intégrable pour le poids w, est limite, au sens de la norme w, de la série $\sum(f \mid q_n)\, q_n$ (où $(q_n)_{n\in\mathbb{N}}$ est la suite des polynômes orthonormés associés à w). Si l'intervalle n'est pas borné, cette propriété de densité peut être vraie ou non. Nous allons voir qu'elle demeure vraie dans les cas des polynômes de Laguerre et d'Hermite.

Nous allons donner une preuve élémentaire du théorème de densité dans le cas de Laguerre que l'on peut modifier facilement pour obtenir le cas de Laguerre généralisé (le cas d'Hermite s'en déduira). Cette preuve (dont l'idée est due à J. von Neumann) est basée sur la série génératrice exponentielle des *fonctions* de Laguerre $(\varphi_n)_{n\in\mathbb{N}}$ $\big(\varphi_n(t) := e^{-t/2}L_n(t)\big)$.

Proposition 60. Soit $t \mapsto w(t) := \mathrm{e}^{-t}$ le poids de Laguerre sur $[0, +\infty[$.
Les fonctions polynomiales forment un sous–espace vectoriel dense dans $L_w(]0, +\infty[)$.

Démonstration. Notons E l'adhérence de $\mathbb{R}[T]$ dans $L_w(]0, +\infty[)$, pour la topologie définie par la norme $\| \|_w$. On note $L_{\{1\}}(]0, +\infty[)$, l'espace des fonctions *continues* de carré sommable sur $]0, +\infty[$ (sans poids, c'est-à-dire avec le poids 1), la norme correspondante étant notée $\| \|$ (qui est la norme L^2 classique). On désigne par F l'adhérence du $\mathbb{R}$-espace vectoriel $\mathbb{R}[T]e^{-t/2}$ dans $L_{\{1\}}(]0, +\infty[)$ (c'est l'adhérence du sous-espace engendré par les fonctions de Laguerre). L'application $f \mapsto f\mathrm{e}^{-t/2}$ est un *isomorphisme* de $L_w(]0, +\infty[)$ sur $L_{\{1\}}(]0, +\infty[)$. C'est clairement une isométrie :

$$\|f\|_w^2 = \int_0^{+\infty} f^2(t)\mathrm{e}^{-t}\,\mathrm{d}t = \int_0^{+\infty} (f(t)e^{-t/2})^2\,\mathrm{d}t.$$

Elle induit un isomorphisme isométrique de $\big(E, \| \|_w\big)$ sur $\big(F, \| \|\big)$. On est donc ramené à montrer que $F = L_{\{1\}}(]0, +\infty[)$, c'est-à-dire que le sous-espace vectoriel engendré par les *fonctions de Laguerre* $(\varphi_n)_{n\in\mathbb{N}}$ est *dense* dans $L_{\{1\}}(]0, +\infty[)$ (pour la topologie définie par la norme $\| \|$).

L'idée de la preuve est de se ramener au cas d'un intervalle borné en utilisant l'inclusion d'ensembles de fonctions $\{t \mapsto \mathrm{e}^{-\gamma t}\}_{\gamma>0} \subset F$. On commence donc par établir cette inclusion (en utilisant la série génératrice exponentielle des polynômes de Laguerre).

Considérons, pour $x \in]-1, 1[$ et $t \in \mathbb{R}$, l'identité :

$$\psi(t, x) := \frac{\mathrm{e}^{-\frac{xt}{1-x}}}{1-x} = \sum_{n=0}^{+\infty} \frac{1}{n!} L_n(t)x^n.$$

En la multipliant par $e^{-t/2}$, on obtient :

$$\varphi(t, x) := \frac{e^{-\frac{1}{2}\frac{1+x}{1-x}t}}{1-x} = \sum_{n=0}^{+\infty} \frac{1}{n!}\varphi_n(t)x^n. \tag{61}$$

La fonction φ est, par définition et par abus de langage, la fonction génératrice exponentielle des *fonctions de Laguerre* $\big($les $(-1)^n \frac{x^n}{n!}\varphi_n(t)$ sont les fonctions de Laguerre orthonormées$\big)$.

Lemme 61. On a les relations :

(i) $\displaystyle\int_0^{+\infty} \varphi(t,x)^2 \,\mathrm{d}t = \frac{1}{1-x^2}$;

(ii) $\displaystyle\int_0^{+\infty} \varphi(t,x)\varphi_n(t) \,\mathrm{d}t = n!x^n$;

(iii) $\displaystyle\int_0^{+\infty} \Big(\varphi(t,x) - \sum_{k=0}^{n} \frac{x^k}{k!}\varphi_k(t)\Big)^2 \,\mathrm{d}t = \frac{1}{1-x^2} - \sum_{k=0}^{n} x^{2k}$.

Démonstration. On utilise la première égalité de (61).
La relation (i) résulte d'un calcul facile d'intégrale.

Montrons (ii). Notons $I_n := \displaystyle\int_0^{+\infty} \varphi(t,x)\varphi_n(t) \,\mathrm{d}t$. On a :

$$I_n = \frac{1}{1-x}\int_0^{+\infty} \mathrm{e}^{-\frac{1}{2}\frac{1+x}{1-x}t}\mathrm{e}^{-t/2}L_n(t)\,\mathrm{d}t = \frac{1}{1-x}\int_0^{+\infty} \mathrm{e}^{-\frac{t}{1-x}}L_n(t)\,\mathrm{d}t.$$

On rappelle :

$$L_n(t) = n!\sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{t^k}{k!}$$

Pour tout $k \in [\![0,n]\!]$, on a (en utilisant $\Gamma(k+1) = k!$) :

$$\frac{1}{1-x}\int_0^{+\infty} \mathrm{e}^{-\frac{t}{1-x}}t^k\,\mathrm{d}t = (1-x)^k\int_0^{+\infty} \mathrm{e}^{-\frac{t}{1-x}}\left(\frac{t}{1-x}\right)^k \mathrm{d}\frac{t}{1-x} = k!(1-x)^k,$$

donc :

$$I_n = n!\sum_{k=0}^{n}\binom{n}{k}(x-1)^k = n!(1+x-1)^n = n!\,x^n.$$

Montrons (iii). On a, pour tous $k,\ k' \in \mathbb{N}$, $k \neq k'$, $\displaystyle\int_0^{+\infty} \varphi_k(t)\varphi_{k'}(t)\,\mathrm{d}t = 0$, donc :

$$\int_0^{+\infty}\Big(\sum_{k=0}^{n}\frac{x^k}{k!}\varphi_k(t)\Big)^2\,\mathrm{d}t = \sum_{k=0}^{n}\frac{x^{2k}}{(k!)^2}\int_0^{+\infty}\varphi_k^2(t)\,\mathrm{d}t = \sum_{k=0}^{n}\frac{x^{2k}}{(k!)^2}\|L_k\|_w^2 = \sum_{k=0}^{n}x^{2k}.$$

Par ailleurs, en utilisant (ii) :

$$\sum_{k=0}^{n}\int_0^{+\infty}\varphi(t,x)\frac{x^k}{k!}\varphi_k(t)\,\mathrm{d}t = \sum_{k=0}^{n}x^{2k}.$$

On en déduit, en utilisant (i) :

$$\int_0^{+\infty}\Big(\varphi(t,x)-\sum_{k=0}^{n}\frac{x^k}{k!}\varphi_k(t)\Big)^2\,\mathrm{d}t=\frac{1}{1-x^2}-2\sum_{k=0}^{n}x^{2k}+\sum_{k=0}^{n}x^{2k}=\frac{1}{1-x^2}-\sum_{k=0}^{n}x^{2k}$$

D'après l'assertion (iii) du lemme ci-dessus, pour tout $x\in\,]-1,1[$ fixé, la série de fonctions $t\mapsto\sum\limits_{n=0}^{+\infty}\frac{x^n}{n!}\varphi_n(t)$ converge vers $t\mapsto\varphi(t,x)$ au sens de la norme $\|\ \|$.

Nous allons en déduire que, pour tout $\gamma\in\mathbb{R}$, strictement positif, la fonction $t\mapsto \mathrm{e}^{-\gamma t}$ est un élément de l'ensemble F.

Si $\gamma>0$, il existe un $x\in\,]-1,1[$ unique tel que $\gamma=\frac{1}{2}\frac{1+x}{1-x}$.

Ainsi, toute fonction de la forme $t\mapsto \mathrm{e}^{-\gamma t}$ ($\gamma\in\mathbb{R}_+^*$) est un élément de $L_{\{1\}}(]0,+\infty[)$, qui est limite de combinaisons linéaires de fonctions de Laguerre, au sens de la norme $\|\ \|$. On a donc $\{\mathrm{e}^{-\gamma t}\}_{\gamma>0}\subset F$.

Nous pouvons maintenant déduire la proposition 60 du résultat analogue pour le poids trivial 1 sur un intervalle *borné* en utilisant un *changement de variable*.

Faisons le *changement de variable* $\tau(t):=\mathrm{e}^{-t}$. L'application τ réalise un difféomorphisme (de classe $\mathcal{C}^1$, et même mieux...) de $[0,+\infty[$ sur $]0,1]$. On a $\frac{\mathrm{d}\tau}{\mathrm{d}t}=-\tau$. Ainsi, si $f\in L_{\{1\}}(]0,+\infty[)$, et $g(\tau):=f(\ln\tau)$, on a :

$$\int_0^{+\infty}f(t)^2\,\mathrm{d}t=\int_0^1 g^2(\tau)\frac{\mathrm{d}\tau}{\tau}=\int_0^1\left(\frac{g(\tau)}{\sqrt{\tau}}\right)^2\mathrm{d}\tau.$$

La fonction $\tau\mapsto\frac{g(\tau)}{\sqrt{\tau}}$ est continue et de carré sommable (pour le poids trivial 1) sur l'intervalle ouvert $]0,1[$. D'après nos résultats dans le cas des intervalles *bornés* (*cf.* la proposition 14 de la page 549), cette fonction est limite de polynômes en τ au sens de la norme $\|\ \|'$, définie par le poids 1 sur $]0,1[$, soit $\big(\|h(\tau)\|'\big)^2:=\displaystyle\int_0^1 h^2(\tau)\,\mathrm{d}\tau$. On a, pour $p\in\mathbb{R}[\tau]$:

$$\int_0^1\Big(\frac{g(\tau)}{\sqrt{\tau}}-p(\tau)\Big)^2d\tau=\int_0^{+\infty}\big(f(t)\,\mathrm{e}^{t/2}-p(\mathrm{e}^{-t})\big)^2\mathrm{e}^{-t}\,\mathrm{d}t=\int_0^{+\infty}\big(f(t)-\mathrm{e}^{-t/2}p(\mathrm{e}^{-t})\big)^2\,\mathrm{d}t.$$

Ainsi f peut être approchée arbitrairement près, au sens de la norme $\|\ \|$ par des expressions de la forme $\mathrm{e}^{-t/2}(a_0+a_1\mathrm{e}^{-t}+a_2\mathrm{e}^{-2t}+\cdots+a_n\mathrm{e}^{-nt})$ $(a_0,\dots,a_n\in\mathbb{R})$; chaque fonction $t\mapsto\mathrm{e}^{-(k+\frac{1}{2})t}$ $(k\in\mathbb{N})$ étant dans F, f est dans l'adhérence de F. Comme F est, par définition, fermé, $f\in F$. Puisque la fonction f a été choisie arbitraire dans $L_{\{1\}}(]0,+\infty[)$, nous avons donc montré que $F=L_{\{1\}}(]0,+\infty[)$. On en déduit que $E=L_w([0,+\infty[)$, ce qui termine la démonstration.

Théorème 62. Soit $w : t \mapsto \mathrm{e}^{-t}$ le poids de Laguerre sur l'intervalle $[0, +\infty[$. Soit $\left(q_n := (-1)^n \frac{L_n}{n!}\right)_{n\in\mathbb{N}}$ la suite des polynômes de Laguerre *orthonormés* associée. Alors, pour toute fonction $f \in L_w([0, +\infty[)$ la suite de polynômes $\sum_{k=0}^{n} (\, f \mid q_k \,)_w \, q_k$ converge vers f au sens de la norme $\| \, \|_w$ et l'on a l'égalité de Parseval :

$$\|f\|_w^2 = \sum_{k=0}^{+\infty} (\, f \mid q_k \,)^2.$$

En utilisant une démarche voisine, on peut étendre ce résultat au cas des polynômes de Laguerre généralisés en utilisant la fonction génératrice de ceux-ci (*cf.* l'exercice III.4.20). On en déduit le résultat suivant, en utilisant la relation entre polynômes d'Hermite et polynômes de Laguerre généralisés de paramètre $\pm 1/2$ (*cf.* la proposition 55 de la page 600).

Théorème 63. Soit $w : t \mapsto \mathrm{e}^{-t^2}$ le poids d'Hermite sur $\mathbb{R}$. Soit $\left(q_n := \dfrac{H_n}{n!\sqrt{\pi}}\right)_{n\in\mathbb{N}}$ la suite des polynômes d'Hermite *orthonormés* associée. Alors, pour toute fonction $f \in L_w(\mathbb{R})$ la suite de polynômes $\sum_{k=0}^{n} (\, f \mid q_k \,)_w \, q_k$ converge vers f au sens de la norme $\| \, \|_w$; on a l'égalité de Parseval :

$$\|f\|_w^2 = \sum_{k=0}^{+\infty} (\, f \mid q_k \,)^2.$$

Terminons par un résultat général sur les développements en série de polynômes orthogonaux *classiques* (i.e. Jacobi, Laguerre généralisés, Hermite).

Théorème 64. Soit w un poids « classique » sur (α, β). Soit Q le polynôme de degré inférieur ou égal à 2 correspondant ($Q > 0$ sur $]\alpha, \beta[$). Soit f une fonction *de classe* $\mathcal{C}^1$ sur $]\alpha, \beta[$, telle que :

$$\int_\alpha^\beta f^2(t) w(t) \, \mathrm{d}t < +\infty \quad \text{et} \quad \int_\alpha^\beta f'^2(t) Q(t) w(t) \, \mathrm{d}t < +\infty.$$

Alors la fonction f admet le développement en polynômes orthonormés : $\sum_{k=0}^{+\infty} (\, f \mid q_k \,)_w \, q_k$ au sens de la norme $\| \, \|_w$, et ce développement *converge uniformément* sur tout segment $[\alpha_1, \beta_1] \subset]\alpha, \beta[$.

La démonstration fait l'objet de l'exercice III.4.24.

3.8 Tables des polynômes orthogonaux classiques

Les tables suivantes rassemblent un certain nombre de formules établies plus haut. Nous avons ajouté, sans preuve, les fonctions génératrices des polynômes de Tchebychev de première et seconde espèce.

3.8.1 Polynômes de Legendre P_n

Intervalle : $[-1,1]$.
Fonction poids : $w=1$. Masse totale : $\mu_0=2$.
Normalisation : $P_n(1):=1$.
Coefficient dominant : $\frac{1}{2^n}\binom{2n}{n}$.
Carré de la norme : $\frac{2}{2n+1}$.
Formules explicites[14] :

$$P_n(t)=\frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}^2(t-1)^{n-k}(t+1)^k=\frac{1}{2^n}\sum_{k=0}^{[n/2]}(-1)^k\binom{n}{k}\binom{2n-2k}{n}t^{n-2k}.$$

Formule de récurrence :

$$\forall n\geqslant 1,\quad (n+1)P_{n+1}(t)=(2n+1)tP_n(t)-nP_{n-1}(t),\quad P_0(t)=1\quad\text{et}\quad P_1(t)=t.$$

Équation différentielle :

$$(1-t^2)y''-2ty'+n(n+1)y=0.$$

Formule d'Olinde Rodriguez :

$$P_n(t)=\frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left((t^2-1)^n\right).$$

Fonction génératrice : $\frac{1}{\sqrt{1-2tz+z^2}}=\sum_{n\in\mathbb{N}}P_n(t)z^n$, $t\in[-1,1]$, $|z|<1$.

3.8.2 Polynômes de Tchebychev de première espèce T_n

Intervalle : $[-1,1]$.
Fonction poids : $w=\frac{1}{\sqrt{1-t^2}}$, Masse totale : $\mu_0=\pi$.
Normalisation : $T_n(1)=1$.
Coefficient dominant : 2^{n-1}.
Carré de la norme : $\begin{cases}\pi/2 & \text{si}\quad n\neq 0\\ \pi & \text{si}\quad n=0.\end{cases}$
Formules explicites ($n\geqslant 1$) :

$$T_n(t)=\frac{n}{2}\sum_{k=0}^{[n/2]}(-1)^k\frac{(n-k-1)!}{k!(n-2k)!}(2t)^{n-2k}=\cos(n\arccos t).$$

Formule de récurrence :

$$\forall n\geqslant 1,\quad T_{n+1}(t)=2tT_n(t)-T_{n-1}(t)\quad\text{et}\quad T_0(t)=1,\ T_1(t)=t.$$

Équation différentielle :

$$(1-t^2)y''-ty'+n^2y=0.$$

[14] On peut montrer l'expression de droite pour P_n en utilisant la formule du binôme et la formule d'Olinde Rodriguez.

Formule d'Olinde Rodriguez :

$$\begin{aligned} T_n(t) &= (-1)^n \sqrt{\pi}\, \frac{\sqrt{1-t^2}}{2^n \Gamma(n+1/2)} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \left[(1-t^2)^{n-1/2} \right] \\ &= (-1)^n \frac{2^n n!}{(2n)!} \sqrt{1-t^2}\, \frac{\mathrm{d}^n}{\mathrm{d}t^n} \left[(1-t^2)^{n-1/2}) \right]. \end{aligned}$$

Fonction génératrice : $\dfrac{1-tz}{1-2tz+z^2} = \sum_{n\in\mathbb{N}} T_n(t) z^n$, $-1 < t < 1$ et $|z| < 1$.

3.8.3 Polynômes de Tchebychev de deuxième espèce U_n

Intervalle : $]-1, 1[$. Masse totale $\pi/2$.
Fonction poids : $\sqrt{1-t^2}$.
Normalisation : $U_n(1) = n+1$.
Coefficient dominant : 2^n.
Carré de la norme : $\pi/2$.
Formules explicites ($n \geqslant 0$) :

$$U_n(t) = \frac{T'_{n+1}(t)}{n+1} = \sum_{k=0}^{[n/2]} (-1)^k \frac{(n-k)!}{k!(n-2k)!} (2t)^{n-2k} = \sum_{h=0}^{[n/2]} (-1)^h \binom{n-h}{h} (2t)^{n-2h}.$$

Formule de récurrence :

$$\forall n \geqslant 1, \quad U_{n+1}(t) = 2tU_n(t) - U_{n-1}(t) \quad \text{et} \quad U_0(t) = 1,\ U_1(t) = 2t.$$

Équation différentielle : $(1-t^2)y'' - 3ty' + n(n+2)y = 0$.
Formule d'Olinde Rodriguez :

$$\begin{aligned} U_n(t) &= (-1)^n \sqrt{\pi} \frac{n+1}{2^{n+1}\Gamma(n+3/2)\sqrt{1-t^2}} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \left[(1-t^2)^{n+1/2} \right] \\ &= (-2)^n \frac{(n+1)!}{(2n+1)!} \frac{1}{\sqrt{1-t^2}} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \left[(1-t^2)^{n+1/2} \right]. \end{aligned}$$

Fonction génératrice : $\frac{1}{1-2tz+z^2} = \sum_{n\in\mathbb{N}} U_n(t) z^n$, $-1 < t < 1$ et $|z| < 1$.

3.8.4 Polynômes de Jacobi $P_n^{a,b}$, $a > -1$ et $b > -1$

Intervalle : $]-1, 1[$. Masse totale $\mu_0 = 2^{a+b+1} \dfrac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$.

Fonction poids : $w = (1-t)^a (1+t)^b$.

Normalisation : $P_n^{(a,b)}(1) = \binom{n+a}{n}$.

Coefficient dominant : $\dfrac{\Gamma(2n+a+b+1)}{2^n n! \Gamma(n+a+b+1)}$.

Carré de la norme : $\dfrac{2^{a+b+1}\Gamma(a+n+1)\Gamma(n+b+1)}{n!(a+b+2n+1)\Gamma(a+b+n+1)}$.

Formules explicites :

$$P_n^{(a,b)}(t) = \frac{1}{2^n}\sum_{h=0}^{n}\binom{n+a}{h}\binom{n+b}{n-h}(t-1)^{n-h}(t+1)^h$$
$$= \frac{\Gamma(a+n+1)}{n!\Gamma(a+b+n+1)}\sum_{k=0}^{n}\binom{n}{k}\frac{\Gamma(a+b+n+k+1)}{\Gamma(a+k+1)}\left(\frac{t-1}{2}\right)^k.$$

Formule de récurrence (avec $s := a+b$) :

$$\forall n \geqslant 2, \quad 2n(n+s)(2n+s-2)P_n^{(a,b)}(t)$$
$$= (2n+s-1)\left[(2n+s)(2n+s-2)t + a^2 - b^2\right]P_{n-1}^{(a,b)}(t)$$
$$- 2(n+a-1)(n+b-1)(2n+s)P_{n-2}^{(a,b)}(t),$$

avec :

$$P_0^{(a,b)}(t) = 1 \qquad \text{et} \qquad P_1^{(a,b)}(t) = \frac{1}{2}(a+b+2)t + \frac{1}{2}(a-b)\cdot$$

Équation différentielle :

$$(1-t^2)y'' + [b-a-(a+b+2)t]y' + n(n+1+a+b)y = 0.$$

Formule d'Olinde Rodriguez :

$$P_n^{(a,b)}(t) = \frac{(-1)^n}{2^n n!}(1-t)^{-a}(1+t)^{-b}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left[(1-t)^{a+n}(1+t)^{b+n}\right].$$

3.8.5 Polynômes de Laguerre L_n

Intervalle : $[0,+\infty[$. Masse totale $\mu_0 = 1$.
Fonction poids : e^{-t}.
Normalisation : coefficient dominant $(-1)^n$.
Coefficient dominant : $(-1)^n$.
Carré de la norme : $(n!)^2$
Formules explicites :

$$L_n(t) = n!\sum_{h=0}^{n}\frac{(-1)^h}{h!}\binom{n}{h}t^h = (n!)^2\sum_{h=0}^{n}\frac{(-1)^h}{(h!)^2(n-h)!}t^h$$
$$= (-1)^n\Big(t^n - \frac{n^2}{1!}t^{n-1} + \frac{n^2(n-1)^2}{2!}t^{n-2} + \cdots + (-1)^n n!\Big).$$

Formule de récurrence :

$$\forall n \geqslant 1, \quad L_{n+1}(t) = \big(2n+1-t\big)L_n(t) - n^2 L_{n-1}(t) \quad \text{et} \quad L_0(t) = 1,\ L_1(t) = 1-t.$$

Équation différentielle :

$$ty'' + (1-t)y' + ny = 0.$$

Formule d'Olinde Rodriguez : $L_n(t) = \mathrm{e}^t\frac{\mathrm{d}^n}{\mathrm{d}t^n}[t^n\mathrm{e}^{-t}]$.

Fonction génératrice exponentielle : $(1-z)^{-1}\mathrm{e}^{\frac{tz}{z-1}} = \sum_{n\in\mathbb{N}} L_n(t)\frac{z^n}{n!}$, $|z| < 1$.

3.8.6 Polynômes de Laguerre $\tilde{L}_n$

Intervalle : $[0, +\infty[$. Masse totale $\mu_0 = 1$
Fonction poids : e^{-t}.
Normalisation : coefficient dominant $(-1)^n/n!$.
Coefficient dominant : $(-1)^n/n!$.
Carré de la norme : 1
Formule explicite : $\tilde{L}_n(t) = \sum_{k=0}^{n} (-1)^k \binom{n}{n-k} \frac{t^k}{k!}$.
Formule de récurrence : $\tilde{L}_0(t) = 1$, $\tilde{L}_1(t) = 1 - t$ et

$$\forall n \geqslant 1, \quad (n+1)\tilde{L}_{n+1}(t) = (2n+1-t)\tilde{L}_n(t) - n\tilde{L}_{n-1}(t).$$

Équation différentielle :

$$ty'' + (1-t)y' + ny = 0.$$

Formule d'Olinde Rodriguez : $\tilde{L}_n(t) = \frac{1}{n!}\mathrm{e}^t \frac{\mathrm{d}^n}{\mathrm{d}t^n}[t^n \mathrm{e}^{-t}]$
Fonction génératrice : $(1-z)^{-1}\mathrm{e}^{\frac{tz}{z-1}} = \sum_{n\in\mathbb{N}} \tilde{L}_n(t) z^n$, $|z| < 1$, $t \in \mathbb{R}$.

3.8.7 Polynômes de Laguerre généralisés $L_n^{(a)}$, $a > -1$

Intervalle : $[0, +\infty[$. Masse totale $\mu_0 = \Gamma(a+1)$.
Fonction poids : $t^a \mathrm{e}^{-t}$.
Normalisation : le coefficient du terme de plus haut degré est $(-1)^n$.
Coefficient dominant : $(-1)^n$.
Carré de la norme : $n!\Gamma(n+a+1)$.
Formule explicite :

$$L_n^{(a)}(t) = n! \sum_{k=0}^{n} \frac{(-1)^k}{k!} \binom{n+a}{n-k} t^k.$$

Formule de récurrence :

$$\forall n \geqslant 1, \quad L_{n+1}^{(a)}(t) = \big(2n+a+1-t\big)L_n^{(a)}(t) - n(n+a)L_{n-1}^{(a)}(t)$$

avec $L_0^{(a)}(t) = 1$ et $L_1^{(a)}(t) = 1 + a - t$.
Équation différentielle :

$$ty'' + (1+a-t)y' + ny = 0.$$

Formule d'Olinde Rodriguez :

$$L_n^{(a)}(t) = t^{-a}\mathrm{e}^t \frac{\mathrm{d}^n}{\mathrm{d}t^n}[t^{n+a}\mathrm{e}^{-t}].$$

Fonction génératrice exponentielle : $(1-z)^{-a-1}\mathrm{e}^{\frac{tz}{z-1}} = \sum_{k\in\mathbb{N}} L_n^{(a)}(t)\frac{z^n}{n!}$, $|z| < 1$.

3.8.8 Polynômes de Laguerre généralisés $\tilde{L}_n^{(a)}$, $a > -1$

Intervalle : $[0, +\infty[$. Masse totale $\mu_0 = \Gamma(a+1)$.
Fonction poids : $t^a \mathrm{e}^{-t}$.
Normalisation : le coefficient du terme de plus haut degré est $(-1)^n/n!$.
Coefficient dominant : $(-1)^n/n!$.
Carré de la norme : $\frac{\Gamma(n+a+1)}{n!}$.
Formule explicite :

$$\tilde{L}_n^{(a)}(t) = \sum_{h=0}^{n} \frac{(-1)^h}{h!} \binom{n+a}{n-h} t^h.$$

Formule de récurrence :

$$\forall n \geqslant 1, \quad (n+1)\tilde{L}_{n+1}^{(a)}(t) = (2n+a+1-t)\tilde{L}_n^{(a)}(t) - (n+a)\tilde{L}_{n-1}^{(a)}(t),$$

avec $\tilde{L}_0^{(a)}(t) = 1$ et $\tilde{L}_1^{(a)}(t) = 1 + a - t$.
Équation différentielle :

$$ty'' + (1+a-t)y' + ny = 0.$$

Formule d'Olinde Rodriguez :

$$\tilde{L}_n^{(a)}(t) = \frac{1}{n!} t^{-a} \mathrm{e}^t \frac{\mathrm{d}^n}{\mathrm{d}t^n}[t^{n+a}\mathrm{e}^{-t}].$$

Fonction génératrice : $(1-z)^{-a-1}\mathrm{e}^{\frac{tz}{z-1}} = \sum_{n\in\mathbb{N}} \tilde{L}_n^{(a)}(t) z^n$, $|z| < 1$.

3.8.9 Polynômes d'Hermite H_n

Intervalle : $]-\infty, +\infty[$. Masse totale $\mu_0 = \sqrt{\pi}$.
Fonction poids : e^{-t^2}.
Normalisation : le coefficient du terme de plus haut degré est 2^n.
Coefficient du terme dominant : 2^n.
Carré de la norme : $2^n\sqrt{\pi}n!$.
Formule explicite :

$$H_n(t) = n! \sum_{k=0}^{[n/2]} \frac{(-1)^k}{k!} \frac{(2t)^{n-2k}}{(n-2k)!}.$$

Formule de récurrence :

$$\forall n \geqslant 1, \quad H_{n+1}(t) = 2tH_n(t) - 2nH_{n-1}(t) \quad \text{et} \quad H_0(t) = 1,\ H_1(t) = 2t.$$

Équation différentielle :

$$y'' - 2ty' + 2ny = 0.$$

Formule d'Olinde Rodriguez : $H_n(t) = (-1)^n \mathrm{e}^{t^2} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \mathrm{e}^{-t^2}$.
Fonction génératrice exponentielle : $\mathrm{e}^{2tz-z^2} = \sum_{n\in\mathbb{N}} H_n(t) \frac{z^n}{n!}$.

4 Compléments

Du point de vue *purement algébrique* les polynômes $\mathscr{M}$-orthogonaux, les récurrences linéaires polynomiales de la forme (2) de la page 537, les approximants de Padé dits du type $[n, n]$ et certaines fractions continues « fonctionnelles » sont quatre « facettes » du même sujet. Nous l'avons vu partiellement avec les relations entre polynômes orthogonaux et récurrences. Dans cette partie, nous compléterons l'étude en introduisant les approximants de Padé et les fractions continues « fonctionnelles ».

Du point de vue *de l'analyse*, les problèmes sont beaucoup plus difficiles. Une présentation assez complète des résultats connus (en particulier des solutions du « problème des moments » de Stieltjes) sortirait du cadre de nos ouvrages et nous nous sommes limités à présenter des résultats importants mais restant élémentaires. En voici un exemple (dû à Tchebychev et son élève Markov et retrouvé par Stieltjes à la fin du XIXe siècle).

Soient w un poids sur un intervalle *borné* (α, β) et $(p_n)_{n\in\mathbb{N}}$ une suite de polynômes orthogonaux pour w. On note, pour tout $n \in \mathbb{N}$, $\{t_{kn}\}_{k\in[\![1,n]\!]}$ les zéros de p_n et $\{\lambda_{kn}\}_{k\in[\![1,n]\!]}$ les nombres de Christoffel correspondants. Alors la suite de fractions rationnelles $\sum\limits_{k=1}^{n} \frac{\lambda_{kn}}{z-t_{kn}}$ converge uniformément sur tout compact de $U := \mathbb{C} \setminus [\alpha, \beta]$ vers une fonction holomorphe sur U dont le développement en série (convergente) à l'infini est la série décalée des moments de w : $\mu_0/z + \mu_1/z^2 + \cdots + \mu_n/z^n + \cdots$.

Ce résultat permet (entre autres...) de « retrouver » un poids w à partir de la seule connaissance de ses moments.

4.1 Déterminants de Hankel, approximants de Padé

4.1.1 Déterminants de Hankel et moments

Nous donnons dans cette partie un certain nombre de formules reliant les polynômes orthogonaux associés à un poids w et les moments μ_n de ce poids, en utilisant des déterminants dits de Hankel. Ces déterminants joueront aussi un rôle important dans la partie suivante (4.1.2).

Une matrice $A = (a_{ij})$ à coefficients dans un corps K est dite *de Hankel* (du nom du mathématicien allemand Hermann Hankel, 1839-1873), si ses *anti-diagonales* (ou diagonales ascendantes) sont constantes, c'est-à-dire si a_{ij} ne dépend que de $i + j$. Le déterminant d'une matrice de Hankel s'appelle *déterminant de Hankel*.

Soit w une fonction poids sur l'intervalle (α, β), de moments $(\mu_n)_{n\in\mathbb{N}}$. On lui associe les *matrices de Hankel* et les *déterminants de Hankel* :

$$H_n := \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & \cdots & \mu_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_n & \mu_{n+1} & \cdots & \mu_{2n} \end{pmatrix} \quad \text{et} \quad D_n := \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & \cdots & \mu_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_n & \mu_{n+1} & \cdots & \mu_{2n} \end{vmatrix}.$$

On a, pour tous $h, k \in \mathbb{N}$, $\mu_{h+k} = (t^h \mid t^k)_w$. La matrice de Hankel H_n est donc la *matrice de Gram* de la famille libre $(1, t, \ldots, t^n)$ (*cf.* le paragraphe *Matrice de Gram* du module *Espaces préhilbertiens* du livre [L2]). Par suite, elle est symétrique et *définie positive*. On a, pour tout $n \in \mathbb{N}$, $D_n > 0$.

Exercice 14.

Soit w une fonction poids sur l'intervalle (α, β), de moments $(\mu_n)_{n\in\mathbb{N}}$. Montrer que pour tous $p, n \in \mathbb{N}$, les déterminants de Hankel :

$$D_{n,p} := \begin{vmatrix} \mu_p & \mu_{p+1} & \cdots & \mu_{p+n} \\ \mu_{p+1} & \mu_{p+2} & \cdots & \mu_{p+n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{p+n} & \mu_{p+n+1} & \cdots & \mu_{p+2n} \end{vmatrix}.$$

sont *strictement positifs*.

Solution. Quitte à permuter les coordonnées, le déterminant $D_{n,p}$ s'interprète comme mineur principal d'une matrice définie positive, il est donc strictement positif (*cf.* la proposition 69 de la page 169 du livre [L2]).

Problème des moments

Stieltjes a montré que si l'on se donne une suite $(\mu_n)_{n\in\mathbb{N}}$, c'est la suite des moments d'une *mesure réelle positive* μ si, et seulement si, pour tout $n \in \mathbb{N}$, les déterminants de Hankel D_n et $D_{n,1}$ sont positifs[15] (la mesure μ n'est pas nécessairement unique et n'est pas nécessairement définie par un poids).

Proposition 65. Soit (q_n) la suite de polynômes orthogonaux *normés* associée au poids w. On a, pour tout entier $n > 0$:

$$q_n(t) = \frac{1}{\sqrt{D_n D_{n-1}}} \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & \cdots & \mu_{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-1} \\ 1 & t & \cdots & t^n \end{vmatrix}.$$

Démonstration.

La preuve la plus simple consiste à vérifier que les seconds membres $s_n(t)$ de l'égalité à démontrer sont des polynômes de degrés respectifs n et qu'ils forment une famille orthonormée. Elle est détaillée dans l'exercice III.4.27. Nous allons donner une autre démonstration un peu plus longue mais permettant de comprendre l'origine des formules.

Soit $n \in \mathbb{N}^*$. Déterminons un polynôme unitaire $u_n = a_0 + a_1 t + \cdots + a_{n-1}t^{n-1} + t^n$, de degré n, de telle sorte que, pour tout $k \in [\![0, n-1]\!]$, $(\, u_n \mid t^k \,) = 0$. On doit avoir :

$$\forall k \in [\![0, n-1]\!], \quad a_0\mu_k + a_1\mu_{k+1} + \cdots + a_{n-1}\mu_{k+n-1} + \mu_{n+k} = 0,$$

c'est-à-dire :

$$\forall k \in [\![0, n-1]\!], \quad a_0\mu_k + a_1\mu_{k+1} + \cdots + a_{n-1}\mu_{k+n-1} = -\mu_{n+k},$$

qui est un système linéaire de n équations à n inconnues $a_0, \ldots, a_{n-1}$, de déterminant $D_{n-1} \neq 0$. On peut le résoudre par les *formules de Cramer* (*cf.* le module *Espaces*

[15] *Cf.* par exemple [Ak].

vectoriels de dimension finie du livre [L1]). On obtient ainsi, en notant V_h, $h \in [\![0, n-1]\!]$, les vecteurs colonnes de la matrice de Hankel H_{n-1} et W le vecteur colonne associé à $(\mu_n, \ldots, \mu_{2n-1})$:

$$\begin{aligned}
\forall h \in [\![1, n-1]\!], \quad a_h &= -\frac{1}{D_{n-1}} \det(V_0, V_1, \ldots, V_{h-1}, W, V_{h+1}, \ldots, V_{n-1}) \\
&= -\frac{(-1)^{n-h}}{D_{n-1}} \det(V_0, V_1, \ldots, V_{h-1}, V_{h+1}, \ldots, V_{n-1}, W) \\
&= -\frac{(-1)^{n-h}}{D_{n-1}} \Delta_h,
\end{aligned}$$

avec $\Delta_h := \det(V_0, V_1, \ldots, V_{h-1}, V_{h+1}, \ldots, V_{n-1}, W)$.

En développant le déterminant de Hankel D_n par rapport à sa dernière ligne, on obtient :

$$\begin{aligned}
D_n &= \sum_{h=0}^{n-1} (-1)^{n+h} \mu_{n+h} \det(V_0, V_1, \ldots, V_{h-1}, V_{h+1}, \ldots, V_{n-1}, W) + \mu_{2n} D_{n-1} \\
&= \sum_{h=0}^{n-1} (-1)^{n+h} \mu_{n+h} \Delta_h + \mu_{2n} D_{n-1}.
\end{aligned}$$

Par ailleurs :

$$\begin{aligned}
\|u_n\|^2 = (\, u_n \mid t^n \,) &= \sum_{h=0}^{n-1} a_h \mu_{n+h} + \mu_{2n} = \frac{1}{D_{n-1}} \left(\sum_{h=0}^{n-1} (-1)^{n+h} \mu_{n+h} \Delta_h + \mu_{2n} D_{n-1} \right) \\
&= \frac{D_n}{D_{n-1}}.
\end{aligned}$$

Donc :

$$q_n = \frac{u_n}{\sqrt{D_n / D_{n-1}}} = \frac{1}{\sqrt{D_n D_{n-1}}} \left(\sum_{h=0}^{n-1} (-1)^{n+h} \Delta_h t^h + D_{n-1} t^n \right).$$

La somme $\sum_{h=0}^{n-1} (-1)^{n+h} \Delta_h t^h + D_{n-1} t^n$ est le développement suivant la dernière ligne du déterminant :

$$\begin{vmatrix}
\mu_0 & \mu_1 & \ldots & \mu_n \\
\mu_1 & \mu_2 & \ldots & \mu_{n+1} \\
\vdots & \vdots & \vdots & \vdots \\
\mu_{n-1} & \mu_n & \ldots & \mu_{2n-1} \\
1 & t & \ldots & t^n
\end{vmatrix},$$

ce qui termine la preuve. ∎

Corollaire 66.

(i) Si les polynômes orthonormés (q_n) associés au poids w admettent $\xi_n t^n$ pour monôme de degré maximum, l'on a $\xi_n = \sqrt{\frac{D_{n-1}}{D_n}}$ et $\frac{D_n}{D_0} = \frac{1}{\xi_1^2 \dots \xi_n^2}$.

(ii) Soit (u_n) la suite de polynômes orthogonaux *unitaires* associée au poids w. On a, pour tout entier $n > 0$:

$$u_n(t) = \frac{1}{D_{n-1}} \begin{vmatrix} \mu_0 & \mu_1 & \dots & \mu_n \\ \mu_1 & \mu_2 & \dots & \mu_{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{n-1} & \mu_n & \dots & \mu_{2n-1} \\ 1 & t & \dots & t^n \end{vmatrix}.$$

Remarque. L'assertion (ii) du corollaire ci-dessus reste vraie pour une suite de polynômes unitaires orthogonaux pour une forme linéaire $\mathscr{M} : \mathbb{R}[X] \to \mathbb{R}$, en remplaçant les μ_n par les $c_{n+1} := \mathscr{M}(t^n)$ et les D_n, par les :

$$\Delta_n := \begin{vmatrix} c_1 & c_2 & \dots & c_{n+1} \\ c_2 & c_3 & \dots & c_{n+2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n+1} & c_{n+2} & \dots & c_{2n+1} \end{vmatrix},$$

si l'on *suppose* $\Delta_n \neq 0$, pour tout $n \in \mathbb{N}$. Le lecteur le vérifiera aisément en utilisant une variante de la preuve de la proposition 65.

Si la forme linéaire $\mathscr{M}$ est *strictement positive*, pour tout $n \in \mathbb{N}$, le déterminant Δ_n est le déterminant d'une matrice de Gram, il est donc *strictement positif*.

Proposition 67. (i) Les polynômes orthonormés (q_n) associés au poids w admettent la représentation intégrale :

$$q_n(t) = \frac{1}{n!\sqrt{D_n D_{n-1}}} \int_{(\alpha,\beta)^n} \prod_{i=1}^{n} (t - \lambda_i) \prod_{1 \leqslant h < k \leqslant n} (\lambda_h - \lambda_k)^2 w(\lambda_1) \cdots w(\lambda_n) \ \mathrm{d}\lambda_1 \dots \mathrm{d}\lambda_n.$$

(ii) Les déterminants de Hankel admettent la représentation intégrale :

$$D_{n-1} = \frac{1}{n!} \int_{(\alpha,\beta)^n} \prod_{1 \leqslant h < k \leqslant n} (\lambda_h - \lambda_k)^2 w(\lambda_1) \dots w(\lambda_n) \ \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_n$$

Ces formules sont dues au mathématicien allemand Heine. Leur démonstration fait l'objet de l'exercice III.4.26. Découvertes au XIXe siècle, et un peu oubliées par la suite, elles ont récemment retrouvé une utilité dans l'étude statistique des écarts des valeurs propres des grandes matrices aléatoires (symétriques ou hermitiennes), qui est un problème important en *mécanique statistique*.

4.1.2 Approximants de Padé

Les approximants de Padé (introduits par le mathématicien français Henri Padé à la fin du XIXe siècle en relation avec des travaux d'Hermite) sont très utilisés dans des contextes divers : théorie de la transcendance, fonctions analytiques, accélération de convergence, physique théorique, mécanique des fluides, théorie du contrôle, implémentation du calcul des fonctions élémentaires sur les calculettes...

Nous donnons ici quelques rudiments de la théorie des approximants de Padé. Ils seront suffisants pour pouvoir expliciter les relations étroites de cette théorie avec celle des polynômes orthogonaux et celle des fractions continues.

La notion élémentaire de *développement limité* (en 0, pour fixer les idées) est centrée sur l'approximation des *fonctions* par des *polynômes*, une fonction g (de la variable ξ) étant considérée comme « petite » si elle est $\mathrm{O}(\xi^{k+1})$ (et d'autant plus petite, que l'entier k est grand). Ainsi, une fonction f est approchée à l'ordre k par un polynôme P si $f(\xi) - P(\xi) = \mathrm{O}(\xi^{k+1})$.

Limitons-nous à des fonctions de classe $\mathcal{C}^\infty$ (qui seront souvent analytiques dans les applications, mais pas toujours) sur un voisinage de 0. Dans ce cas, la condition $g(\xi) = \mathrm{O}(\xi^{k+1})$ est équivalente à :

$$g(0) = g'(0) = \frac{g''(0)}{2!} = \cdots = \frac{g^{(k)}(0)}{k!} = 0.$$

Si l'on considère la *série de Taylor* de g, qui est, par définition, la série *formelle* :

$$\hat{g} = \sum_{h\in\mathbb{N}} \frac{g^{(h)}(0)}{h!}\xi^h \in \mathbb{R}[[\xi]]$$

(on notera que, vu nos hypothèses, cette série n'a *pas nécessairement un rayon de convergence non nul*[16]), la condition $g = \mathrm{O}(\xi^{k+1})$ équivaut à la nullité des $k+1$ premiers termes, c'est-à-dire à la condition $\omega(\hat{g}) \geqslant k+1$ (où $\omega(\hat{g})$ est *l'ordre* de la série $\hat{g}$).

Lemme 68. Soient g une fonction de classe $\mathcal{C}^\infty$ sur un intervalle ouvert contenant l'origine et $\hat{g} \in \mathbb{R}[[\xi]]$ son développement de Taylor à l'origine. Soit $k \in \mathbb{N}$, on a $g(\xi) = \mathrm{O}(\xi^{k+1})$ si, et seulement si, $\omega(\hat{g}) \geqslant k+1$.

Chez certains auteurs, la condition $\omega(\hat{g}) \geqslant k+1$, pour une série formelle $\hat{g}$, est écrite $\hat{g} = \mathrm{O}(\xi^{k+1})$.

Au lieu de n'utiliser que des polynômes P pour approcher une fonction, au sens rappelé ci-dessus, on peut aussi utiliser des *fractions rationnelles* P/Q ($P \in \mathbb{R}[\xi]$, $Q \in \mathbb{R}[\xi] \setminus \{0\}$) et dire qu'une fonction f est approchée à l'ordre k par P/Q si $f(\xi) - P(\xi)/Q(\xi) = \mathrm{O}(\xi^{k+1})$.

Les fonctions de classe $\mathcal{C}^\infty$ restant bornées au voisinage de l'origine, il est naturel de n'utiliser, pour l'approximation de ces fonctions, que des fractions rationnelles *sans pôle à l'origine*. Une telle fraction rationnelle s'écrit P/Q avec $Q(0) \neq 0$, et, quitte à multiplier P et Q par une même constante, non nulle, on peut supposer que $Q(0) = 1$. Si $\sum_{h\in\mathbb{N}} b_h\xi^h$ est le

[16] Dans ce chapitre, une série entière de rayon de convergence nul sera dite *divergente*.

développement en série de P/Q à l'origine, la condition $f(\xi) - \frac{P}{Q}(\xi) = \mathrm{O}(\xi^{k+1})$ est équivalente à la nullité des $k+1$ premiers termes de la série formelle $\hat{g} = \sum_{h=0}^{+\infty} \left(\frac{f^{(h)}(0)}{h!} - b_h\right)\xi^h$, c'est-à-dire à la condition $\omega(\hat{g}) \geqslant k+1$. On peut donc oublier les fonctions et utiliser une notion *purement algébrique* d'approximation des séries formelles par des fractions rationnelles : c'est le principe de la théorie des *approximants de Padé*. Cette théorie s'appliquera ensuite automatiquement aux fonctions, via leurs développements de Taylor (en utilisant le lemme 68).

Dans le cadre algébrique, on peut travailler avec un corps (commutatif) arbitraire, le développement de Taylor (ou, si l'on préfère, le développement en série) de P/Q à l'origine étant remplacé par la division de P par Q suivant les puissances croissantes (*cf.* le théorème 68 de la page 324 du module *Polynômes et fractions rationnelles* du livre [L1]). Compte tenu des applications que nous avons en vue, nous allons simplement présenter les choses dans le cadre *complexe* (le cas *réel* est similaire). La variable complexe sera notée ξ, comme ci-dessus, et pas z comme on pourrait s'y attendre, car, dans les applications étudiées plus loin, nous travaillerons souvent au voisinage de l'infini en une variable complexe z (respectivement réelle t), et donc au voisinage de 0 en $\xi = 1/z$ (respectivement $\xi = 1/t$).

Définition 10. Soient $m, n \in \mathbb{N}$. Soit $\hat{f}(\xi) = \sum_{h\in\mathbb{N}} c_h \xi^h \in \mathbb{C}[[\xi]]$ une série entière formelle à coefficients complexes.
Un *approximant de Padé* d'ordre $[m/n]$ de $\hat{f}$ est une *fraction rationnelle* P/Q, telle que $\deg P \leqslant m$, $\deg Q \leqslant n$, $Q(0) \neq 0$ et $\omega\left(\hat{f} - \frac{P}{Q}\right) \geqslant m+n+1$.

Dans $\omega\left(\hat{f} - \frac{P}{Q}\right) \geqslant m+n+1$, $\frac{P}{Q}$ est, par abus de notation, identifié à son développement en série. On impose en général $Q(0) = 1$, ce que nous ferons dans la suite sauf mention explicite du contraire. Dans les applications, nous considérerons surtout les approximants, dits *diagonaux*, obtenus pour $m = n$. On a une définition similaire dans le cas réel ; dans ce cas nous appellerons, par définition, approximants de Padé d'une fonction f (de classe $\mathcal{C}^\infty$) ceux de son développement de Taylor $\hat{f}$ à l'origine.

Puisque $Q(0) \neq 0$, la condition $\omega\left(\hat{f} - \frac{P}{Q}\right) \geqslant m+n+1$ est équivalente à la condition

$$\omega(Q\hat{f} - P) \geqslant m+n+1. \tag{62}$$

Cette dernière est bien plus simple à utiliser, car, la série $\hat{f}$ étant donnée, les inconnues du problème (i.e. les coefficients de P et Q) y figurent *linéairement*.

Proposition 69. Soient $m, n \in \mathbb{N}$. Soit $\hat{f}(\xi)$ une série entière formelle à coefficients complexes. Si elle admet un approximant de Padé d'ordre $[m/n]$, alors celui-ci est *unique*.

Démonstration. Si P_1/Q_1 et P_2/Q_2 sont deux approximants de Padé d'ordre $[m/n]$

de $\hat{f}(\xi)$, on a $\omega\left(\frac{P_1}{Q_1} - \frac{P_2}{Q_2}\right) \geqslant m+n+1$, donc $\omega(P_1Q_2 - P_2Q_1) \geqslant m+n+1$. Comme $\deg(P_1Q_2 - P_2Q_1) \leqslant m+n$, cela implique $P_1Q_2 - P_2Q_1 = 0$, donc $\frac{P_1}{Q_1} = \frac{P_2}{Q_2}$. ■

Exemples. (i) Soit $f(\xi) := \exp \xi$. Voici quelques approximants de Padé :

$$[1/0] : \ 1+\xi, \qquad [0/1] : \ \frac{1}{1-\xi}, \qquad [1/1] : \ \frac{1+\frac{\xi}{2}}{1-\frac{\xi}{2}},$$

$$[1/2] : \ \frac{1+\frac{\xi}{3}}{1-\frac{2\xi}{3}+\frac{\xi^2}{6}}, \qquad [2/2] : \ \frac{1+\frac{\xi}{2}+\frac{\xi^2}{12}}{1-\frac{\xi}{2}+\frac{\xi^2}{12}}.$$

Les vérifications, à partir de $\hat{f} = 1+\xi+\xi^2/2+\cdots+\xi^h/h!+\cdots$, à l'aide de l'équation (62), sont faciles et laissées au lecteur.
(ii) Soit $f(\xi) = \sin \xi$. Il existe un approximant de Padé d'ordre $[2/2]$: $\frac{\xi}{1+\frac{\xi^2}{6}}$.

Le développement de Taylor de $\sin \xi$ à l'origine est la série formelle $\hat{f} = \xi - \frac{\xi^3}{6} + \frac{\xi^5}{120} \cdots$ L'équation (62) pour l'approximant d'ordre $[2/2]$ s'écrit :

$$\omega\Big((1+b_1\xi+b_2\xi^2)\big(\xi - \frac{\xi^3}{6} + \frac{\xi^5}{120}\big) - (a_0+a_1\xi+a_2\xi^2)\Big) \geqslant 5.$$

Le lecteur terminera le calcul, que nous détaillerons plus loin sous forme matricielle.

(iii) Soit $f(\xi) = \sqrt{\frac{1+\frac{\xi}{2}}{1+2\xi}}$. Son développement de Taylor à l'origine est la série formelle $\hat{f}(\xi) = 1 - \frac{3}{4}\xi + \frac{39}{32}\xi^2 + \cdots$. L'équation (62), pour un approximant d'ordre $[1/1]$, s'écrit ici $\omega\Big((1+b_1\xi)(1-\frac{3}{4}\xi+\frac{39}{32}\xi^2) - (a_0+a_1\xi)\Big) \geqslant 3$.

Ainsi $\frac{1+\frac{7}{8}\xi}{1+\frac{13}{8}\xi}$ est l'approximant d'ordre $[1/1]$.

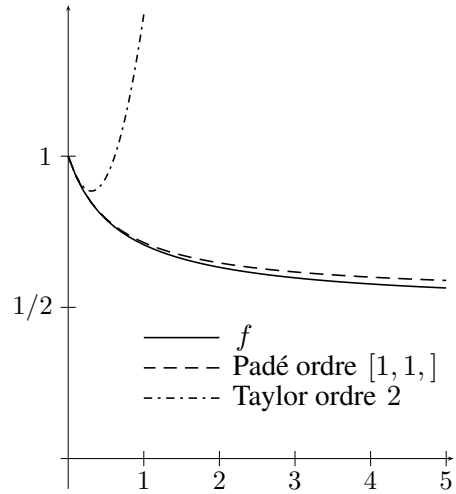
Il est intéressant de comparer les *graphes* (pour ξ réel) des trois fonctions :

$$\xi \mapsto 1 - \frac{3}{4}\xi + \frac{39}{32}\xi^2$$

$$\xi \mapsto \frac{1+\frac{7}{8}\xi}{1+\frac{13}{8}\xi}$$

$$\xi \mapsto f(\xi) = \sqrt{\frac{1+\frac{\xi}{2}}{1+2\xi}}$$

(*cf.* ci-contre) et de constater que l'approximation de Padé est « bien meilleure » que l'approximation de Taylor (les deux derniers graphes ont en particulier une asymptote horizontale, respectivement aux ordonnées $1/2 = 0.5$ et $7/13 = 0.53846\ldots$, tandis que le premier est une parabole).



(iv) Soit $f(\xi) = 1 + \xi^2 + \xi^4$. On cherche un approximant d'ordre $[1/1]$:

$$\omega\Big((1 + b_1\xi)(1 + \xi^2 + \xi^4) - (a_0 + a_1\xi)\Big) \geqslant 3.$$

Cela s'écrit $1 - a_0 + (b_1 - a_1)\xi + \xi^2 = 0$, ce qui est impossible. Un tel approximant *n'existe donc pas*.

(v) On considère le poids $w = 1$ sur l'intervalle $[-1, 1]$ (cas de Legendre) et, aux moments $(\mu_h)_{h\in\mathbb{N}}$ correspondants, on associe la série entière :

$$\sum_{h\in\mathbb{N}} \mu_h \xi^{h+1} = 2\big[\xi + \frac{\xi^3}{3} + \frac{\xi^5}{5} + \cdots\big]$$

(qui est évidemment convergente). On calcule les trois premiers approximants diagonaux : $[0/0]$: 1, $[1/1]$: $2\frac{\xi}{1}$ et $[2/2]$: $2\frac{3\xi-1}{3-\xi^2}\cdot$ Posant $\xi = 1/t$, l'on obtient $[0/0](t) = \frac{1}{1}$ $[1/1](t)$: $\frac{2}{t}$ et $[2/2](t)$: $2\frac{3-t}{3t^2-1}$: $\frac{3-t}{\frac{1}{2}(3t^2-1)}$, et l'on constate que les *dénominateurs* de ces fractions sont *les trois premiers polynômes de Legendre*. Le lecteur vérifiera, à titre d'exercice, que le dénominateur de l'approximant $[3/3]$, considéré comme fonction de t, est le quatrième polynôme de Legendre. En fait, pour tout $n \in \mathbb{N}$, le dénominateur de 'approximant $[n/n](t)$ est le polynôme de Legendre $P_n(t)$. Nous verrons pourquoi plus loin.

(vi) On considère le poids e^{-t} sur $[0, +\infty[$ (cas de Laguerre).

Aux moments $(\mu_h = h!)_{h\in\mathbb{N}}$ correspondants, on associe la série entière formelle (divergente) :

$$\sum_{h\in\mathbb{N}} \mu_h \xi^{h+1} = \sum_{h\in\mathbb{N}} h!\xi^{h+1} = \xi + \xi^2 + 2!\xi^3 + \cdots$$

On calcule les trois premiers approximants diagonaux : $[0/0]$: 1, $[1/1]$ et $[2/2]$. Posant $\xi = 1/t$, on obtient :

$$[0/0](t) : \frac{1}{1} \qquad [1/1](t) : \frac{2}{t} \qquad \text{et} \qquad [2/2](t) : 2\frac{3-t}{3t^2-1}$$

et l'on constate que les *dénominateurs* de ces fractions sont *les trois premiers polynômes de Laguerre*. On peut poursuivre comme dans l'exemple (v).

Remarques.

(i) Quand un approximant de Padé existe, il est *unique*. Par contre, l'exemple (iv) ci-dessus montre *qu'il n'existe pas toujours*.

(ii) Dans les exemples (v) et (vi), on voit apparaître des polynômes orthogonaux classiques dans un contexte purement algébrique, à partir de suites simples et indépendamment de toute notion d'intégration. D'où une question naturelle : pourquoi la théorie des approximants de Padé appliquée à certaines séries entières formelles fournit-elle des suites de polynômes orthogonaux ? Nous y répondrons plus loin et nous mettrons en évidence des liens étroits avec la transformation de Stieltjes et la théorie des fractions continues, renouant ainsi avec la principale source historique de la théorie des polynômes orthogonaux.

Existence et calcul des approximants de Padé

Soit $\hat{f} := \sum_{k\in\mathbb{N}} c_k\xi^k$ une série formelle.
Si $P(\xi) = a_0 + \cdots + a_m\xi^m$ et $Q(\xi) = b_0 + \cdots + b_n\xi^n$, la condition (62), de la page 622, s'écrit :

$$\omega\big((b_0 + \cdots b_n\xi^n)\hat{f} - (a_0 + \cdots a_m\xi^m)\big) \geqslant m + n + 1. \tag{63}$$

On a nécessairement $Q(0) \neq 0$, nous supposerons donc $Q(0) = b_0 := 1$. On écrit les $m + n + 1$ relations correspondant à l'annulation des coefficients des monômes de degré inférieur ou égal à $m + n$ du premier menbre. On regroupe ces relations en deux paquets : l'un formé par les relations d'annulation des coefficients des monômes de degrés $m + 1, \ldots, m + n$ d'une part, l'autre par les relations d'annulation des coefficients des monômes de degré inférieur ou égal à m d'autre part. Notons que les a_i n'apparaissent pas dans le premier paquet. On obtient ainsi deux systèmes. Le premier est :

$$\begin{cases} c_{m-n+1}b_n + c_{m-n+2}b_{n-1} + \cdots + c_m b_1 + c_{m+1} = 0 \\ c_{m-n+2}b_n + c_{m-n+3}b_{n-1} + \cdots + c_{m+1}b_1 + c_{m+2} = 0 \\ \vdots \\ c_m b_n + c_{m+1}b_{n-1} + \cdots + c_{m+n-1}b_1 + c_{m+n} = 0 \end{cases}$$

et le second :

$$\begin{cases} a_0 & = c_0 \\ a_1 & = c_1 + b_1 c_0 \\ \vdots & \\ a_m & = c_m + \sum_{k=1}^{\inf(m,n)} b_k c_{m-k}. \end{cases}$$

Le premier système s'écrit sous forme matricielle :

$$\begin{pmatrix} c_{m-n+1} & c_{m-n+2} & \cdots & c_m \\ c_{m-n+2} & c_{m-n+3} & \cdots & c_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ c_m & c_{m+1} & \cdots & c_{m+n-1} \end{pmatrix} \begin{pmatrix} b_n \\ b_{n-1} \\ \vdots \\ b_1 \end{pmatrix} = \begin{pmatrix} -c_{m+1} \\ -c_{m+2} \\ \vdots \\ -c_{m+n} \end{pmatrix},$$

en posant (si nécessaire) $c_h := 0$, si $h < 0$. Si l'on a une solution $(b_0, \ldots, b_n)$ de ce système, on obtient $(a_0, \ldots, a_m)$ en reportant les b_k $(k = 0, \ldots, n)$ dans le second système, et l'on en déduit l'approximant de Padé d'ordre $[m/n]$ de $\hat{f}$. D'où le résultat suivant (avec $c_h := 0$, pour $h < 0$).

Théorème 70. Soit $\hat{f} := \sum_{k\in\mathbb{N}} c_k\xi^k$ une série entière formelle à coefficients complexes. Soient $m, n \in \mathbb{N}$. Si le *déterminant de Hankel* :

$$\begin{vmatrix} c_{m-n+1} & c_{m-n+2} & \cdots & c_m \\ c_{m-n+2} & c_{m-n+3} & \cdots & c_{m+1} \\ \vdots & \vdots & \vdots & \vdots \\ c_m & c_{m+1} & \cdots & c_{m+n-1} \end{vmatrix}$$

est *non nul*, la série $\hat{f}$ admet un approximant de Padé d'ordre $[m/n]$. En particulier si le déterminant de Hankel :

$$\Delta_{n-1} := \begin{vmatrix} c_1 & c_2 & \cdots & c_n \\ c_2 & c_3 & \cdots & c_{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ c_n & c_{n+1} & \cdots & c_{2n-1} \end{vmatrix}$$

est non nul, la série $\hat{f}$ admet un approximant de Padé d'ordre $[n/n]$.

Exercice 15.
Calculer, par la méthode matricielle ci-dessus, l'approximant d'ordre $[2, 2]$ de $\sin\xi$.
Solution. Le premier système s'écrit :

$$\begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix}\begin{pmatrix} b_2 \\ b_1 \end{pmatrix} = \begin{pmatrix} -c_3 \\ -c_4 \end{pmatrix} \quad \text{soit} \quad \begin{pmatrix} 1 & 0 \\ 0 & -\frac{1}{6} \end{pmatrix}\begin{pmatrix} b_2 \\ b_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{6} \\ 0 \end{pmatrix}.$$

D'où :

$$\begin{pmatrix} b_2 \\ b_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -6 \end{pmatrix}\begin{pmatrix} \frac{1}{6} \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{6} \\ 0 \end{pmatrix}.$$

Le deuxième système s'écrit alors :

$$\begin{pmatrix} c_0 & 0 & 0 \\ c_1 & c_0 & 0 \\ c_2 & c_1 & c_0 \end{pmatrix}\begin{pmatrix} 1 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \quad \text{soit} \quad \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ \frac{1}{6} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

L'approximant d'ordre $[2/2]$ cherché est donc $\frac{\xi}{1+\frac{\xi^2}{6}}$.

Soit $P := a_0 + a_1\xi + \cdots + a_n\xi^n \in \mathbb{R}_n[\xi]$ un polynôme *de degré au plus* n. On lui associe le polynôme $p(t) := t^n P(1/t) = a_0t^n + a_1t^{n-1} + \cdots + a_n \in \mathbb{R}[T]$. On dit que p est le *polynôme réciproque* de P (cette notion dépend du choix de n). Si $P(0) = 1$, p est *unitaire*.

Proposition 71. Soit $\hat{f} := \sum_{k\in\mathbb{N}} c_k\xi^k$ une série entière formelle à coefficients *réels*. On suppose que le déterminant de Hankel Δ_n est non nul pour tout $n \in \mathbb{N}$. On définit une forme linéaire $\mathscr{M} : \mathbb{R}[T] \to \mathbb{R}$ par $\mathscr{M}(t^n) := c_{n+1}$, pour tout $n \in \mathbb{N}$. Alors la suite $(q_n)_{n\in\mathbb{N}}$ des *polynômes réciproques des dénominateurs* Q_n, $n \in \mathbb{N}$, des approximants de Padé d'ordre $[n/n]$ de $\hat{f}$ est une suite de polynômes $\mathscr{M}$-orthogonaux.

Démonstration. Soit $Q_n := 1 + b_1\xi + \cdots + b_n\xi^n$. On reprend, avec $m = n$, le premier système de la preuve du théorème 70 :

$$\begin{cases} c_1b_n + c_2b_{n-1} + \cdots + c_nb_1 + c_{n+1} = 0 \\ c_2b_n + c_3b_{n-1} + \cdots + c_{n+1}b_1 + c_{n+2} = 0 \\ \vdots \\ c_nb_n + c_{n+1}b_{n-1} + \cdots + c_{2n-1}b_1 + c_{2n} = 0. \end{cases}$$

On note $q_n(t) := t^n + b_1t^{n-1} + \cdots + b_n$ le polynôme réciproque de Q_n. En remplaçant dans le système ci-dessus les c_i ($i \in \mathbb{N}^*$) par $\mathscr{M}(t^{i-1})$, on obtient :

$$\begin{cases} \mathscr{M}(1)b_n + \mathscr{M}(t)b_{n-1} + \cdots + \mathscr{M}(t^{n-1})b_1 + \mathscr{M}(t^n) = 0 \\ \mathscr{M}(t)b_n + \mathscr{M}(t^2)b_{n-1} + \cdots + \mathscr{M}(t^n)b_1 + \mathscr{M}(t^{n+1}) = 0 \\ \vdots \\ \mathscr{M}(t^{n-1})b_n + \mathscr{M}(t^n)b_{n-1} + \cdots + \mathscr{M}(t^{2n-2})b_1 + \mathscr{M}(t^{2n-1}) = 0, \end{cases}$$

que l'on peut réécrire

$$\begin{cases} \mathscr{M}(q_n) = \mathscr{M}(b_n + tb_{n-1} + \cdots + t^{n-1}b_1 + t^n) = 0 \\ \mathscr{M}(tq_n) = \mathscr{M}(tb_n + t^2b_{n-1} + \cdots + t^nb_1 + t^{n+1}) = 0 \\ \vdots \\ \mathscr{M}(t^{n-1}q_n) = \mathscr{M}(t^{n-1}b_n + t^nb_{n-1} + \cdots + t^{2n-2}b_1 + t^{2n-1}) = 0. \end{cases}$$

Par suite q_n est $\mathscr{M}$-orthogonal à $\mathbb{R}_{n-1}[T]$. ∎

Soient P, $Q \in \mathbb{R}_n[\xi]$ des polynômes et p, $q \in \mathbb{R}[T]$ leurs polynômes réciproques. Si $\xi t = 1$, $P(\xi)/Q(\xi) = p(t)/q(t)$.

Si les P_n/Q_n sont les approximants de Padé d'ordre $[n/n]$ de la série $\sum_{k\in\mathbb{N}} c_k\xi^k$ et si p_n, q_n sont les polynômes réciproques de P_n et Q_n, on dira que les p_n/q_n sont les approximants de Padé d'ordre $[n/n]$ de la série $\sum_{k\in\mathbb{N}} c_kt^{-k}$.

Approximants de Padé, récurrences linéaires d'ordre deux et polynômes $\mathscr{M}$-orthogonaux

Nous allons donner quelques compléments sur les approximants de Padé d'ordre $[n/n]$ en mettant en évidence les relations entre ces approximants, certaines relations de récurrence d'ordre deux et les polynômes $\mathscr{M}$-orthogonaux. Cela répond, en particulier, à la question posée dans la remarque (ii) de la page 624.

Le théorème suivant décrit le passage d'une récurrence d'ordre deux de la forme (2) aux approximants de Padé d'ordre $[n/n]$ d'une série entière formelle convenable. Pour tous les polynômes la variable est t.

Théorème 72. Soient $(B'_n)_{n\in\mathbb{N}^*}$ et $(C'_n)_{n\in\mathbb{N}^*}$ deux suites de nombres réels. On suppose $C'_1 \neq 0$. On définit par récurrence deux suites de polynômes $(p_n)_{n\in\mathbb{N}}$ et $(q_n)_{n\in\mathbb{N}}$:

1. $p_n = (t + B'_n)p_{n-1} - C'_n p_{n-2}$ si $n \geqslant 2$, et $p_0 = 0$, $p_1 = C'_1$;
2. $q_n = (t + B'_n)q_{n-1} - C'_n q_{n-2}$ si $n \geqslant 2$, et $q_0 = 1$, $q_1 = t + B'_1$.

(i) Les $\{p_n\}_{n\in\mathbb{N}^*}$ sont de degré $n-1$, les $\{q_n\}_{n\in\mathbb{N}^*}$ sont *unitaires* et de degré n et il existe une forme linéaire $\mathscr{M}$ sur $\mathbb{R}[T]$ telle que les polynômes $\{q_n\}_{n\in\mathbb{N}}$ soient $\mathscr{M}$-orthogonaux). On a alors, pour tous $n \in \mathbb{N}$, $\mathscr{M}(q_n^2) = C'_1 \cdots C'_{n+1}$ et $\mathscr{M}(t^{n+1}q_n) = -C'_1 \cdots C'_{n+1}(B'_1 + \cdots + B'_{n+1})$.

(ii) Il existe une série entière formelle unique $\hat{f} := c_1/t + \cdots + c_n/t^n + \cdots$ telle que pour tout $n \in \mathbb{N}$, p_n/q_n soit l'approximant de Padé d'ordre $[n/n]$ de $\hat{f}$.

(iii) Pour tout $n \in \mathbb{N}$, $\frac{q_n(t)-q_n(\tau)}{t-\tau}$ est un polynôme à deux variables t et τ et l'on a :

$$p_n(t) = \mathscr{M}^{\{\tau\}}\left(\frac{q_n(t) - q_n(\tau)}{t-\tau}\right). \tag{64}$$

(iv) Si $\mathscr{M}$ est la forme linéaire définie en (i), on a, pour tout $n \in \mathbb{N}$, $\mathscr{M}(t^n) = c_{n+1}$.

(v) La forme linéaire $\mathscr{M}$ définie en (i) est *strictement positive* si, et seulement si, $C'_n > 0$ pour tout $n \in \mathbb{N}$. S'il en est ainsi, pour tout $n \in \mathbb{N}$, alors les zéros de q_n sont *simples et réels*.

Dans (iii) la notation $\mathscr{M}^{\{\tau\}}$ signifie que l'on applique la forme linéaire $\mathscr{M}$ au polynôme en τ : $\dfrac{q_n(t) - q_n(\tau)}{t-\tau}$, à t fixé (*cf.* la remarque de la page 541).

Démonstration. On déduit (i) de l'exercice 3 de la page 542.

Montrons (ii). D'après l'exercice 3 de la page 542, on a, pour tout $n \in \mathbb{N}$:

$$p_{n+1}q_n - p_n q_{n+1} = C'_1 \cdots C'_{n+1} \qquad \text{donc} \qquad \frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n} = \frac{C'_1 \cdots C'_{n+1}}{q_n q_{n+1}}.$$

Le développement en série à l'infini (c'est-à-dire en $1/t$) de la fraction rationnelle $\frac{C'_1 \cdots C'_{n+1}}{q_n q_{n+1}}$ commence par le terme $\frac{C'_1 \cdots C'_{n+1}}{t^{2n+1}}$; par suite les développements en série à l'infini des fractions rationnelles $\frac{p_n}{q_n}$ et $\frac{p_{n+1}}{q_{n+1}}$ coïncident jusqu'au terme en $\frac{1}{t^{2n}}$ compris.

On en déduit l'existence d'une série formelle unique $\hat{f} := c_1/t + \cdots + c_n/t^n + \cdots$ telle que, pour tout $n \in \mathbb{N}$, $\omega\big(\frac{p_n}{q_n} - \frac{c_1}{t} - \frac{c_2}{t^2} - \ldots\big) \geqslant 2n+1$.

Pour tout $n \in \mathbb{N}$, p_n/q_n est l'approximant de Padé d'ordre $[n/n]$ de $\hat{f}$.

L'assertion (iii) résulte de la remarque de la page 541.

Montrons (iv). On a :

$$\frac{1}{t-\tau} = \frac{1}{t}\frac{1}{1-\frac{\tau}{t}} = \frac{1}{t} + \frac{\tau}{t^2} + \cdots + \frac{\tau^{2n-1}}{t^{2n}} + \frac{\tau^{2n}}{t^{2n}(t-\tau)},$$

donc :

$$\frac{q_n(t)-q_n(\tau)}{(t-\tau)q_n(t)} = \left(\frac{1}{t} + \frac{\tau}{t^2} + \cdots + \frac{\tau^{2n-1}}{t^{2n}}\right)\left(1 - \frac{q_n(\tau)}{q_n(t)}\right) + \frac{\tau^{2n}}{t^{2n}q_n(t)}\frac{q_n(t)-q_n(\tau)}{t-\tau}.$$

Le polynôme en t, à coefficients dans $\mathbb{C}[\tau]$: $\frac{q_n(t)-q_n(\tau)}{t-\tau}$ est de degré $n-1$, par suite le développement en série formelle en $1/t$ de $\mathscr{M}^{(\tau)}\left(\frac{\tau^{2n}}{t^{2n}q_n(t)}\frac{q_n(t)-q_n(\tau)}{t-\tau}\right)$ commence par un terme en $1/t^{2n+1}$. Par ailleurs, puisque $\mathscr{M}^{(\tau)}\big(\tau^k q_n(\tau)\big) = 0$ pour tout $k < n$, on a :

$$\mathscr{M}^{(\tau)}\left(\left(\frac{1}{t} + \frac{\tau}{t^2} + \cdots + \frac{\tau^{2n-1}}{t^{2n}}\right)\frac{q_n(\tau)}{q_n(t)}\right) = \mathscr{M}^{(\tau)}\left(\left(\frac{\tau^n}{t^{n+1}} + \frac{\tau}{t^2} + \cdots + \frac{\tau^{2n-1}}{t^{2n}}\right)\frac{q_n(\tau)}{q_n(t)}\right),$$

donc le développement en série formelle en $1/t$ de $\mathscr{M}^{(\tau)}\left(\left(\frac{1}{t} + \frac{\tau}{t^2} + \cdots + \frac{\tau^{2n-1}}{t^{2n}}\right)\frac{q_n(\tau)}{q_n(t)}\right)$ commence également par un terme en $1/t^{2n+1}$.

Par suite le développement de $\mathscr{M}^{(\tau)}\left(\frac{q_n(t)-q_n(\tau)}{(t-\tau)q_n(t)}\right)$ jusqu'à l'ordre $2n$ est :

$$\frac{\mathscr{M}^{(\tau)}(1)}{t} + \frac{\mathscr{M}^{(\tau)}(\tau)}{t^2} + \cdots + \frac{\mathscr{M}^{(\tau)}(\tau^{2n-1})}{t^{2n}} = \mathscr{M}^{(\tau)}\left(\frac{1}{t} + \frac{\tau}{t^2} + \cdots + \frac{\tau^{2n-1}}{t^{2n}}\right).$$

D'après (iii) on a $\frac{p_n(t)}{q_n(t)} = \mathscr{M}^{\{\tau\}}\left(\frac{q_n(t)-q_n(\tau)}{(t-\tau)q_n(t)}\right)$ et, d'après (ii), le développement de $\frac{p_n(t)}{q_n(t)}$ jusqu'à l'ordre $2n$ est $\frac{c_1}{t} + \cdots + \frac{c_{2n}}{t^{2n}}$.

On en déduit que, pour tout $k \leqslant 2n-1$, $\mathscr{M}^{(\tau)}(\tau^k) = c_{k+1}$, ou, de façon équivalente, $\mathscr{M}(t^k) = c_{k+1}$.

L'assertion (v) résulte de l'exercice 1 de la page 536 et du théorème 9 de la page 545. ■

Si, dans la proposition ci-dessus, les suites données $(B'_n)_{n\in\mathbb{N}^*}$ et $(C'_n)_{n\in\mathbb{N}^*}$ sont celles associées à un poids w comme dans (2) (avec la convention $C'_1 := \mu_0$), alors la suite $(q_n)_{n\in\mathbb{N}}$ est la suite $(u_n)_{n\in\mathbb{N}}$ des polynômes orthogonaux unitaires et la suite $(p_n)_{n\in\mathbb{N}}$ est la suite associée $(r_n)_{n\in\mathbb{N}}$ des polynômes de deuxième espèce. On prendra garde au fait que les suites $(B'_n)_{n\in\mathbb{N}^*}$ et $(C'_n)_{n\in\mathbb{N}^*}$ ainsi associées à un poids *ne sont pas arbitraires* (en particulier les C'_n sont nécessairement strictement positifs).

Corollaire 73. On se place dans les conditions de la proposition ci-dessus et l'on suppose que les zéros de q_n (dans $\mathbb{C}$) sont *simples*. On note $\{z_{k,n}\}_{k\in[\![1,n]\!]}$ ces zéros. On a (en étendant la forme linéaire $\mathscr{M}$ à $\mathbb{C}[Z]$ de la façon évidente) :

$$\forall\, p \in \mathbb{R}[Z], \quad \text{tel que} \quad \deg p \leqslant 2n-1, \quad \mathscr{M}(p) = \sum_{k=1}^{n} \frac{p(z_{k,n})}{q_n'(z_{k,n})} \mathscr{M}\left(\frac{q_n(z)}{z - z_{k,n}}\right) \tag{65}$$

et :

$$\frac{p_n(t)}{q_n(t)} = \sum_{k=1}^{n} \frac{1}{q_n'(z_{k,n})} \mathscr{M}\left(\frac{q_n(z)}{z - z_{k,n}}\right) \frac{1}{t - z_{k,n}}\cdot \tag{66}$$

Notant $\lambda_{k,n} := \mathscr{M}\left(\dfrac{q_n(z)}{q_n'(z_{k,n})(z - z_{k,n})}\right)$, on obtient une généralisation des nombres de Christoffel (*cf.* la proposition-définition 16 de la page 550). La formule (65) s'écrit : $\mathscr{M}(p) = \sum\limits_{k=1}^{n} \lambda_{k,n} p(z_{k,n})$. Elle généralise la formule de quadrature de Gauß (*cf.* la proposition-définition 16 de la page 550).

La formule (66) s'écrit $\frac{p_n(t)}{q_n(t)} = \sum\limits_{k=1}^{n} \frac{\lambda_{k,n}}{t - z_{k,n}}\cdot$. Elle généralise la formule (9).

Démonstration. Nous avons montré la formule (65) dans le cas d'un poids. Pour le cas général on procède de façon similaire en utilisant la formule d'interpolation de Lagrange. Celle-ci donne :

$$p(t) = \sum_{k=1}^{n} \frac{p(t_{k,n})}{q_n'(t_{k,n})} \frac{q_n(t)}{t - t_{k,n}} + q_n(t)s(t),$$

où s est un polynôme de degré inférieur ou égal à $n-1$.

On obtient (65) en appliquant $\mathscr{M}$ à cette égalité en utilisant $\mathscr{M}(q_n s) = 0$.

On montre la formule (66) en appliquant (65) au polynôme (en z à t fixé) :

$$p(z) := \frac{q_n(z) - q_n(t)}{(z-t)q_n(t)}.$$

■

Exercice 16.

Montrer que si la forme $\mathscr{M}$ est *strictement positive*, les nombres de Christoffel généralisés définis ci-dessus sont *strictement positifs*.

Solution. Si $\mathscr{M}$ est *strictement positive*, les zéros de q_n sont distincts et réels (*cf.* le théorème 9 de la page 545). On les note $\{t_{k,n}\}_{k\in[\![1,n]\!]}$. Pour tout $h \in [\![1,n]\!]$, on définit le polynôme $s_h := (t - t_{1,n})^2 \dots (t - t_{h-1,n})^2 (t - t_{h+1,n})^2 \dots (t - t_{n,n})^2$. Les s_h sont de degré $2n-2$, positifs et non nuls. On peut donc leur appliquer la formule (65). On obtient $\mathscr{M}(s_h) = \lambda_{h,n} s_h(t_{h,n})$. Puisque $\mathscr{M}(s_h) > 0$ et $s_h(t_{h,n}) > 0$, on a $\lambda_{h,n} > 0$.

Considérons maintenant un problème « inverse » de celui « résolu » par le théorème 72 de la page 628. On part d'une série formelle *sans terme constant* $\hat{f} := c_1/t + \dots + c_n/t^n + \cdots$ vérifiant les hypothèses du théorème 70 (c'est-à-dire la non nullité des déterminants de Hankel

pour le cas $[n/n]$) et l'on considère ses approximants de Padé d'ordre $[n/n]$ notés p_n/q_n (les q_n étant supposés *unitaires*). On se propose de montrer que les suites $(p_n)_{n\in\mathbb{N}}$ et $(q_n)_{n\in\mathbb{N}}$ vérifient des récurrences d'ordre deux comme dans le théorème 72.

Théorème 74. Soit $\hat{f} := c_1/t + \cdots + c_n/t^n + \cdots$ une série formelle *sans terme constant* telle que, pour tout $n \in \mathbb{N}$ le déterminant de Hankel Δ_n du théorème 70 soit non nul. On note, pour tout $n \in \mathbb{N}$, p_n/q_n l'approximant de Padé d'ordre $[n/n]$ de $\hat{f}$, avec q_n unitaire.
On définit une forme linéaire $\mathscr{M}$ sur $\mathbb{R}[T]$ par $\mathscr{M}(t^n) = c_{n+1}$ pour tout $n \in \mathbb{N}$. On a alors les résultats suivants.

(i) Les $\{q_n\}_{n\in\mathbb{N}}$ sont $\mathscr{M}$-orthogonaux.

(ii) Pour tout $n \in \mathbb{N}^*$, q_n est donné par la formule :

$$q_n(t) = \frac{1}{\Delta_{n-1}} \begin{vmatrix} c_1 & c_2 & \dots & c_{n+1} \\ c_2 & c_3 & \dots & c_{n+2} \\ \vdots & \vdots & \vdots & \vdots \\ c_n & c_{n+1} & \dots & c_{2n} \\ 1 & t & \dots & t^n \end{vmatrix}$$

et l'on a, pour tout $n \in \mathbb{N}^*$, $\mathscr{M}(q_n^2) = \frac{\Delta_n}{\Delta_{n-1}} \neq 0$.

(iii) On définit, par récurrence, deux suites réelles $(C'_n)_{n\in\mathbb{N}^*}$ et $(B'_n)_{n\in\mathbb{N}^*}$ par les conditions :

$$\mathscr{M}(q_n^2) = C'_1 \cdots C'_{n+1} \quad \text{et} \quad \mathscr{M}(t^{n+1}q_n) = -C'_1 \cdots C'_{n+1}(B'_1 + \cdots + B'_{n+1}),$$

alors les suites $(p_n)_{n\in\mathbb{N}}$ et $(q_n)_{n\in\mathbb{N}}$ vérifient les relations de récurrence :

1. $p_n = (t + B'_n)p_{n-1} - C'_n p_{n-2}$ si $n \geqslant 2$, et $p_0 = 0$, $p_1 = C'_1$;
2. $q_n = (t + B'_n)q_{n-1} - C'_n q_{n-2}$ si $n \geqslant 2$, et $q_0 = 1$, $q_1 = t + B'_1$.

(iii) Les conditions suivantes sont équivalentes.

(a) Pour tout $n \in \mathbb{N}$, $\Delta_n > 0$.

(b) La forme linéaire $\mathscr{M}$ est *strictement positive*.

(c) Pour tout $n \in \mathbb{N}$, $C'_n > 0$.

Démonstration. L'assertion (i) résulte de la proposition 71. On prouve l'assertion (ii) par la méthode utilisée pour montrer la proposition 65 de la page 618 (*cf.* la remarque de la page 620).
Montrons (iii). On définit les suites $(C'_n)_{n\in\mathbb{N}^*}$ et $(B'_n)_{n\in\mathbb{N}^*}$ comme indiqué.
On a $q_0 = 1$ et $q_1 = t + B'_1$. Montrons que les $\{q_n\}$ vérifient, pour tout $n \geqslant 2$, les relations de récurrence :

$$q_n = (t + B'_n)q_{n-1} - C'_n q_{n-2}.$$

Pour tout $n \in \mathbb{N}^*$, on peut écrire de manière unique :

$$q_{n+1} = (t + b'_{n+1})q_n - c'_{n+1}q_{n-1} + d_{n-2}q_{n-2} + \cdots + d_0 q_0,$$

où $b'_{n+1},\ c'_{n+1},\ d_{n-2},\dots,d_0 \in \mathbb{R}$.
On a, pour tout $k \in [\![0, n-2]\!]$, $0 = \mathscr{M}(q_k q_{n+1}) = d_k\,\mathscr{M}(q_k^2)$, donc $d_k = 0$. Par suite :

$$q_{n+1} = (t + b'_{n+1})q_n - c'_{n+1}q_{n-1}.$$

Donc, d'après l'exercice 3 de la page 542, on a, pour tout $n \in \mathbb{N}$ $\mathscr{M}(q_n^2) = C'_1 c'_2 \dots c'_{n+1}$ et $\mathscr{M}(t^{n+1}q_n) = -C'_1 c'_2 \dots c'_{n+1}(B'_1 + b'_2 + \dots + b'_{n+1})$. On en déduit, par récurrence, $c'_n = C'_n$ et $b'_n = B'_n$, pour tout $n \in \mathbb{N}^*$. Par suite les $\{q_n\}$ vérifient, pour tout $n \geqslant 2$, les relations de récurrence :

$$q_n = (t + B'_n)q_{n-1} - C'_n q_{n-2}.$$

En appliquant le théorème 72 de la page 628, on en déduit qu'il existe une forme linéaire $\tilde{\mathscr{M}}$ telle que les $\{q_n\}_{n\in\mathbb{N}}$ soient $\tilde{\mathscr{M}}$-orthogonaux et que $\tilde{\mathscr{M}}(1) = C'_1$. Les $\{q_n\}_{n\in\mathbb{N}}$ formant une base de $\mathbb{R}[T]$, on a nécessairement $\tilde{\mathscr{M}} = \mathscr{M}$, donc, en particulier, $\tilde{\mathscr{M}}(t^n) = c_{n+1}$, pour tout $n \in \mathbb{N}$.
On définit une suite de polynômes $\{\tilde{p}_n\}_{n\in\mathbb{N}}$ par :

$$\tilde{p}_n = (t + B'_n)\tilde{p}_{n-1} - C'_n\tilde{p}_{n-2} \quad \text{si} \quad n \geqslant 2, \quad \text{et} \quad \tilde{p}_0 = 0,\ \tilde{p}_1 = C'_1.$$

D'après le théorème 72,, les $\tilde{p}_n/q_n$ sont les approximants de Padé d'ordre $[n, n]$ de la série $\hat{f}$. D'après l'unicité des approximants de Padé, on a, pour tout $n \in \mathbb{N}$, $p_n = \tilde{p}_n$.
La preuve de (iii) est facile et laissée au lecteur. ■

Exercice 17.

On considère la suite $(p_n)_{n\in\mathbb{N}}$ de polynômes définie par la relation de récurrence

$$p_{n+1} = (2n+1)\,tp_n + p_{n-1} \quad \text{et} \quad p_0 = 1, \quad p_1 = t + 1,$$

Ces polynômes sont les polynômes de Bessel (*cf.* les pages 556, 568 et 580 et l'exercice III.4.9 de la page 652).

1. Calculer le coefficient χ_n du terme dominant de p_n. Déterminer la récurrence vérifiée par les polynômes unitaires $u_n := \frac{1}{\chi_n}p_n$.
2. Soit $c_1 \in \mathbb{R}^*$. Il existe une unique forme linéaire $\mathscr{M} : \mathbb{R}[T] \to \mathbb{R}$ telle que les polynômes $\{p_n\}_{n\in\mathbb{N}}$ soient orthogonaux par rapport à $\mathscr{M}$ et que $\mathscr{M}(p_0) := c_1$. Calculer, pour tout $n \in \mathbb{N}^*$, $\mathscr{M}(p_n^2)$.
3. Montrer qu'il n'existe pas de poids w tel que les polynômes de Bessel soient orthogonaux par rapport à w.

Solution.

1. On a $\chi_{n+1} = (2n+1)\chi_n$, donc $\chi_n = (2n-1)(2n-3)\dots 3 = \frac{(2n-1)!}{2^{n-1}(n-1)!}$.
   Par suite :
   $(2n+1)(2n-1)\dots 3\,u_{n+1} = (2n+1)(2n-1)\dots 3\,tu_n + (2n-3)(2n-5)\dots 3\,u_{n-1}$
   c'est-à-dire :
   $$u_{n+1} = tu_n + \frac{1}{(2n+1)(2n-1)}\,u_{n-1}.$$
   Ainsi $u_{n+1} = tu_n - C'_{n+1}u_{n-1}$, avec $C'_{n+1} := -\frac{1}{(2n+1)(2n-1)}$.

2. D'après le théorème 72, on a $\mathscr{M}(u_n^2) = C'_{n+1} \cdots C'_2 c_1 = (-1)^n \frac{c_1}{(2n+1)\chi_n^2}$ et, par suite, $\mathscr{M}(p_n^2) = \chi_n^2 \mathscr{M}(u_n^2) = (-1)^n \frac{c_1}{2n+1}$.
3. S'il existait un poids w tel que les polynômes de Bessel soient orthogonaux par rapport à w, la forme linéaire associée $\mathscr{M}_w$ serait proportionnelle à $\mathscr{M}$ et les $\mathscr{M}(p_n^2)$ seraient tous de même signe, ce qui n'est pas.

Série génératrice des moments

Toutes les mesures considérées dans ce paragraphe sont, comme précédemment, des mesures définies par un poids w sur un intervalle (α, β) (dont tous les moments sont définis) ou des mesures (réelles) à support fini sur $\mathbb{R}$. À une telle mesure μ, de moments $(\mu_n)_{n\in\mathbb{N}}$, on peut associer la série génératrice des moments correspondante : $\sum\limits_{n=0}^{+\infty} \mu_n \xi^n$. Pour des raisons qui apparaîtront plus loin, on utilise plutôt la série des moments *décalée* $\sum\limits_{n=0}^{+\infty} \mu_n \xi^{n+1}$, et l'on pose $\xi = 1/z$. On écrira donc souvent la série génératrice décalée « au voisinage de l'infini » $\sum\limits_{n=0}^{+\infty} \mu_n z^{-n-1}$.

Proposition 75. Si l'intervalle (α, β) est borné, ou si μ est une mesure à support fini, la série génératrice des moments est une série entière de rayon de convergence (en ξ) non nul. Dans le premier cas le rayon de convergence est supérieur ou égal à $1/A$, avec $A := \max(|\alpha|, |\beta|)$.

Démonstration. Commençons par le cas d'une mesure à support fini. Il suffit, par linéarité, de montrer le résultat lorsque μ est une masse de Dirac. Soit $\mu := \delta_c$ la masse de Dirac en $c \in \mathbb{R}$. On a, pour tout $n \in \mathbb{N}$, $\mu_n = \delta_c(t^n) = c^n$. La série génératrice correspondante est $\sum\limits_{n=0}^{+\infty} c^n \xi^n$; son rayon de convergence est $1/|c| > 0$.

Passons au cas d'une mesure définie par un poids w sur l'intervalle borné $]\alpha, \beta]$. On a, pour tout $n \in \mathbb{N}$, $\mu_n = \displaystyle\int_\alpha^\beta t^n w(t)\,\mathrm{d}t$, donc $|\mu_n| \leqslant \displaystyle\int_\alpha^\beta |t|^n w(t)\,\mathrm{d}t$.

On pose $A := \max\big(|\alpha|, |\beta|\big)$. On a alors : $|\mu_n| \leqslant A^n \displaystyle\int_\alpha^\beta w(t)\,\mathrm{d}t \leqslant A^n |\mu_0|$, ce qui prouve la convergence de la série génératrice correspondante (son rayon de convergence est au moins égal à $1/A$). ∎

Le résultat précédent n'est pas en général vrai pour un poids sur un intervalle *non borné*. Considérons par exemple le poids définissant les polynômes de Laguerre : $w(t) = e^{-t}$ pour $t \in [0, +\infty[$. On a, pour tout $n \in \mathbb{N}$, $\mu_n = n!$ et la série génératrice des moments correspondante est $\sum\limits_{n=0}^{+\infty} n!\xi^n$, qui est clairement divergente (*i.e.* de rayon de convergence nul).

Si w est un poids sur un intervalle (α, β), les déterminants de Hankel associés D_n sont non nuls ($n \in \mathbb{N}$) (*cf.* la page 617), on peut donc appliquer le théorème 70 de la page 626 à la série génératrice des moments *décalée* $\hat{f}(\xi) = \mu_o\xi + \mu_1\xi^2 + \cdots + \mu_n\xi^{n+1} + \cdots$ ($\mu_n := c_{n+1}$). D'où le résultat suivant.

Proposition 76.
Soient w un poids sur un intervalle (α, β) et $\hat{f}(\xi) = \mu_o\xi + \mu_1\xi^2 + \cdots + \mu_n\xi^{n+1} + \cdots$ la série génératrice des moments décalée.
Pour tout $n \in \mathbb{N}$, cette série admet un approximant de Padé d'ordre $[n/n]$.

En utilisant le théorème 72, on montre facilement le résultat suivant qui explicite la relation entre, d'une part, les polynômes orthogonaux pour un poids w et les polynômes de seconde espèce associés, et, d'autre part, les approximants de Padé d'ordre $[n/n]$ de la série décalée des moments de w d'autre part.

Corollaire 77. Soient w un poids et $(\mu_n)_{n\in\mathbb{N}}$ ses moments. On note $(r_n)_{n\in\mathbb{N}}$ la suite des polynômes de seconde espèce associée à la suite $(u_n)_{n\in\mathbb{N}}$ des polynômes *unitaires* orthogonaux pour w. Alors, pour tout $n \in \mathbb{N}$, r_n/u_n est l'approximant de Padé d'ordre $[n/n]$ de la série décalée des moments $\mu_0 t^{-1} + \mu_1 t^{-2} + \cdots + \mu_{n-1} t^{-n} + \cdots$

Les polynômes orthogonaux pour w apparaissent ainsi comme les *dénominateurs des approximants de Padé d'ordre* $[n/n]$ de la série décalée des moments de w. C'est essentiellement la première définition « historique » des polynômes orthogonaux.

Nous avons vu (*cf.* (9)) que $\frac{r_n(t)}{u_n(t)} = \sum\limits_{k=0}^{n} \frac{\lambda_{k,n}}{t-t_{k,n}}$, où les $t_{k,n}$ sont les zéros de u_n et les $\lambda_{k,n}$ les nombres de Christoffel. En passant en variable complexe, on obtient $\frac{r_n(z)}{u_n(z)} = \sum\limits_{k=0}^{n} \frac{\lambda_{k,n}}{z-t_{k,n}}$.

Le résultat suivant résume les liens entre série décalés des moments, polynômes orthogonaux et approximants de Padé.

Corollaire 78. Soient w un poids et $(\mu_n)_{n\in\mathbb{N}}$ la suite de ses moments. Alors, pour tout $n \in \mathbb{N}$, la fraction rationnelle $\sum\limits_{k=0}^{n} \frac{\lambda_{k,n}}{t-t_{k,n}}$ est l'approximant de Padé d'ordre $[n/n]$ de la série décalée des moments $\mu_0 t^{-1} + \mu_1 t^{-2} + \cdots + \mu_{n-1} t^{-n} + \cdots$.

Ainsi la théorie de Padé nous dit que les fractions rationnelles $\sum\limits_{k=0}^{n} \frac{\lambda_{k,n}}{z-t_{k,n}}$ sont, en un sens algébrique, une « bonne approximation » de la série formelle :

$$\mu_0 t^{-1} + \mu_1 t^{-2} + \cdots + \mu_{n-1} t^{-n} + \cdots$$

Il est naturel de se demander si, quand la série décalée des moments converge, les *fonctions* définies par ces fractions sont, *au sens de l'analyse*, une « bonne approximation » de la somme de cette série (au voisinage de l'infini dans le plan complexe). La réponse à cette question (dans le cas où (α, β) est borné) est basée sur la transformation dite de Stieltjes. Elle est l'objet de la prochaine partie (*cf.* le théorème 81 de la page 639).

4.1.3 Transformation de Stieltjes

Soient w un poids sur (α, β) et $(p_n)_{n\in\mathbb{N}}$ une suite de polynômes orthogonaux pour w. On note, pour tout $n \in \mathbb{N}$, $\{t_{k,n}\}_{k\in[\![1,n]\!]}$ les zéros de p_n et $\{\lambda_{k,n}\}_{k\in[\![1,n]\!]}$ les nombres de Christoffel correspondants. Nous considérons ici, pour simplifier, seulement le cas où (α, β) est *borné*.

Comme nous l'avons dit plus haut, on souhaite étudier la limite de la suite de fractions rationnelles $\sum\limits_{k=1}^{n} \frac{\lambda_{k,n}}{z-t_{k,n}}$. On part des deux observations suivantes.

1. On rappelle la notation $\mu^{(n)} := \sum\limits_{k=1}^{n} \lambda_{kn}\delta_{t_{kn}}$. Pour tout $f \in \mathcal{C}\big([\alpha, \beta]), \mathbb{C}\big)$, la suite de réels $\big(\mu^{(n)}(f)\big)_{n\in\mathbb{N}^*} := \sum\limits_{k=1}^{n} f(t_k)\lambda_{k,n}$ converge vers $\mu(f) := \displaystyle\int_\alpha^\beta f(t)w(t)\,\mathrm{d}t$.
2. Pour tout $n \in \mathbb{N}$, et tout $z \in \mathbb{C} \setminus [\alpha, \beta]$, on a $\mu^{(n)}\left(\frac{1}{z-t}\right) = \sum\limits_{k=1}^{n} \frac{\lambda_{k,n}}{z-t_{k,n}}$.

Si f est à valeurs réelles, l'observation 1. est vraie d'après le théorème 20 de la page 554, et l'on passe au cas complexe en considérant $\operatorname{Re} f$ et $\operatorname{Im} f$. L'observation 2. est vraie car les zéros $\{t_{k,n}\}$ des polynômes orthogonaux appartiennent à $]\alpha, \beta[$, ce qui entraîne la continuité sur $[\alpha, \beta]$ de la fonction $t \mapsto \sum\limits_{k=1}^{n} \frac{\lambda_{kn}}{z-t_{kn}}$, pour tout $z \in \mathbb{C} \setminus [\alpha, \beta]$.

Ces observations suggèrent d'introduire une fonction S de la variable complexe z associée à w, $S : z \in \mathbb{C} \setminus [\alpha, \beta] \mapsto S(z) := \displaystyle\int_\alpha^\beta \frac{w(t)}{z-t}\,\mathrm{d}t$ et de montrer que c'est la limite cherchée.

Cela motive l'introduction de la notion de *transformée de Stieltjes* d'une mesure réelle positive. Nous nous limiterons au cas des mesures à support fini ou définie par un poids.

Si μ est une mesure à support fini ou une mesure définie par un poids dont tous les moments sont définis, on peut, pour tout nombre complexe z, non réel, définir $\mu\left(\frac{1}{z-t}\right)$ en utilisant le fait que, à z fixé, la fonction $t \mapsto \frac{1}{z-t}$ est continue (à valeurs complexes) sur $[\alpha, \beta]$. On procède comme suit.

1. Soit μ la masse de Dirac en $c \in \mathbb{R}$, on a $\mu\left(\frac{1}{z-t}\right) = \frac{1}{z-c}$ et l'on passe au cas d'une mesure à support fini par linéarité.
2. Soient μ la mesure définie par le poids w sur $]\alpha, \beta[$ et $\alpha < \alpha' < \beta' < \beta$.

   L'intégrale $\displaystyle\int_{\alpha'}^{\beta'} \frac{w(t)}{z-t}\,\mathrm{d}t$ existe puisque la fonction $t \mapsto \frac{w(t)}{z-t}$ (à valeurs complexes) est, à z fixé, non réel, continue en t sur $[\alpha', \beta']$. Notons $d > 0$ la distance de z à l'axe réel. On a $\displaystyle\int_{\alpha'}^{\beta'} \left|\frac{w(t)}{z-t}\right| \mathrm{d}t = \int_{\alpha'}^{\beta'} \frac{w(t)}{|z-t|}\,\mathrm{d}t \leqslant \frac{1}{d}\mu_0$. L'intégrale $\displaystyle\int_\alpha^\beta \frac{w(t)}{z-t}\,\mathrm{d}t$ est donc absolument convergente. On définit alors $\mu\left(\frac{1}{z-t}\right) := \displaystyle\int_\alpha^\beta \frac{w(t)}{z-t}\,\mathrm{d}t$.

On remarque que $\mu\left(\frac{1}{z-t}\right)$ peut encore être définie de façon analogue si z est réel, pourvu que z n'appartienne pas au support de μ dans le cas d'une mesure à support fini (respectivement que z n'appartienne pas à l'adhérence de (α,β) $\big($notée $[\alpha,\beta]\big)$ dans le cas d'un poids).

Définition 11. Soit μ une mesure à support fini $\Sigma \subset \mathbb{R}$, ou définie par un poids sur un intervalle (α,β). La *transformée de Stieltjes* de μ est la fonction définie pour $z \in \mathbb{C}\setminus\Sigma$ (respectivement $z \in \mathbb{C}\setminus[\alpha,\beta]$) par $z \mapsto S(\mu)(z) = \mu\left(\frac{1}{z-t}\right)$.

La transformée de Stieltjes dans le cadre défini ci-dessus avait déjà été considérée par Tchebychev. Stieltjes a étendu la définition au cas où μ est une mesure positive sur $\mathbb{R}_+$ (apparemment sans connaitre le travail de son prédécesseur...).

Proposition 79. Soit μ une mesure réelle à support fini $\Sigma \subset \mathbb{R}$, ou définie par un poids sur un intervalle (α,β).
(i) La *transformée de Stieltjes* de μ est *holomorphe* sur $P^1(\mathbb{C})\setminus\Sigma = (\mathbb{C}\setminus\Sigma)\cup\{\infty\}$ (respectivement $\mathbb{C}\setminus[\alpha,\beta]$).
(ii) Si l'intervalle $]\alpha,\beta[$ est *borné*, cette transformée se prolonge en une fonction « holomorphe à l'infini » (i.e. une fonction holomorphe de $\xi = 1/z$ au voisinage de $\xi = 0$), dont le développement en série en $\xi = 0$ est *la série génératrice des moments décalée* de μ.
(iii) On a $S(\mu)(\bar{z}) = \overline{S(\mu)(z)}$.
(iv) Si μ est *positive*, l'image par la transformée de Stieltjes $S(\mu)$ du demi-plan supérieur $\{\operatorname{Im} z > 0\}$ est contenue dans le demi-plan-inférieur $\{\operatorname{Im} z < 0\}$.

Démonstration. (i) Dans le cas d'une mesure à support fini, la transformée de Stieltjes est une fonction fraction rationnelle, à pôles dans Σ ; elle est donc holomorphe sur $P^1(\mathbb{C})\setminus\Sigma$. Dans le cas d'un poids, supposons d'abord que w est un poids sur un *segment*. Considérons, à t fixé, $z \mapsto \frac{w(t)}{z-t}$ comme une fonction de deux variables réelles x et y ($z = x+iy$), définie sur $\mathbb{C}\setminus\{t\}$ à valeurs dans $\mathbb{R}^2$:

$$\frac{w(t)}{z-t} = w(t)\Big(\operatorname{Re}\frac{1}{z-t} + i\operatorname{Im}\frac{1}{z-t}\Big) =: P(t,x,y) + iQ(t,x,y).$$

Les fonctions $(t,x,y) \mapsto \frac{\partial}{\partial x}\frac{w(t)}{x+iy-t}$ et $(t,x,y) \mapsto \frac{\partial}{\partial y}\frac{w(t)}{x+iy-t}$ sont continues en (t,x,y) sur le domaine $[\alpha,\beta]\times\mathbb{C}\setminus[\alpha,\beta]$, on a donc :

$$\frac{\partial}{\partial x}\int_\alpha^\beta \frac{w(t)}{x+iy-t}\,\mathrm{d}t = \int_\alpha^\beta \frac{\partial}{\partial x}\frac{w(t)}{x+iy-t}\,\mathrm{d}t$$
$$\frac{\partial}{\partial y}\int_\alpha^\beta \frac{w(t)}{x+iy-t}\,\mathrm{d}t = \int_\alpha^\beta \frac{\partial}{\partial y}\frac{w(t)}{x+iy-t}\,\mathrm{d}t$$

et ces fonctions sont continues en (x,y) (*cf.* le théorème de continuité d'une fonction définie par une intégrale du module *Intégration* du livre [L2]). De plus, le couple (P,Q) vérifie

les équations de Cauchy-Riemann. On en déduit que les parties réelle et imaginaire de la fonction $(x,y) \mapsto \displaystyle\int_\alpha^\beta \frac{w(t)}{x+iy-t}\,\mathrm{d}t$ vérifient aussi les conditions de Cauchy-Riemann. Ainsi, la fonction $z \mapsto S(\mu)(z)$ est holomorphe sur $\mathbb{C} \setminus [\alpha,\beta]$.

Considérons maintenant le cas d'un intervalle ouvert $]\alpha,\beta[$. Soient $z_0 \in \mathbb{C}\setminus\mathbb{R}$ et $d>0$ la distance de z_0 à $\mathbb{R}$. On note D le disque fermé de centre z_0 et de rayon $d/2$.

Soient $\alpha',\beta' \in \mathbb{R}$, tels que $\alpha<\alpha'<\beta'<\beta$. On note $I_{\alpha',\beta'}(z) = \displaystyle\int_{\alpha'}^{\beta'} \frac{w(t)}{z-t}\,\mathrm{d}t$; c'est une fonction holomorphe de z sur $\mathbb{C}\setminus\mathbb{R}$, daprès l'étude du cas d'un segment. On a, pour tout $z\in D$, $|I_{\alpha',\beta'}(z)| \leqslant \frac{2}{d}\mu_0$. La famille de fonctions holomorphes $\{I_{\alpha',\beta'} | \alpha<\alpha'<\beta'<\beta\}$ est donc une *famille normale* sur D. On en déduit que $S(\mu)(z)$ est holomorphe sur l'intérieur de D ($S_\mu(z)$ est, pour tout $z\in D$, la limite de la suite $I_{\alpha+1/n,\beta-1/n}(z)$, quand n tend vers $+\infty$).

(ii) Soit $G(\mu)(z)$ la somme de la série génératrice des moments *décalée* de μ : $\sum\limits_{n=0}^{+\infty} \mu_n z^{-n-1}$.

Elle est définie et holomorphe pour z non nul et « assez grand » : $|z|>A$, avec $A := \sup(|\alpha|,|\beta|)$ dans le cas d'un poids et $A := \sup\{|z| \mid z\in\Sigma\}$ dans le cas d'une mesure à support fini. Montrons que $F(\mu)$ et $G(\mu)$ sont égales pour $|z|>\rho$. Nous ne détaillerons que le cas d'un poids (celui d'une mesure à support fini est trivial).

On a, pour tout $n\in\mathbb{N}$, $\frac{1}{z-t} = z^{-1}\frac{1}{1-\frac{t}{z}} = z^{-1}\left[1+\frac{t}{z}+\cdots+\frac{t^n}{z^n}+\frac{t^{n+1}}{z^{n+1}}\frac{1}{1-\frac{t}{z}}\right]$, donc :

$$S(\mu)(z) = \int_\alpha^\beta \frac{w(t)}{z-t}\,\mathrm{d}t = \mu_0 z^{-1} + \mu_1 z^{-2} + \cdots + \mu_n z^{-n-1} + z^{-n-1}\int_\alpha^\beta t^{n+1}\frac{w(t)}{z-t}\,\mathrm{d}t.$$

Si $z\in\mathbb{C}$ est fixé, tel que $|z|>A$, on a :

$$\Big|\int_\alpha^\beta t^{n+1}\frac{w(t)}{z-t}\,\mathrm{d}t\Big| \leqslant A^{n+1}\int_\alpha^\beta \frac{w(t)}{|z|-|t|}\,\mathrm{d}t \leqslant \mu_0\frac{A^{n+1}}{|z|-A}$$

$$\text{et}\quad \Big|z^{-n-1}\int_\alpha^\beta t^{n+1}\frac{w(t)}{z-t}\,\mathrm{d}t\Big| \leqslant \mu_0\Big(\frac{A}{|z|}\Big)^{n+1}\frac{1}{|z|-A}.$$

La quantité au second membre de l'inégalité ci-dessus tend vers 0 quand n tend vers $+\infty$; ainsi, pour $|z|>A$, la série génératrice des moments décalée $\mu_0 z^{-1}+\mu_1 z^{-2}+\cdots+\mu_n z^{-n-1}+\cdots$ converge et sa somme $G(\mu)(z)$ est égale à $S(\mu)(z)$. La fonction $\xi\mapsto \mu_0\xi+\mu_1\xi^2+\cdots+\mu_n\xi^n+\cdots$ est évidemment holomorphe ; elle prolonge la fonction $\xi\mapsto S(\mu)(1/\xi)$.

(iii) La relation $S(\mu)(\bar z) = \overline{S(\mu)(z)}$ est évidente (la mesure μ est réelle).

(iv) Supposons maintenant la mesure *p*ositive et $\operatorname{Re} z>0$.

Dans le cas d'un poids, on a, en posant $z=x+iy$ $(x,y\in\mathbb{R})$:

$$S(\mu)(z) = \int_\alpha^\beta \frac{w(t)\,\mathrm{d}t}{z-t} = \int_\alpha^\beta \frac{x-t-iy}{(x-t)^2+y^2}w(t)\,\mathrm{d}t,$$

donc $\operatorname{Im} S(\mu)(z) = -y \displaystyle\int_\alpha^\beta \frac{1}{(x-t)^2+y^2} w(t)\,\mathrm{d}t$ et $\operatorname{Im} S(\mu)(z) \leqslant 0$. Par ailleurs :

$$\int_\alpha^\beta \frac{1}{(x-t)^2+y^2} w(t)\,\mathrm{d}t \geqslant \int_\alpha^\beta \frac{1}{y^2} w(t)\,\mathrm{d}t = \mu_0/y^2 > 0,$$

donc $\operatorname{Im} S(\mu)(z) > 0$.
Dans le cas d'une masse de Dirac δ_c, on a :

$$S(\delta_c)(z) = \frac{1}{z-c} = \frac{x-c-iy}{(x-c)^2+y^2} \qquad \text{et} \qquad \operatorname{Im} S(\delta_c)(z) = -\frac{y}{(x-c)^2+y^2} < 0.$$

Le lecteur traitera le cas des mesures positives à support fini à titre d'exercice. ■

Exemple. Soient $w = 1$ sur le segment $[-1,1]$ (cas de Legendre) et μ la mesure associée. On a $S(\mu)(z) := \displaystyle\int_{-1}^1 \frac{\mathrm{d}t}{z-t}\cdot$ Calculons d'abord $S(\mu)(z)$ pour z réel ($z = x \notin [-1,1]$). On a alors :

$$S(\mu)(x) = \big[\ln|x-t|\big]_1^{-1} = \ln|x+1| - \ln|x-1| = \ln\frac{x+1}{x-1}\cdot$$

Passons au cas où z est complexe ($z \notin [-1,1]$).
On suppose d'abord que $z \in \mathbb{C} \setminus [-1,+\infty[$; on peut alors, pour tout $t \in [-1,1]$ choisir la détermination principale du logarithme pour $\ln(z-t)$ qui est égale à $\ln(t-z) + i\pi$ si $z \in]-\infty,-1[$.
On a alors $\frac{1}{z-t} = -\frac{\mathrm{d}}{\mathrm{d}t}\ln(z-t)$ et $S(\mu)(z) = \big[\ln(z-t)\big]_1^{-1} = \ln(z+1) - \ln(z-1)$. Nous avons vu dans l'exercice 6 de la page 853 du module *Fonctions analytiques* du livre [L2] que la fonction :

$$\begin{array}{rcl} \mathbb{C} \setminus [-1,+\infty[& \longrightarrow & \mathbb{C} \\ z & \longmapsto & \ln(z+1) - \ln(z-1) \end{array}$$

se prolonge en une fonction continue sur $\mathbb{C} \setminus [-1,1]$ et que cette fonction est holomorphe sur ce domaine.
Finalement $S(\mu)(z) = \ln\frac{z+1}{z-1}\cdot$ Quand $|z| > 1$, on a $S(\mu)(z) = \ln\frac{1+1/z}{1-1/z}$, que l'on peut développer en série de $1/z$ (c'est-à-dire au voisinage de l'infini). On trouve :
$S(\mu)(z) = 2\big(\frac{1}{z} + \frac{1}{3z^3} + \frac{1}{5z^5} + \cdots + \frac{1}{(2n+1)z^{2n+1}} + \cdots\big)$, qui est bien la série génératrice des moments décalée de la mesure de Legendre.

Proposition 80. Soient $[\alpha, \beta]$ un segment et w un poids sur $]\alpha, \beta[$. Soient μ la mesure correspondante de densité w et $(\mu^{(n)})_{n\in\mathbb{N}^*}$ la suite de mesures discrètes associée (voir le théorème 20 de la page 554). Alors la suite de fonctions holomorphes $\left(S\big(\mu^{(n)}\big)\right)_{n\in\mathbb{N}^*}$ converge *simplement* vers $S(\mu)$ sur $\mathbb{C} \setminus [\alpha, \beta]$. Pour tout $n \in \mathbb{N}^*$, la fonction $S\big(\mu^{(n)}\big)$ est une *fraction rationnelle* dont les *pôles* sont les *zéros* du $n^{\text{ème}}$ polynôme orthogonal p_n associé à w. Plus précisément, on a $S\big(\mu^{(n)}\big)(z) = \sum\limits_{k=1}^{n} \frac{\lambda_{k,n}}{z-t_{k,n}}$, où les $t_{k,n}$ sont les zéros du $n^{\text{ème}}$ polynôme orthogonal pour w et les $\lambda_{k,n}$ les nombres de Christoffel correspondants..

Démonstration. Si l'on fixe $z \in \mathbb{C} \setminus [\alpha, \beta]$, la fonction $t \mapsto \frac{1}{z-t}$ est continue sur $[\alpha, \beta]$, donc (*cf.* le théorème 20 de la page 554) la suite de nombre réels $\mu^{(n)}\left(\frac{1}{z-t}\right) = S(\mu_n)(z)$ converge vers $\mu\left(\frac{1}{z-t}\right) = S(\mu)(z)$. ∎

En utilisant la proposition 79 de la page 636, on obtient le résultat suivant, qui est le résultat essentiel de cette partie.

Théorème 81. Soient $[\alpha, \beta]$ un *segment* et w un poids sur $]\alpha, \beta[$.
On note $A := \max(|\alpha|, |\beta|)$. La série décalée des moments de w :

$$\hat{f} = \frac{\mu_0}{z} + \frac{\mu_1}{z^2} + \cdots + \frac{\mu_n}{z^{n+1}} + \cdots$$

est convergente sur un domaine de la forme $\{z \in \mathbb{C} \mid |z| > 1/A\}$ et la suite de ses approximants de Padé converge simplement sur $\mathbb{C}\setminus[\alpha, \beta]$ vers la transformée de Stieltjes $S(w\,\mathrm{d}t)(z) = \displaystyle\int_\alpha^\beta \frac{w(t)}{z-t}\,\mathrm{d}t$ de w, qui est une fonction holomorphe sur $\mathbb{C} \setminus [\alpha, \beta]$ dont la restriction à $\{z \in \mathbb{C} \mid |z| > \max(|\alpha|, |\beta|)\}$ est la somme de la série $\hat{f}$.

On peut en fait montrer que la convergence des approximants de Padé de la série décalée des moments est *uniforme* sur tout fermé de $\mathbb{C} \setminus [\alpha, \beta]$ (théorème de Markoff, 1896), par exemple en utilisant les propriétés des familles normales (*cf.* le module *Fonctions analytiques* du livre [L2]). C'est l'objet de l'exercice III.4.30).

Remarque. Dans le cas d'un poids sur un intervalle (α, β) *non borné*, la transformée de Stieltjes est une fonction holomorphe sur $\mathbb{C}\setminus\mathbb{R}$, mais elle n'est plus en général holomorphe au voisinage de l'infini et l'on ne peut plus l'interpréter comme la somme de la série génératrice des moments décalée (qui est en général *divergente*). On peut, dans certains cas, surmonter cette difficulté en utilisant la théorie des *fractions continues* fonctionnelles, que nous verrons plus loin (*cf.* 4.2.2). Nous traiterons en détail le cas particulier des polynômes de Laguerre (*cf.* 4.2.3).

Inversion de la transformation de Stieltjes

On peut retrouver une mesure positive μ à partir de la restriction de sa transformée de Stieltjes au demi-plan supérieur $\{\operatorname{Im} z > 0\}$, par une formule du type « valeur au bord », faisant intervenit la *fonction de répartition* F de la mesure (*cf.* la page 554).

Théorème 82 (Formule d'inversion de Stieltjes-Perron). Soient μ une mesure positive définie par un poids sur un intervalle ou à support fini et F sa fonction de répartition. Soit $S(\mu)$ sa transformée de Stieltjes.
Pour tous $c, d \in \mathbb{R}$, $c < d$, on a :

$$\frac{F(d) + F(d^-)}{2} - \frac{F(c) + F(c-)}{2} = \frac{1}{\pi} \lim_{\substack{y \to 0 \\ y > 0}} \int_c^d \operatorname{Im} S(\mu)(t + iy)\, \mathrm{d}t.$$

Nous admettrons ce résultat, nous contentant de le prouver pour deux exemples (pour le cas général, *cf.* l'exercice III.4.28).

Exemples.

1. Soit $\mu := \delta$ la masse de Dirac à l'origine. Sa fonction de répartition est la fonction de Heaviside H. Soient $c, d \in \mathbb{R}^*$. On définit (après vérification de l'existence des limites) :

$$G(c,d) := \frac{1}{\pi} \lim_{\substack{y \to 0 \\ y > 0}} \int_c^d \operatorname{Im} S(\mu)(t + iy)\, \mathrm{d}t = \frac{1}{\pi} \lim_{\substack{y \to 0 \\ y > 0}} \int_c^d \operatorname{Im} \frac{1}{t + iy}\, \mathrm{d}t$$
$$= -\frac{1}{\pi} \lim_{\substack{y \to 0 \\ y > 0}} \int_c^d \frac{y}{t^2 + y^2}\, \mathrm{d}t = -\frac{1}{\pi} \lim_{\substack{y \to 0 \\ y > 0}} \int_{c/y}^{d/y} \frac{du}{u^2 + 1}\, \mathrm{d}t = -\frac{1}{\pi} \Big[\arctan u \Big]_{\pm\infty}^{\pm\infty}.$$

   On a $G(c,d) = 0, 1, -1$ selon les signes respectifs de c et d. On vérifie que $G(c,d) = H(c) - H(d)$. Les cas où $c = 0$ ou $d = 0$ sont laissés au lecteur.

2. Soient $\alpha = -1$, $\beta = 1$ et $w = 1$ (le cas des polynômes de Legendre). On vérifie que si $\ln z$ est la détermination principale du logarithme :

$$\frac{1}{\pi} \lim_{\substack{y \to 0 \\ y > 0}} \operatorname{Im} \ln(x + iy) = \begin{cases} 0 & \text{si} \quad x > 0 \\ 1 & \text{si} \quad x < 0 \\ \frac{1}{2} & \text{si} \quad x = 0. \end{cases}$$

   On a vu que $S(\mu)(z) = \ln(z+1) - \ln(z-1)$. On vérifie que, pour $t \in \,]-1, 1[$:

$$1 = -\frac{1}{2i\pi} \left[\lim_{\substack{y \to 0 \\ y > 0}} \Big(S(\mu)(t + iy) - S(\mu)(t - iy) \Big) \right]$$

   et, pour $t \notin [-1, 1]$:

$$0 = -\frac{1}{2i\pi} \left[\lim_{\substack{y \to 0 \\ y > 0}} \Big(S(\mu)(t + iy) - S(\mu)(t - iy) \Big) \right]$$

   Les cas où $t = 1$ ou $t = -1$ sont laissés au lecteur.

Remarque.

Certains auteurs utilisent, au lieu de la transformation de Stieltjes, la *transformation de Cauchy*, qui s'écrit (en variable ξ) $C(\mu')(\xi) = \frac{1}{2i\pi}\,\mu'_{(u)}\left(\frac{1}{u-\xi}\right)$. Nous laissons le lecteur comparer les deux transformations à titre d'exercice pour $0 \leqslant \alpha < \beta$ (par définition $\mu'_{(u)}$ est la mesure de densité $\omega(u)du$, avec $\omega(u) = \frac{1}{u}w\left(\frac{1}{u}\right)$).

L'interprétation en termes de transformation de Cauchy permet de comprendre le lien entre la transformation de Stieltjes et la classique *formule de Cauchy* en théorie des fonctions holomorphes.

4.2 Polynômes orthogonaux et fractions continues

La théorie des fractions continues est initialement une méthode d'approximation des nombres réels par des nombres rationnels. Au XIXe siècle, Euler a créé une méthode analogue pour l'approximation des fonctions analytiques au voisinage de l'infini. Du point de vue mathématique, c'est essentiellement dans ce cadre qu'ont été découverts et étudiés les polynômes orthogonaux d'Euler à Stieltjes (par Gauß, Jacobi, Tchebychev, Hermite, Padé, Laguerre...). Le point de vue de l'orthogonalité n'a été clairement dégagé que plus tard. L'approche par fractions continues en relation avec la transformée de Stieltjes (ou de façon équivalente la transformée dite de Cauchy, *cf.* la remarque ci-dessus) est, après un quasi oubli, revenue d'actualité dans les recherches contemporaines sur les polynômes orthogonaux et leurs applications, avec ce que l'on appelle la *Méthode de Riemann Hilbert* (*cf.* [Dei]). Cette dernière partie du module est consacrée à une brève présentation du sujet. Les résultats essentiels sont le théorème 87 et le corollaire 88.

4.2.1 Fractions continues

On considère, dans ce paragraphe, des fractions continues (finies ou infinies) *générales* à coefficients dans le corps $\mathbb{C}(Z)$ des fractions rationnelles sur $\mathbb{C}$ (et plus précisément dans l'anneau $\mathbb{C}[Z]$), *cf.* le module I.1, page 59 :

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \cdots}}} \tag{67}$$

où les $(b_n)_{n\in\mathbb{N}}$ et les $(a_n)_{n\in\mathbb{N}^*}$ sont des polynômes : $b_n \in \mathbb{C}[Z]$, $a_n \in \mathbb{C}[Z]^*$.

Nous simplifierons la notation de cette fraction en $b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots + \frac{a_n|}{|b_n} + \cdots$

La valeur d'une fraction continue *finie* est définie par récurrence descendante comme :

$$b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots + \frac{a_n|}{|b_n} = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2|}{|b_2} + \cfrac{a_3|}{|b_3} + \cdots + \cfrac{a_n|}{|b_n}}.$$

On appelle *réduites* de la fraction continue (67) les $r_k := b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots + \frac{a_k|}{|b_k}$: r_k est la $k^{ème}$ réduite. On a, en utilisant une homographie : $b + \frac{a}{r} = \frac{br+a}{r} = \begin{bmatrix} b & a \\ 1 & 0 \end{bmatrix}(r)$, où (r)

désigne le vecteur colonne $\begin{pmatrix} r \\ 1 \end{pmatrix}$. On en déduit :

$$b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots + \frac{a_n|}{|b_n} = \begin{bmatrix} b_0 & a_1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} b_1 & a_2 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} b_{n-1} & a_n \\ 1 & 0 \end{bmatrix} \begin{bmatrix} b_n & 1 \\ 1 & 0 \end{bmatrix} (\infty)$$

avec $(\infty) := \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Il n'y a pas de problème de définition car tous les déterminants $\begin{vmatrix} b_k & a_{k+1} \\ 1 & 0 \end{vmatrix} = -a_{k+1}$ ($k \in [\![0, n-1]\!]$) sont non nuls. En effet les matrices $\begin{bmatrix} b_n & 1 \\ 1 & 0 \end{bmatrix}$ sont alors inversibles et l'on ne peut pas avoir $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ dans la suite des calculs, sinon le terme précédent serait aussi $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ et l'on aboutirait à une contradiction. Ainsi tous les résultats sont des vecteurs colonnes définissant des éléments de la droite projective $\mathbf{P^1}(\mathbb{C}(Z)) = \mathbb{C}(Z) \cup \{\infty\}$ sur le corps $\mathbb{C}(Z)$.

La notion de $k^{ème}$ réduite s'étend de façon évidente à une fraction continue *infinie* (son calcul ne fait intervenir que les $(k+1)$ premiers b_i et les k premiers a_i).

D'après le théorème 52 de la page 59 et la proposition 54 du module I.8 (avec un changement de notation), on a le résultat suivant.

Théorème 83. Soit $b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots + \frac{a_n|}{|b_n} + \cdots$ une fraction continue (finie ou infinie). Définissons deux suites (R_n) et (S_n) ($n \in [\![0, k]\!]$, respectivement $n \in \mathbb{N}$) par les relations :

$$R_0 = b_0, \quad S_0 = 1, \quad R_1 = b_0 b_1 + a_1, \quad S_1 = b_1, \tag{68}$$

et, pour $n > 1$:

$$R_n = b_n R_{n-1} + a_n R_{n-2}, \quad S_n = b_n S_{n-1} + a_n S_{n-2}. \tag{69}$$

Alors les réduites sont données par les formules R_n/S_n. Elles vérifient les relations :

$$\frac{R_n}{S_n} - \frac{R_{n-1}}{S_{n-1}} = (-1)^{n-1} \frac{a_1 a_2 \ldots a_n}{S_n S_{n-1}}. \tag{70}$$

4.2.2 Polynômes orthogonaux et réduites

Soit $b_0 + \frac{a_1|}{|b_1} + \frac{a_2|}{|b_2} + \cdots + \frac{a_n|}{|b_n} + \cdots$ une fraction continue *infinie* de $\mathbb{C}(Z)$, vérifiant les conditions du paragraphe précédent.

La comparaison des formules de récurrence pour les réduites de cette fraction continue (*cf.* les relations (68) et (69) ci-dessus) avec les formules de récurrence pour les polynômes orthogonaux associés à un poids w (ou plus généralement à une forme linéaire *strictement positive* $\mathscr{M}$) :

$$q_n = (A_n t + B_n) q_{n-1} - C_n q_{n-2} = 0$$

suggère de poser $b_0 := 0$, $a_1 := 1$, $b_n := A_n z + B_n$ ($n > 0$), $a_n := -C_n$ ($n > 1$) et de

considérer la fraction continue $\mathbf{F}_w$ (respectivement $\mathbf{F}_{\mathscr{M}}$) (de $\mathbb{R}[Z]$) :

$$\mathbf{F}_w := \frac{1|}{|A_1 z + B_1} - \frac{C_2|}{|A_2 z + B_2} + \cdots - \frac{C_n|}{|A_n z + B_n} - \cdots$$

(on a vu que $C_n \neq 0$, pour tout $n \in \mathbb{N}^*$, on a donc $a_n \neq 0$).
Nous choisissons ici les polynômes orthogonaux $\{q_n\}_{n\in\mathbb{N}}$ *orthonormés* (il est facile de généraliser pour une autre suite en utilisant les χ_n au lieu des ξ_n).
On note, comme plus haut, μ_n les moments du poids w et $D_1 := \mu_0\mu_2 - \mu_1^2$ le déterminant de Hankel de rang 2. Dans le cas d'une forme linéaire $\mathscr{M}$, on note $c_n := \mathscr{M}(t^{n-1})$, pour tout $n \in \mathbb{N}^*$ et $\Delta_1 := c_1c_3 - c_2^2$.
Nous traiterons le cas d'un poids. Celui d'une forme linéaire strictement positive se traite de façon tout à fait similaire (tant que l'on reste dans un cadre algébrique...), il suffit de remplacer dans les formules μ_0 par c_1 et D_1 par Δ_1.

Proposition 84. Soit $\mathbf{F}_w$ la fraction continue associée à la suite $\{q_n\}_{n\in\mathbb{N}}$ des polynômes orthogonaux *orthonormés* pour w. On note $\{r_n\}_{n\in\mathbb{N}}$ la suite des polynômes de seconde espèce associée à $\{q_n\}_{n\in\mathbb{N}}$.
Les réduites $R_n/S_n \in \mathbb{R}(Z)$ de la fraction continue $\mathbf{F}_w$ sont déterminées par les formules :

$$(i) \quad R_n(z) = \sqrt{\frac{D_1}{\mu_0^{\frac{3}{2}}}} \int_\alpha^\beta \frac{q_n(z) - q_n(\tau)}{z - \tau}\, w(\tau)d\tau = \sqrt{\frac{D_1}{\mu_0^{\frac{3}{2}}}}\, r_n,$$

$$(ii) \quad S_n(z) = \sqrt{\mu_0}\, q_n(z).$$

Ainsi les dénominateurs des réduites forment une suite de polynômes orthogonaux pour le poids w (en remplaçant z par t).

Démonstration. L'assertion (ii) est immédiate : les S_n et les q_n vérifient les mêmes formules de récurrence, et la formule est vraie pour $n = 0$ et $n = 1$.
L'assertion (i) est vraie pour $n = 0$.
On a $q_1(z) = \xi_1 z + \xi_1'$, donc $\displaystyle\int_\alpha^\beta \frac{q_1(z) - q_1(\tau)}{z - \tau} w(\tau)d\tau = \xi_1\mu_0$, et, par ailleurs $\sqrt{D_1} = \frac{\sqrt{\mu_0}}{\xi_1}$ (*cf.* le corollaire 66 de la page 620), donc $\frac{\sqrt{D_1}}{\mu^{\frac{3}{2}}}\xi_1\mu_0 = 1 = R_1$. L'assertion (i) est donc vraie pour $n = 1$. On termine la preuve en utilisant la proposition 5 de la page 540. ∎

On considère, pour tout $n \in \mathbb{N}$, les développements en série à l'infini des réduites :

$$R_n(z)/S_n(z) = d_{0n}z^{-1} + d_{1n}z^{-2} + \cdots + d_{kn}z^{-k} + \cdots$$

Proposition 85. Pour tout $n \in \mathbb{N}$, La fraction rationnelle $\frac{\mu_0^2}{\sqrt{D_1}}\frac{R_n}{S_n}$ est l'approximant de Padé d'ordre $[n/n]$ de la série génératrice des moments décalée : $\sum\limits_{n\in\mathbb{N}} \mu_n/z^n$.

Démonstration. D'après la proposition 84, on a $\frac{R_n}{S_n} = \frac{\mu_0^2}{\sqrt{D_1}}\frac{r_n}{q_n}$. Les r_n/q_n sont les ap-

proximants de Padé d'ordre $[n, n]$ de la série décalée des moments de w. Nous l'avons montré pour les polynômes unitaires (*cf.* le corollaire 77 de la page 634), mais $\frac{r_n}{q_n}$ est évidemment indépendant de la suite de polynômes orthogonaux choisis. ■

On prendra garde au fait que les séries entières formelles :

$$d_0 z^{-1} + d_1 z^{-2} + \cdots + d_k z^{-(k+1)} + \cdots$$

figurant dans la proposition ci-dessus ne sont pas en général convergentes. Elles le sont si (α, β) est *borné*.

Proposition 86. Les réduites R_n/S_n de $\mathbf{F}_w$ admettent la décomposition en éléments simples :

$$\frac{R_n(z)}{S_n(z)} = \mu_0^{-2}\sqrt{D_1} \sum_{k=1}^{n} \frac{\lambda_{kn}}{z - t_{kn}},$$

où les t_{kn} sont les zéros des polynômes orthogonaux associés à w et les λ_{kn} les nombres de Christoffel correspondants.

Démonstration.
On a $\frac{R_n(z)}{S_n(z)} = \mu_0^{-2}\sqrt{D_1}\,\frac{r_n}{q_n}$, d'où le résultat en utilisant le corollaire 17 de la page 551. ■

On dit qu'une fraction continue $\mathbf{F}$ converge en $z \in \mathbb{C}$ si, pour tout $n \in \mathbb{N}$, $S_n(z) \neq 0$ et si la suite $\big(R_n(z)/S_n(z)\big)_{n\in\mathbb{N}}$ converge. On note alors $\mathbf{F}(z)$ la limite.
Tous les résultats établis ci-dessus pour un poids sont des résultats *d'algèbre* et restent valables, *mutatis mutandis*, pour une forme linéaire *strictement positive*. Il n'en est pas en général de même pour l'important résultat ci-dessous qui est basé sur des résultats *d'analyse*.

Théorème 87 (Stieltjes). Soit w un poids sur un intervalle *borné* (α, β). La suite des réduites (R_n/S_n) de la fraction continue $\mathbf{F}_w$ est une suite de fonctions holomorphes sur $\mathbb{C} \setminus [\alpha, \beta]$. Cette suite *converge simplement* sur cet ouvert vers $\mu_0^{-2}\sqrt{D_1}S(\mu)(z)$, où $S(\mu)(z)$ est la transformée de Stieltjes de μ. Ainsi la fraction continue $\mathbf{F}_w$ converge sur $\mathbb{C} \setminus [\alpha, \beta]$ et sa limite est égale à $\mu_0^{-2}\sqrt{D_1}S(\mu)$.

Démonstration. Les réduites sont des fractions rationnelles. Pour tout $n \in \mathbb{N}$, les zéros de S_n sont ceux de q_n, ils appartiennent donc à $]\alpha, \beta[$. Les réduites sont donc holomorphes sur $\mathbb{C} \setminus [\alpha, \beta]$. Le théorème est alors une conséquence immédiate du résultat ci-dessus et du théorème 20 de la page 554. ■

Notons que le domaine de convergence de la fraction continue $\mathbf{F}_w$ est *toujours strictement plus grand* que le disque de convergence de la série des moments. Cela sera encore plus frappant dans le cas de la mesure des polynômes de Laguerre, dont la série des moments est divergente tandis que la fraction continue associée converge sur $\mathbb{C} \setminus \mathbb{R}_+$, comme nous le verrons dans le paragraphe suivant.

On peut en fait montrer que la convergence de (R_n/S_n) vers $\mu_0^{-2}\sqrt{D_1}S(\mu)$ est *uniforme* sur tout fermé de $\mathbb{C} \setminus [\alpha, \beta]$ (*cf.* ci-dessus la page 639 et l'exercice III.4.30).

Corollaire 88. Soit w un poids sur un intervalle *borné* (α, β).
On note $A := \max(|\alpha|, |\beta|)$. Pour tout $z \in \mathbb{C}$, tel que $|z| > 1/A$, on a :

$$\mathbf{F}_w(z) = \int_\alpha^\beta \frac{w(t)}{z-t}\,\mathrm{d}t = \sum_{n\in\mathbb{N}} \frac{\mu_n}{z^{n+1}}\cdot$$

4.2.3 Série d'Euler et intégrale de Laguerre

Quand l'intervalle (α, β) n'est pas borné, les résultats du paragraphe précédent peuvent ne plus être vrais. Nous avons vu, par exemple, que la série génératrice des moments peut diverger. Toutefois, on peut montrer que, sous certaines hypothèses, bien que cette série diverge, il est possible que la suite des réduites dans l'écriture en fractions continues correspondante converge vers la transformée de Stieltjes $S(\mu)$. Cela fournit un procédé de « resommation » de la série divergente des moments, généralisant la sommation des séries entières convergentes. L'idée de ce procédé remonte à Euler. Nous ne traiterons ici qu'un exemple, celui des polynômes de Laguerre[17]. Nous allons énoncer un certain nombre de résultats dont les preuves feront l'objet d'une série d'exercices.

Dans le texte original de Laguerre, l'intégrale (dite aujourd'hui de Laguerre) est écrite $\displaystyle\int_x^{+\infty} \frac{\mathrm{e}^{-x}\,\mathrm{d}x}{x}\cdot$

Avec la notation actuelle, on écrit cette intégrale $\displaystyle\int_x^{+\infty} \frac{\mathrm{e}^{-u}\,\mathrm{d}u}{u}$ $(x > 0)$. Pour rester en accord avec nos notations précédentes, nous étudierons la variante $\displaystyle\int_{-x}^{+\infty} \frac{\mathrm{e}^{-u}\,\mathrm{d}u}{u}$ $(x < 0)$. On introduit $\mathrm{e}^{-x}I_0(x) := \mathrm{e}^{-x}\displaystyle\int_{-x}^{+\infty} \frac{\mathrm{e}^{-u}\,\mathrm{d}u}{u} = \int_{-x}^{+\infty} \frac{\mathrm{e}^{-(u+x)}\,\mathrm{d}u}{u}\cdot$ En posant $t := u + x$, on obtient :

$$\mathrm{e}^{-x}I_0(x) = \int_0^{+\infty} \frac{\mathrm{e}^{-t}}{t-x}\,\mathrm{d}t.$$

On observe que l'intégrale $\displaystyle -\int_0^{+\infty} \frac{\mathrm{e}^{-t}}{z-t}\,\mathrm{d}t$ converge, pour tout $z \in \mathbb{C} \setminus \mathbb{R}_+$, et définit la transformée de Stieltjes $S(\mu)(z)$ de la mesure μ définie par le poids de Laguerre $w : t \mapsto w(t) = \mathrm{e}^{-t}$, sur $[0, +\infty[$:
Le but de Laguerre était d'étudier, pour $x \in \mathbb{R}$, $x < 0$, à la suite d'Euler, la série entière formelle $\displaystyle\sum_{n\in\mathbb{N}} \frac{n!}{x^{n+1}}$ (dite d'Euler[18]) et de montrer que cette série, bien que divergente, peut fournir de bonnes approximations de $S(\mu)(x)$, par une réécriture sous forme de fraction continue convergente. Nous avons déjà rencontré plus haut la série d'Euler, c'est la série des moments décalée de μ quand la variable est réelle ($\xi = 1/x$).
On introduit les fonctions polynomiales en $1/x$:

$$f_0 = 0 \quad \text{et, pour } n \in \mathbb{N}^* \quad x \mapsto f_n(x) = \frac{1}{x} + \frac{1!}{x^2} + \cdots + \frac{(n-1)!}{x^n}\cdot$$

[17] Historiquement cet exemple est dû à Laguerre. Nous reprenons essentiellement sa méthode en "modernisant" les notations.
[18] Euler appelle *série de Wallis* la série numérique obtenue en faisant $x = -1$.

La fonction f_n est la somme des $n-1$ premiers termes de la série d'Euler. Contrairement au cas d'un poids sur un intervalle borné, la série des moments décalée n'est plus le développement en série de $S(\mu)(z)$. Les f_n sont toutefois les parties régulières de développements limités de $S(\mu)$, au voisinage de $-\infty$ (*cf.* l'exercice III.4.31 de la page 661).

On introduit les fonctions $I_n(x) := \displaystyle\int_{-x}^{+\infty} \frac{\mathrm{e}^{-u}}{u^{n+1}}\,\mathrm{d}u$ et $J_n(x) = n!\displaystyle\int_{-x}^{+\infty} \frac{(u+x)^n \mathrm{e}^u}{u^{n+1}}\,\mathrm{d}u$ $(n \in \mathbb{N})$.

On vérifie que, pour tout $n \in \mathbb{N}$, J_n est combinaison linéaire des $x \mapsto x^p I_p$, $p \leqslant n$.

On peut montrer (*cf.* les exercices III.4.32, III.4.33, III.4.34) les résultats suivants.

1. On a : $xJ_n' = nJ_n - n^2 J_{n-1}$, $J_n' = n(-J_{n-1} + J_{n-1}')$ et la suite $(J_n)_{n\in\mathbb{N}}$ vérifie la relation de récurrence des polynômes de Laguerre (*cf.* 54), page 600) :
$$\forall n \in \mathbb{N}^*, \quad J_{n+1} + (x - 2n - 1)J_n + n^2 J_{n-1} = 0.$$
2. Pour tout $n \in \mathbb{N}^*$, J_n est solution de l'équation différentielle de Laguerre :
$$xy'' + (1-x)y' + ny = 0.$$
3. On a, pour tout $n \in \mathbb{N}^*$:
a) $J_n = L_n I_0 - h_n \mathrm{e}^x$, où L_n est le polynôme de Laguerre d'ordre n et h_n est un polynôme déterminé de manière unique. Le polynôme h_n est de degré $n-1$, le coefficient de son terme de plus haut degré est $(-1)^{n+1}$, et la suite (h_n) vérifie la relation de récurrence des polynômes de Laguerre :
$$\forall n \geqslant 1, \quad h_{n+1} + (x - 2n - 1)h_n + n^2 h_{n-1} = 0\,;$$
b) la fonction polynomiale h_n, de la variable complexe z, est la *partie entière* de la fraction rationnelle $L_n f_n$ et la fraction rationnelle $\frac{h_n}{L_n}$ est la réduite d'ordre n de la fraction continue associée à la mesure de Laguerre μ.
4. Le développement en série à l'infini de la fraction continue :
$$\frac{1|}{|-z+1} - \frac{1|}{|-z+3} - \cdots - \frac{(n-1)^2|}{|-z+(2n-1)} + \cdots$$
est la série formelle $-\sum\limits_{n\geqslant 0} n!/z^{n+1}$, opposée de la série génératrice des moments décalée.

On peut montrer le résultat suivant.

Théorème 89. La suite des réduites $\dfrac{h_n}{L_n}$ de la fraction continue associée à la mesure de Laguerre μ, de densité $w(t) = \mathrm{e}^{-t}$ (sur $\mathbb{R}_+$) converge simplement, dans le demi-plan $\{\operatorname{Re} z < 0\}$, vers la transformée de Stieltjes $z \mapsto S_\mu(z) = \displaystyle\int_0^{+\infty} \frac{\mathrm{e}^{-t}}{z-t}\,\mathrm{d}t$.

Si z est réel et $z = \operatorname{Re} z > 0$, le théorème se déduit de l'exercice III.4.35. Pour z complexe la preuve est similaire (en reprenant *mutatis mutandis* les exercices III.4.32, III.4.33, III.4.34, III.4.35). On peut écrire, pour $\operatorname{Re} z < 0$:
$$\sum_{n\geqslant 0} n!/z^{n+1} = \frac{1|}{|z-1} - \frac{1|}{|z-3} - \cdots - \frac{(n-1)^2|}{|z-(2n-1)} - \cdots,$$
en convenant que le premier membre est défini par le second, qui converge sur le demi-plan $\{\operatorname{Re} z < 0\}$ (*cf.* l'exercice de la page 65).

En appliquant cela pour $z := -1$, on *définit* comme « resommée » de la série numérique *divergente* :

$$\sum_{n=0}^{+\infty}(-1)^n n! = 1 - 1! + 2! - 3! + \cdots$$

(qu'Euler appelait série de Wallis) la somme α de la fraction continue numérique convergente :

$$\frac{1|}{|2} - \frac{1|}{|4} - \cdots - \frac{(n-1)^2|}{|2n} + \cdots$$

On a $\alpha = \displaystyle\int_0^{+\infty} \frac{\mathrm{e}^{-t}}{1+t}\,\mathrm{d}t$. C'est l'une des valeurs que proposait Euler en 1740 comme somme de la série de Wallis (*cf.* l'exercice III.4.35).

Exercice type corrigé

III.4.0

Le but de l'exercice est de montrer que si l'équation différentielle $\psi'' + (2\lambda + 1 - t^2)\psi = 0$ admet une solution ψ non triviale de carré intégrable sur $\mathbb{R}$, alors $\lambda \in \mathbb{N}$ et $\psi = c\psi_n$, où ψ_n est la $n^{\text{ème}}$ fonction d'Hermite et $c \in \mathbb{R}$. Du point de vue physique, cela permet de déterminer les niveaux d'énergie de l'oscillateur harmonique en mécanique quantique en liaison avec les polynômes d'Hermite (*cf.* la partie 3.4.1). L'exercice permet de comprendre *pourquoi* les polynômes d'Hermite apparaissent dans la quantification de l'oscillateur harmonique. Le cas similaire de l'apparition des polynômes de Laguerre (généralisés) dans la quantification de l'atome d'hydrogène a été traité dans le cours, *cf.* la partie 3.5.1.

Pour alléger l'énoncé et (surtout) la solution, on emploiera la locution : « pour tout t assez grand, l'on a », à la place de « il existe $t_0 > 0$, tel que, pour tout $t \geqslant t_0$ l'on ait ». Le lecteur scrupuleux pourra ajuster les constantes t_0 au fur et à mesure, si nécessaire.

(o) Question préliminaire. Soit $\gamma \in \mathbb{R}$, avec $0 < \gamma$.

Soit $p \in \mathbb{R}[T]$. Montrer que, pour tout $|t|$ assez grand, l'on a $\mathrm{e}^{\gamma t^2} - |p(t)| \geqslant \frac{1}{2}\mathrm{e}^{\gamma t^2}$.

(i) Soit $y : t \mapsto y(t)$ une solution maximale de l'équation différentielle $y'' - 2ty' + 2\lambda y = 0$ ($\lambda \in \mathbb{R}$). Montrer qu'elle est la restriction à $\mathbb{R}$ d'une fonction entière $z \mapsto y(z)$. Montrer que l'espace vectoriel des solutions de l'équation différentielle admet une base formée d'une fonction paire et d'une fonction impaire.

(ii) Soit $\gamma \in \mathbb{R}$, $\frac{1}{2} < \gamma < 1$. Soit $y : t \mapsto y(t)$ une solution, *non identiquement nulle*, de l'équation différentielle $y'' - 2ty' + 2\lambda y = 0$ ($\lambda \in \mathbb{R}$). Montrer les résultats suivants.

a) On suppose la fonction y paire ou impaire. Alors :

- ⋆ ou bien : il existe $C > 0$, tel que $|y(t)| > C\mathrm{e}^{\gamma t^2}$, pour tout $|t|$ assez grand.
- ⋆ ou bien : $\lambda = n \in \mathbb{N}$ et $y = \alpha_1 H_n$, où $\alpha_1 \in \mathbb{R}$ est une constante et H_n le $n^{\text{ème}}$ polynôme d'Hermite.

b) On ne fait plus d'hypothèse supplémentaire sur y (toujours supposée non identiquement nulle). Alors :

- ⋆ ou bien : il existe $C > 0$ tel que $|y(t)| > C\mathrm{e}^{\gamma t^2}$, pour $t > 0$ assez grand,

- ⋆ ou bien : il existe $C > 0$ tel que $|y(t)| > C\mathrm{e}^{\gamma t^2}$, pour $-t > 0$ assez grand.
- ⋆ ou bien : $\lambda = n \in \mathbb{N}$ et $y = \alpha_1 H_n$, où $\alpha_1 \in \mathbb{R}$ est une constante et H_n le $n^{ème}$ polynôme d'Hermite.

(iii) Soit ψ une solution, non identiquement nulle, de carré intégrable (respectivement bornée) sur $\mathbb{R}$, de l'équation différentielle $\psi'' + (2\lambda + 1 - t^2)\psi = 0$ ($\lambda \in \mathbb{R}$). Montrer qu'alors $\lambda = n \in \mathbb{N}$ et $\psi = \alpha_1\psi_n$, où α_1 est une constante et ψ_n la fonction d'Hermite d'ordre n.

Solution.

(o) Soit $p \in \mathbb{R}[T]$. On écrit $\mathrm{e}^{\gamma t^2} - |p(t)| = \mathrm{e}^{\gamma t^2}\big(1 - |p(t)|\mathrm{e}^{-\gamma t^2}\big)$. La fonction $t \mapsto |p(t)|\mathrm{e}^{-\gamma t^2}$ tend vers 0, quand $|t|$ tend vers $+\infty$, donc $1 - |p(t)|\mathrm{e}^{-\gamma t^2} > 1/2$, et $\mathrm{e}^{\gamma t^2} - |p(t)| \geqslant \frac{1}{2}e^{\gamma 1 t^2}$, pour $|t|$ assez grand.

(i) Soient $y(0) = a_0$ et $y'(0) = a_1$. Nous savons que la solution y est déterminée de façon unique par ces conditions et que, si elle est maximale, elle existe sur $\mathbb{R}$ tout entier. Nous allons montrer que y est développable en série entiére à l'origine et que cette série entière a un rayon de convergence infini.

On cherche une série entière *formelle* solution de la forme $\sum\limits_{k\in\mathbb{N}} a_k t^k$, où les a_k sont des réels inconnus pour $k \geqslant 2$. En reportant dans l'équation et en identifiant les coefficients des monômes t^n, on trouve les conditions *nécessaires* : $a_{k+2} = 2\frac{k-\lambda}{(k+1)(k+2)}a_k$, pour $k \geqslant 0$. Ces conditions se séparent en *deux* relations de récurrence. Les a_{2h} ($h \in \mathbb{N}^*$) sont déterminés de façon unique par a_0 et les a_{2h+1} ($h \in \mathbb{N}^*$) sont déterminés de façon unique par a_1. Quand k tend vers l'infini, $2\frac{k-\lambda}{(k+1)(k+2)} \sim \frac{2}{k}\cdot$ Ainsi, d'après la règle de d'Alembert, les séries entières (en u) $\sum\limits_{h=0}^{+\infty} a_{2h}u^h$ et $\sum\limits_{h=0}^{+\infty} a_{2h+1}u^h$ ont un rayon de convergence infini.

Si $a_0 := 1$, la fonction $\sum\limits_{h=0}^{\infty} a_{2h}t^{2h}$ ($t \in \mathbb{R}$), déterminée ci-dessus, est une solution paire.

Si $a_1 := 1$, la fonction $\sum\limits_{h=0}^{+\infty} a_{2h+1}t^{2h}$ ($t \in \mathbb{R}$), déterminée ci-dessus, est une solution impaire. Elles correspondent respectivement aux conditions initiales :

$$y(0) = a_0 = 1, y'(0) = a_1 = 0 \quad \text{et} \quad y(0) = a_0 = 0, y'(0) = a_1 = 1$$

et forment donc une base de l'espace vectoriel des solutions

(ii) a) Il s'agit maintenant de *minorer* les sommes des séries, ce qui est plus délicat.

Quand k tend vers l'infini, $\frac{k(k-\lambda)}{(k+1)(k+2)}$ tend vers 1, il existe donc $k_1 \in \mathbb{N}^*$ (que l'on peut supposer pair : $k_1 := 2h_1$), avec $k_1 > \lambda$ et $k_1 > 2$, tel que, pour tout $k \geqslant k_1$, l'on ait $\gamma < \frac{k(k-\lambda)}{(k+1)(k+2)}$, et donc $\frac{2\gamma}{k} < 2\frac{k-\lambda}{(k+1)(k+2)}\cdot$

On suppose d'abord que y est paire. On peut donc se limiter à $t > 0$. On a $a_1 = 0$ et l'on peut supposer $a_0 \neq 0$, sinon il n'y a rien à démontrer. Quitte à remplacer y par $-y$, on peut supposer que $a_{2h_1} \geqslant 0$.

On suppose d'abord que λ *n'est pas un entier positif pair*. Alors, quel que soit $h \in \mathbb{N}$, $2h - \lambda$ n'est *jamais nul* et, d'après la relation de récurrence, les a_{2h} ($h \in \mathbb{N}$) sont *tous non nuls*. En particulier $a_{2h_1} > 0$. Pour tout $k \geqslant k_1$, $a_{k+2} > \frac{2\gamma}{k}a_k$.

On a :

$$\sum_{h=h_1}^{+\infty} a_{2h}t^{2h} \geqslant a_{2h_1}\big(t^{2h_1} + \frac{2\gamma}{2h_1}t^{2h_1+2} + \frac{2^2\gamma^2}{(2h_1)(2h_1+2)}t^{2h_1+4} + \cdots\big)$$
$$= \frac{(h_1-1)!a_{2h_1}}{\gamma^{h_1-1}}t^2 \sum_{h=h_1-1}^{+\infty} \frac{(\gamma t^2)^h}{h!}.$$

Pour t assez grand, on a $\frac{(h_1-1)!a_{2h_1}}{\gamma^{h_1-1}}t^2 > 1$, donc $\sum\limits_{h=h_1}^{+\infty} a_{2h}t^{2h} \geqslant \sum\limits_{h=h_1-1}^{+\infty} \frac{(\gamma t^2)^h}{h!}$.

On a, toujours pour t assez grand :

$$\sum_{h=h_1}^{+\infty} a_{2h}t^{2h} \geqslant \mathrm{e}^{\gamma t^2} - \sum_{h=0}^{h_1-2} \frac{(\gamma t^2)^h}{h!},$$

puis :

$$\sum_{h=0}^{+\infty} a_{2h}t^{2h} \geqslant \mathrm{e}^{\gamma t^2} - \sum_{h=0}^{h_1-2} \frac{(\gamma t^2)^h}{h!} + \sum_{h=0}^{h_1-1} a_{2h}t^{2h} = \mathrm{e}^{\gamma t^2} - P(t),$$

où P est un polynôme. En utilisant (o), on en déduit que, pour t assez grand,

$$\sum_{h=0}^{+\infty} a_{2h}t^{2h} \geqslant \frac{1}{2}\mathrm{e}^{\gamma t^2}.$$

Il reste à traiter le cas où $\lambda = 2p$, avec $p \in \mathbb{N}$. Alors les a_{2h}, pour $h > p$ sont tous nuls et la solution $\sum\limits_{h=0}^{+\infty} a_{2h}t^{2h}$ est un polynôme de degré $2p$. Celui-ci est proportionnel au polynôme d'Hermite H_{2p} : $y(t) = \frac{a_0}{H_{2p}(0)}H_{2p}$, puisque ce dernier est une solution paire.

Le cas où y est impaire se traite de façon analogue. On peut supposer $a_1 > 0$ et se ramener au cas $t > 0$. On suppose d'abord que λ n'est pas un entier positif impair.

On a :

$$\sum_{h=h_1}^{+\infty} a_{2h+1}t^{2h+1} \geqslant a_{2h_1+1}\Big(t^{2h_1+1} + \frac{2\gamma}{2h_1+1}t^{2h_1+3} + \frac{2^2\gamma^2}{(2h_1+1)(2h_1+3)}t^{2h_1+5} + \cdots\Big)$$
$$\geqslant a_{2h_1+1}\Big(t^{2h_1+1} + \frac{\gamma}{h_1+1}t^{2h_1+3} + \frac{\gamma^2}{(h_1+1)(h_1+2)}t^{2h_1+5} + \cdots\Big)$$
$$= \frac{h_1!a_{2h_1+1}}{\gamma^{h_1}}t \sum_{h=h_1}^{+\infty} \frac{(\gamma t^2)^h}{h!}.$$

On termine comme dans le cas pair. Pour $|t|$ assez grand, on a $|y(t)| \geqslant \frac{1}{2}e^{\gamma t^2}$.

Si $\lambda = 2p+1$ est un entier positif impair, la solution impaire y est proportionnelle au polynôme d'Hermite H_{2p+1}. Cela termine la démonstration de (ii) a).

b) Supposons d'abord que λ ne soit pas un entier positif. La solution générale s'écrit $y := f + g$, où f est une solution paire et g une solution impaire, que l'on peut supposer toutes deux non nulles (sinon il n'y a rien à démontrer). Quitte à remplacer y par $-y$, on peut supposer $a_0 := y(0) = f(0) > 0$. Alors, pour $|t|$ assez grand, on a $f(t) \geqslant \frac{1}{2}e^{\gamma t^2}$. Supposons $a_1 > 0$, alors, pour t assez grand, on a $g(t) \geqslant \frac{1}{2}e^{\gamma t^2}$. Ainsi, pour t assez grand, on a $y(t) \geqslant e^{\gamma t^2}$. On prendra garde au fait que l'on n'obtient *aucune information* sur le comportement de $y(t)$ quand $t \to -\infty$ (l'estimation peut, dans ce cas être *fausse*). Enfin, si $a_1 < 0$, on raisonne de façon symétrique, en supposant que $t \to -\infty$. On a, pour $-t$ assez grand, $f(t) \geqslant \frac{1}{2}e^{\gamma t^2}$ et $g(t) \geqslant \frac{1}{2}e^{\gamma t^2}$, donc $y(t) \geqslant e^{\gamma t^2}$.

Supposons maintenant que $\lambda = 2p$ soit un entier positif pair. La solution générale de notre équation différentielle s'écrit $y = \alpha_1 H_{2p} + g$, où g est une solution *impaire*. Supposons $g \neq 0$, comme λ n'est pas un entier impair, on sait, d'après (ii), que, pour $|t|$ assez grand $|g(t)| \geqslant \frac{1}{2}e^{\gamma t^2}$. On a :

$$|y(t)| \geqslant |g(t)| - |\alpha_1 H_{2p}(t)| \geqslant \frac{1}{2}\big(e^{\gamma t^2} - |2\alpha_1 H_{2p}(t)|\big) \geqslant \frac{1}{4}e^{\gamma t^2},$$

dans les même conditions, d'après (o).

Si $\lambda = 2p + 1$ est un entier positif impair, on raisonne de façon symétrique.

(iii) Soit ψ est une solution, de l'équation différentielle $\psi'' + (2\lambda + 1 - t^2)\psi = 0$ ($\lambda \in \mathbb{R}$), alors elle s'écrit, nous l'avons vu, $\psi(t) = e^{-t^2/2}y(t)$, ou y est une solution de l'équation différentielle $y'' - 2ty' + 2\lambda y = 0$. Si λ n'est pas un entier positif, on en déduit, en utilisant (ii), que $|\psi(t)| \geqslant \frac{1}{2}e^{(\gamma-\frac{1}{2})t^2}$, pour t assez grand, *ou* pour $-t$ assez grand, donc $|\psi(t)|$ tend vers $+\infty$ quand $t \to \infty$ *ou* quand $t \to \infty$ (ce n'est pas nécessairement vrai des deux côtés à la fois !). Ainsi ψ ne peut pas être de carré intégrable, ni bornée.

Si $\lambda := n$ est un entier positif, il existe une droite de solutions, engendrée par la $n^{ème}$ fonction d'Hermite $\psi_n(t) = e^{-t^2/2}H_n(t)$. Si ψ est une solution n'appartenant pas à ce sous-espace, nous avons vu que, pour $|t|$ assez grand, on a $y(t) := e^{t^2/2}\psi(t) \geqslant \frac{1}{4}e^{\gamma t^2}$. Il en résulte que $\psi(t) \geqslant \frac{1}{4}e^{(\gamma-\frac{1}{2})t^2}$ et que $|\psi(t)|$ tend vers $+\infty$, quand t tend vers $\pm\infty$. Ainsi ψ n'est pas de carré intégrable, ni, non plus, bornée.

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 2

III.4.1 Soit $(p_n)_{n\in\mathbb{N}}$ une suite de polynômes orthogonaux pour le poids w sur l'intervalle borné $]\alpha,\beta[$. Soient $a,b \in \mathbb{R}, a \neq 0$. On définit une nouvelle suite $(r_n)_{n\in\mathbb{N}}$ de polynômes par $r_n(t) = p_n(at+b)$ $(n \in \mathbb{N})$.
a) Montrer qu'il existe un poids ω sur un intervalle convenable tel que $(r_n)_{n\in\mathbb{N}}$ soit une suite de polynômes orthogonaux pour ω.
b) On note μ_n (respectivement ν_n) les moments de w (respectivement ω). Calculer la suite des ν_n en fonction de celle des μ_n.

III.4.2 Soit $(p_n)_{n\in\mathbb{N}}$ une suite de polynômes orthogonaux. On désigne, pour $n > 0$, par γ'_n le coefficient de t^{n-1} dans p_n. Montrer que le coefficient B_n de la formule de récurrence du théorème 3 de la page 537 s'écrit $B_n = A_n\left(\frac{\gamma'_n}{\gamma_n} - \frac{\gamma'_{n-1}}{\gamma_{n-1}}\right)$.

III.4.3 On utilise dans cet exercice les notations du théorème 4 de la page 538. Montrer que, pour tout $n \in \mathbb{N}$:

$$q_n(t) = q_0 \begin{vmatrix} A_1t+B_1 & -\sqrt{C_2} & 0 & \dots & 0 & 0 \\ -\sqrt{C_2} & A_2t+B_2 & -\sqrt{C_3} & \dots & 0 & 0 \\ 0 & -\sqrt{C_3} & A_3t+B_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_{n-1}t+B_{n-1} & -\sqrt{C_n} \\ 0 & 0 & 0 & \dots & -\sqrt{C_n} & A_nt+B_n \end{vmatrix}$$

III.4.4 Soit w un poids, sur un intervalle $J \subset \mathbb{R}$. On considère la relation de récurrence (1) de la page 537 pour les polynômes orthonormés $(q_n)_{n\in\mathbb{N}}$ correspondants. Montrer que le $\mathbb{R}$-espace vectoriel des suites $(p_n)_{n\in\mathbb{N}}$ de polynômes vérifiant cette relation et $\deg p_n \leqslant n$, pour tout $n \in \mathbb{N}$, est de dimension 3. En donner une base en utilisant les (q_n).

III.4.5
On considère le poids $w : t \mapsto w(t) = \sqrt{\frac{1-t}{1+t}}$ sur l'intervalle $]-1,1[$.

a) Pour tout $n \in \mathbb{N}$, on pose $J_n = \displaystyle\int_0^{\pi/2} (\cos 2\theta)^n \,\mathrm{d}\theta$. Trouver, pour $n \geqslant 2$ une relation entre J_n et J_{n-2}. Calculer J_n.
b) Calculer les moments μ_n $(n \in \mathbb{N})$ de w, en exprimant μ_n en fonction de J_n et J_{n+1}.
c) On définit, pour tout $n \in \mathbb{N}$, et $t \in]-1,1[$: $p_n(t) := \dfrac{\sin\left((2n+1)\frac{\arccos t}{2}\right)}{\sin\frac{\arccos t}{2}}$.

Montrer que p_n est une fonction polynomiale de degré n. Montrer que les p_n sont orthogonaux pour le poids $w(t)$.

d) Trouver une suite de polynômes orthogonaux pour le poids $\omega : t \mapsto \omega(t) := \sqrt{\frac{1+t}{1-t}}$ sur l'intervalle $]-1,1[$.

III.4.6 Soient $[\alpha,\beta]$ un segment et $c \in [\alpha,\beta]$, montrer qu'il n'existe pas de fonction $g \in \mathcal{C}([\alpha,\beta],\mathbb{R})$ telle que la mesure $\mu_g := g(t)\,\mathrm{d}t$ définie par g soit égale à δ_c.

Exercices sur la section 3

III.4.7 Soit $n \in \mathbb{N}^*$. On désigne par $\mathcal{U}_n$ l'ensemble des polynômes unitaires de $\mathbb{R}_n[T]$.
Calculer $\alpha_n := \inf\limits_{P\in\mathcal{U}_n} \displaystyle\int_{-1}^{1} \frac{P^2(t)}{\sqrt{1-t^2}}\,\mathrm{d}t$.

III.4.8 Soit $\lambda \in \mathbb{R}$. Mettre l'équation différentielle $t^2y'' + (2t+2)y' - \lambda y = 0$ sous forme de Sturm-Liouville.

III.4.9 $**$ **(Polynômes de Bessel)**
Soit $\lambda \in \mathbb{R}$.
a) Montrer que l'équation différentielle $t^2y''+(2t+2)y'-\lambda y = 0$ admet une solution polynomiale si, et seulement si, $\lambda = n(n+1)$, avec $n \in \mathbb{N}$. On note alors y_n l'unique solution polynomiale telle que $y_n(0) = 1$ (y_n est, par définition, le $n^{\text{ème}}$ polynôme de Bessel).
b) Montrer que $y_n(t) = \sum\limits_{k=0}^{n} \frac{(n+k)!}{(n-k)!k!}\left(\frac{t}{2}\right)^k$. Écrire les six premiers polynômes de Bessel.
c) Soient $n, m \in \mathbb{N}$, $n \neq m$.
Montrer, en utilisant l'exercice III.4.8 ci-dessus, que $\displaystyle\int_U y_n(z)y_m(z)e^{-2/z}dz = 0$, où U est le cercle unité du plan complexe. On dit que y_n et y_m sont des polynômes *orthogonaux sur le cercle unité*.
d) Montrer que le polynôme $p_n(z) := e^{2/z}\frac{\mathrm{d}^n}{\mathrm{d}z^n}(z^{2n}e^{-2/z})$ est orthogonal sur U aux monômes z^k, pour $k \in [\![0, n-1]\!]$. Peut-on en déduire une formule de type Olinde Rodrigues pour y_n ?.
e) Soit $f : \mathbb{R}^* \to \mathbb{R}$ une fonction de classe $\mathcal{C}^\infty$.
On définit une fonction $\varphi : \mathbb{R}^* \to \mathbb{R}$ par $\varphi(\xi) := f(1/\xi)$ ($t\xi = 1$). Montrer que, pour tout $n \in \mathbb{N}$:

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n(f) = (-1)^n\left(\xi^2\frac{\mathrm{d}}{\mathrm{d}\xi}\right)^n(\varphi) = (-1)^n\xi^{n+1}\left(\frac{\mathrm{d}}{\mathrm{d}\xi}\right)^n\xi^{n-1}(\varphi). \tag{71}$$

f) Montrer, en utilisant la question précédente, une formule de type Olinde Rodrigues pour y_n.
g) Montrer que les polynômes de Bessel satisfont une relation de récurrence du type :

$$y_{n+1} = a_n t y_n + b_n y_{n-1} \qquad \text{avec} \qquad a_n, b_n \in \mathbb{N}.$$

h) On appelle moment d'ordre n du poids $e^{-2/z}$ sur U le nombre $\mu_n := \frac{1}{2i\pi}\displaystyle\int_U z^n e^{-2/z}dz$.
Calculer μ_n, pour $n \in \mathbb{N}$. Calculer le déterminant de Hankel correspondant.

III.4.10 $**$ **(Polynômes de Poisson-Charlier)**

Soit $a > 0$ (fixé dans tout l'exercice).

On définit, pour $t \in \mathbb{R}_+$, $j(t) := e^{-a}a^t/\Gamma(t+1)$.

On étudie en probabilités la distribution de Poisson définie par $j(h) = e^{-a}a^h/h!$, $h \in \mathbb{N}$. Le but de l'exercice est de construire une suite de polynômes orthogonaux pour cette distribution c'est-à-dire (par définition) pour la forme linéaire $\mathscr{M} : p \in \mathbb{R}[T] \mapsto \sum\limits_{h\in\mathbb{N}} p(h)j(h)$ (les polynômes de *Poisson-Charlier*).

On note, pour $y \in \mathbb{R}$ et $h \in \mathbb{N}$, $\binom{y}{h} := \frac{y(y-1)\dots(y-h+1)}{h!}$.

On définit, pour $n \in \mathbb{N}, t \in \mathbb{R}$, $p_n(t) = a^{n/2}(n!)^{-1/2} \sum\limits_{h=0}^{n} (-1)^{n-h}\binom{n}{h}h!a^{-h}\binom{t}{h}$. Vérifier que p_n définit un polynôme de degré n en t ; c'est, par définition, le polynôme de Poisson-Charlier d'ordre n.

1. a) Montrer que $\sum\limits_{h=0}^{+\infty} j(h) = 1$. Montrer que, pour tout $n \in \mathbb{N}$, la série $\sum\limits_{h\in\mathbb{N}} h^n j(h)$ est convergente. Montrer que, pour tout polynôme $P \in \mathbb{R}[T]$ la série $\sum\limits_{h\in\mathbb{N}} P(h)j(h)$ est convergente. On note $\mu(P)$ sa somme, et, pour $P, Q \in \mathbb{R}[T]$ on pose $(P \mid Q)_\mu = \mu(PQ)$. Si $(P \mid Q)_\mu = 0$, on dira que P et Q sont orthogonaux pour la mesure de Poisson (de paramètre a).

b) On note, pour $x, t \in \mathbb{R}$, avec $t > 0$ et $|x| < a$, $G(t,x) = e^{-x}(1 + a^{-1}x)^t$.

Montrer que la série $\sum\limits_{h\in\mathbb{N}} a^{-h}\binom{t}{h}x^h e^{-x}$ converge et que sa somme est $G(t,x)$.

c) Montrer que pour tous $x, t \in \mathbb{R}$, avec $t > 0$ et $|x| < a$, la série double :

$$\sum_{h\in\mathbb{N}}\sum_{n\geqslant h} \frac{(-1)^{n-h}}{n!}\binom{n}{h}h!a^{-h}\binom{t}{h}x^n$$

converge absolument, et que l'on a $G(t,x) = \sum\limits_{h=0}^{+\infty}\sum\limits_{n=h}^{+\infty} \frac{(-1)^{n-h}}{n!}\binom{n}{h}h!a^{-h}\binom{t}{h}x^n$.

En déduire, en appliquant le théorème de Fubini, que, dans les même conditions :

$$G(t,x) = \sum_{n=0}^{+\infty} a^{-n/2}(n!)^{-1/2}p_n(t)x^n.$$

d) Montrer que, pour tous $u, v \in \mathbb{R}$, $0 < |u| < a$, $0 < |v| < a$, l'on a :

$$e^{a^{-1}uv} = \sum_{n=0}^{+\infty} j(n)G(n,u)G(n,v)$$

(on utilisera l'identité $a(1 + a^{-1}u)(1 + a^{-1}v) = a + (u + v) + a^{-1}uv$).

e) Montrer que :

$$e^{a^{-1}uv} = \sum_{h=0}^{+\infty}\sum_{k=0}^{+\infty}\Big[\sum_{n=0}^{+\infty} j(n)a^{-h/2}(h!)^{-1/2}p_h(n)a^{-k/2}(k!)^{-1/2}p_k(n)\Big]u^h v^k.$$

En déduire que, pour tous $h, k \in \mathbb{N}$:

$$\sum_{n=0}^{+\infty} j(n)a^{-h/2}(h!)^{-1/2}p_h(n)a^{-k/2}(k!)^{-1/2}p_k(n) = a^{-h}(h!)^{-1}\delta_{hk}$$

(on utilisera l'exercice III.4.23, ci-dessous).

2. a) Soit f une fonction sur $\mathbb{R}$, on définit la fonction Δf par : $\Delta f(t) := f(t+1) - f(t)$. Montrer que, pour tout $n \in \mathbb{N}^*$, l'on a $\Delta^n f(t) = f(t+n) - \binom{n}{1} f(t+n-1) + \cdots + (-1)^n f(t)$.
b) On définit, pour $t \in \mathbb{R}_+$, $j(t) := e^{-a} a^t / \Gamma(t+1)$. Montrer que, pour tout $n \in \mathbb{N}$ et tout $t \geqslant n$, $p_n(t) = a^{n/2}(n!)^{-1/2}(-1)^n j(t)^{-1} \Delta^n j(t-n)$ (c'est une version discrète de la formule d'Olinde-Rodrigues).
c) Soit, pour tous $n \in \mathbb{N}$ et $t \in \mathbb{R}$, L_n^{t-n} le polynôme de Laguerre d'ordre n de paramètre $t-n$. Comparer $p_n(t)$ et $L_n^{t-n}(a)$.

III.4.11 On désigne par $(P_n)_{n\in\mathbb{N}}$ les polynômes de Legendre.
a) Montrer, pour tout $n \in \mathbb{N}$, la relation : $(1-t^2)P_n' + ntP_n = nP_{n-1}$.
b) En déduire les relations :

$$\forall n \in \mathbb{N}^*, \quad P_{n-1}'(t) = tP_n'(t) - nP_n(t) \quad \text{et} \quad P_{n+1}'(t) = tP_n'(t) + (n+1)P_n(t).$$

III.4.12 a) Montrer directement, sans utiliser la théorie générale, que, pour tout $n \in \mathbb{N}$, et tout $a \in \mathbb{R}$, l'équation différentielle $ty'' + (1+a-t)y' + ny = 0$ admet une solution polynomiale non triviale. Montrer que cette solution est de degré n, et qu'elle est unique à une constante multiplicative près.
b) Même question pour $y'' - 2ty' + 2ny = 0$, pour tout $n \in \mathbb{N}$.

III.4.13 Montrer que, pour tous $n \in \mathbb{N}^*$, a, $b > -1$, $\sum\limits_{h=0}^{n} \binom{n+a}{h}\binom{n+b}{n-h} = \binom{2n+a+b}{n}$ (on pourra utiliser un développement en série).

III.4.14 Le but de cet exercice est de prouver la proposition 42 de la page 585.
On pose, pour $t, \tau \in \mathbb{R}$, avec $t \neq \tau$, $K(t,\tau) = \frac{1}{t-\tau}$.

a) Montrer que $\dfrac{\partial}{\partial t}\left((1-t^2)\dfrac{\partial K}{\partial t}\right) = \dfrac{\partial}{\partial \tau}\left((1-\tau^2)\dfrac{\partial K}{\partial \tau}\right)$.

b) Montrer que, pour tout $n \in \mathbb{N}$, la fonction $t \mapsto Q_n(t) := \frac{1}{2}\displaystyle\int_{-1}^{1} \frac{P_n(\tau)}{t-\tau} d\tau$ est de classe $\mathcal{C}^\infty$ sur $]-\infty, -1[$ et $]1, +\infty]$ et que, sur chacun de ces intervalles, elle est une solution de l'équation différentielle de Legendre d'indice n : $(1-t^2)y'' - 2ty' + n(n+1)y = 0$.

c) Montrer que $Q_n = \frac{1}{2}\ln\left|\frac{t+1}{t-1}\right| P_n - W_n$, avec $W_n \in \mathbb{R}_{n-1}[T]$.

d) Montrer que les fonctions P_n et Q_n sont linéairement indépendantes.

III.4.15 On considère le laplacien en dimension $n \geqslant 1$: $\Delta := \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$.
a) Montrer que $\Delta = \operatorname{div} \overrightarrow{\operatorname{grad}}$, en déduire que le laplacien est invariant par isométrie et s'écrit de la même façon dans tous les repères orthonormés.
b) On suppose $n = 2$. Écrire le laplacien en coordonnées polaires.
c) On suppose $n = 3$. Écrire le laplacien en coordonnées cylindriques, puis en coordonnées sphériques.

III.4.16 On considère le laplacien en dimension 2. On utilise des coordonnées cartésiennes (x, y) et des coordonnées polaires (r, θ), et l'on note $\Delta_\theta := \frac{\partial^2}{\partial\theta^2}$.
a) Soit $k \in \mathbb{N}$. On désigne par $\mathcal{F}_k \subset \mathbb{R}[X, Y]$ le $\mathbb{R}$-espace vectoriel des polynômes homogènes de degré k. Quelle est sa dimension ?
b) Montrer que, si f est une fonction polynomiale homogène de degré k, on a $\Delta_\theta f = r^2\Delta f - k^2 f$.
c) On note $z = x + iy$, $\bar{z} = x - iy$ $(x, y \in \mathbb{R})$. Montrer que toute fonction polynomiale homogène de degré k sur $\mathbb{R}^2$, à valeurs complexes, est une combinaison linéaire d'expressions $z^a\bar{z}^b$ $(a, b \in \mathbb{N},\ a + b = k)$. On note $a - b = m$, on a : $z^a\bar{z}^b = \rho^k e^{im\theta}$. Donner une base de l'espace des polynômes homogènes de degré k.
d) Une fonction polynomiale f, de deux variables réelles (à valeurs réelles ou complexes) est dite harmonique si $\Delta f = 0$. Soit $k \in \mathbb{N}$. Quelle est la dimension (sur $\mathbb{R}$, respectivement $\mathbb{C}$) de l'espace des fonctions polynomiales homogènes harmoniques de degré k ? En trouver une base (on traitera les cas des valeurs rélles et complexes). Vérifier que les éléments u de cette base vérifient $\Delta_\theta u = -k^2 u$.

III.4.17 $**$ On considère maintenant le laplacien en dimension 3. On utilise des coordonnées cartésiennes (x, y, z) et des coordonnées sphériques (r, θ, φ). Un polynôme $f \in \mathbb{R}[X, Y, Z]$ est dit harmonique si $\Delta f = 0$.

1. a) On désigne par $\mathcal{F}_k \subset \mathbb{R}[X, Y, Z]$ le $\mathbb{R}$-espace vectoriel des polynômes homogènes de degré k. Quelle est sa dimension ?
   b) On désigne par $\mathcal{H}_k$ le $\mathbb{R}$-espace vectoriel des polynômes homogènes harmoniques de degré k. Soit $k \geqslant 2$. Montrer que $\Delta(\mathcal{F}_k) \subset \mathcal{F}_{k-2}$. En déduire que $\dim \mathcal{H}_k \geqslant 2k + 1$.
   c) On note $r := \sqrt{x^2 + y^2 + z^2}$. Soient $k, m \in \mathbb{N}$, avec $k \geqslant 2m$. Soit $g \in \mathcal{F}_{k-2m}$. Calculer $\Delta(r^{2m} g)$.
   d) Soient $k \geqslant 2$ et $f \in \mathcal{H}_k$. On suppose qu'il existe $g \in \mathcal{F}_{k-2}$ tel que $f = r^2 g$. Montrer que, pour tout $m \in \mathbb{N}^*$, tel que $2m \leqslant k$, f s'écrit $f = r^{2m} g_m$, avec $g_m \in \mathcal{F}_{k-2m}$. En déduire $f = 0$.
   e) Montrer que $\dim \mathcal{H}_k \leqslant \dim \mathcal{F}_k - \dim \mathcal{F}_{k-2}$. Calculer $\dim \mathcal{H}_k$.
2. a) On définit Δ_S par $\Delta_S = \frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}$.
   Montrer que, si $f \in \mathcal{F}_k$, on a $\Delta_S f = r^2\Delta f - k(k+1)f$.
   b) Montrer que toute fonction polynomiale homogène de degré k sur $\mathbb{R}^3$, à valeurs complexes, est une combinaison linéaire d'expressions :
   $$(x + iy)^a (x - iy)^b z^c \qquad \text{avec} \qquad a, b, c \in \mathbb{N} \quad \text{et} \quad a + b + c = k.$$
   Vérifier que : $(x + iy)^a (x - iy)^b z^c = r^k e^{i(a-b)\varphi} \sin^{a+b}\theta P(\cos\theta)$, où P est un polynôme. Quelle est la dimension complexe de l'espace $\mathcal{F}_{\mathbb{C},k}$ des fonctions polynomiales homogènes, à valeurs complexes, de degré k ? Que peut-on en conclure ?
   c) Soit $k \in \mathbb{N}$. Quelle est la dimension complexe de l'espace $\mathcal{H}_{\mathbb{C},k}$ des polynômes homogènes harmoniques, à valeurs complexes, de degré k ?
   d) Soit $P \in \mathbb{C}[T]$ un polynôme. On suppose que $f(x, y, z) = r^k e^{im\varphi} P(\cos\theta)$ définit un polynôme en (x, y, z). Montrer que ce polynôme est homogène. Montrer qu'il est harmonique si, et seulement si :
   $$(1 - t^2)P''(t) - 2tP'(t) + \Big(k(k+1) - \frac{m^2}{1 - t^2}\Big)P(t) = 0.$$

Trouver une base de $\mathcal{H}_k$.
Vérifier que les éléments u de cette base vérifient $\Delta_S u = -k(k+1)u$.

III.4.18 Soient E un espace préhilbertien réel et a un endomorphisme de E possédant un adjoint noté ici a^+. On note $N := a^+a$. On *suppose* que $[a, a^+] := aa^+ - a^+a = \mathrm{Id}$ et que l'endomorphisme N de E possède au moins une valeur propre λ. On note v un vecteur propre unitaire associé.
a) Montrer que λ est un nombre réel et que $\lambda \geqslant 0$.
Montrer que $\|a(v)\| = \sqrt{\lambda}$ et que $\|a^+(v)\| = \sqrt{\lambda+1}$.
b) Montrer que $[N, a] = -a$ et $[N, a^+] = a^+$. En déduire que $a(v)$ est, ou bien nul, ou bien un vecteur propre de N associé à la valeur propre $\lambda - 1$ et que $a^+(v)$ est un vecteur propre de N associé à la valeur propre $\lambda + 1$.
c) Montrer que, pour tout $n \in \mathbb{N}$, $a^n(v)$ est, ou bien nul, ou bien un vecteur propre de N associé à la valeur propre $\lambda - n$. Montrer qu'il existe $m \in \mathbb{N}$ tel que $a^m(v) \neq 0$ et $a^{m+1}(v) = 0$. Montrer que $\lambda = m$ et que 0 est une valeur propre de v. En déduire que tout $n \in \mathbb{N}$ est une valeur propre de N. Montrer que si l'on connaît un vecteur propre associé à *une* valeur propre n_0, on peut trouver une suite de vecteurs propres associés respectivement à *toutes* les valeurs propres $n \in \mathbb{N}$.
d) Donner un exemple de triplet (E, a, a^+) satisfaisant aux conditions de cet exercice.

III.4.19 Soit $(t, x) \mapsto \psi(t, x) := \frac{e^{-\frac{x}{1-x}t}}{1-x}$ la fonction génératrice des polynômes de Laguerre.

a) Montrer que $(1-x)^2 \frac{\partial \psi(t,x)}{\partial x} = (1 - x - t)\psi(t, x)$. Retrouver la formule de récurrence pour les polynômes de Laguerre.

b) Soit t donné dans $]-1, 1[$. En utilisant l'expression explicite du polynôme L_n, montrer que la série entière $\displaystyle\sum_{n\in\mathbb{N}} L'_n(t)\frac{x^n}{n!}$ a un rayon de convergence non nul.

c) Montrer que $(1-x)\frac{\partial \psi(t,x)}{\partial t} = -x\psi(t, x)$. En déduire, en utilisant la question b), une relation entre L'_n, L'_{n-1} et L_{n-1} pour $n \geqslant 1$.
Retrouver l'équation différentielle des polynômes de Laguerre.

III.4.20 Soit ψ la fonction génératrice des polynômes de Laguerre (*cf.* l'exercice précédent). Montrer que la fonction :

$$\psi^a(t, x) := (1-x)^{-a}\psi(t, x)\,(a > -1, x \in]-1, 1[, t \in \mathbb{R})$$

est la fonction génératrice des polynômes de Laguerre généralisés L_n^a.

III.4.21 $**$ Montrer la relation de récurrence des polynômes de Jacobi : proposition 39 de la page 582, en utilisant les indications données à la page 582.

III.4.22

1. Soit $n \in \mathbb{N}$. Montrer, en utilisant les résultats de l'exercice 6 de la page 571 la formule :

$$P_n^{(a,b)}(t) = \frac{\Gamma(a+n+1)}{n!\Gamma(a+b+n+1)} \sum_{k=0}^{n} \binom{n}{k} \frac{\Gamma(a+b+n+k+1)}{\Gamma(a+k+1)} \left(\frac{t-1}{2}\right)^k. \qquad (72)$$

2. Retrouver la formule (72) en utilisant la formule de l'assertion (iii) de la proposition 31 de la page 577 et le résultat de l'exercice III.4.13.
3. Soit $\lambda > -1/2$. On définit les polynômes de Gegenbauer $(C_n^\lambda)_{n\in\mathbb{N}}$ par :

$$C_n^\lambda := \frac{\Gamma(\lambda+1/2)}{\Gamma(2\lambda)} \frac{\Gamma(n+2\lambda)}{\Gamma(n+\lambda+1/2)} P_n^{(\lambda-1/2,\lambda-1/2)}.$$

Montrer les formules :

$$\forall n \in \mathbb{N}, \quad C_n^\lambda(t) = \sum_{h=0}^{[n/2]} (-1)^h \frac{\Gamma(n+\lambda-h)}{\Gamma(\lambda)h!(n-2h)!}(2t)^{n-2h}. \tag{73}$$

4. Montrer les formules :

$$\forall n \in \mathbb{N}^*, \quad T_n(t) = \frac{n}{2} \sum_{h=0}^{[n/2]} (-1)^h \frac{(n-h-1)!}{h!(n-2h)!}(2t)^{n-2h}$$

et

$$\forall n \in \mathbb{N}, \quad U_n(t) = \sum_{h=0}^{[n/2]} (-1)^h \frac{(n-h)!}{h!(n-2h)!}(2t)^{n-2h} = \sum_{h=0}^{[n/2]} (-1)^h \binom{n-h}{h}(2t)^{n-2h}$$

pour les polynômes de Tchebychev de première et deuxième espèce $\{T_n\}_{n\in\mathbb{N}^*}$ et $\{U_n\}_{n\in\mathbb{N}}$.

III.4.23 Soit $r \in \mathbb{R}_+^*$. On considère une série double réelle $\sum\limits_{h,k\in\mathbb{N}} a_{h,k}u^h v^k$. On suppose qu'elle converge absolument pour $|u|, |v| \leqslant r$.
a) Montrer qu'il existe une suite de polynômes homogènes de deux variables $(P_n)_{n\in\mathbb{N}}$ telle que, pour tous $u, v \in \mathbb{R}$ tels que $|u|, |v| \leqslant r$, la série $\sum\limits_{n\in\mathbb{N}} P_n(u,v)$ soit convergente et que :

$$\sum_{h=0}^{+\infty}\sum_{k=0}^{+\infty} a_{h,k}u^h v^k = \sum_{n=0}^{+\infty} P_n(u,v).$$

b) Soit $\lambda \in \mathbb{R}$. Montrer que, pour tout $u \in \mathbb{R}$ tel que $|u| \leqslant r$, la série entière $\sum\limits_{n\in\mathbb{N}} P_n(1,\lambda)u^n$ a un rayon de convergence strictement positif.
c) On suppose que, pour tous $u, v \in \mathbb{R}$ tels que $|u|, |v| \leqslant r$, $\sum\limits_{h=0}^{+\infty}\sum\limits_{k=0}^{+\infty} a_{h,k}u^h v^k = 0$. Montrer que, pour tout $n \in \mathbb{N}$, le polynôme P_n est nul, en déduire que, pour tous $h, k \in \mathbb{N}$, $a_{h,k} = 0$ (on pourra raisonner par l'absurde et se ramener à une série entière d'une variable).
d) Justifier l'identification des coefficients, page 605.

III.4.24 $**$ Le but de cet exercice est de montrer le théorème 64 de la page 611, dont on reprend les notations et les hypothèses. On rappelle que les polynômes q_n ($n \in \mathbb{N}$) sont respectivement solutions d'équations différentielles de la forme $(Qwy')' + \lambda_n wy = 0$ (*cf.* (20) de la page 566).

a) Montrer que la fonction f' admet le développement en polynômes orthogonaux :

$$\sum_{k=1}^{+\infty} (\, f' \mid q'_k \,)_{Qw} \, \frac{q'_k}{\|q'_k\|^2_{Qw}},$$

(aux sens respectifs de la norme $\| \; \|_{Qw}$ et du produit scalaire $(\; \mid \;)_{Qw}$).
b) Soient $\alpha < \alpha' < \beta' < \beta$. Montrer que :

$$\int_{\alpha'}^{\beta'} f'(t) q'_n(t) Q(t) w(t) \, \mathrm{d}t = \big[f(t) q'_n(t) Q(t) w(t) \big]_{\alpha'}^{\beta'} + \lambda_n \int_{\alpha'}^{\beta'} f(t) q_n(t) w(t) \, \mathrm{d}t.$$

En déduire que les limites $\lim\limits_{t \to \alpha} f(t) q'_n(t) Q(t) w(t)$ et $\lim\limits_{t \to \beta} f(t) q'_n(t) Q(t) w(t)$ existent. Montrer ensuite que ces limites sont nulles.
c) Comparer $(\, f' \mid q'_n \,)^2_{Qw}$ et $(\, f \mid q_n \,)^2_w$. En déduire que, pour tout $n \in \mathbb{N}$, $\lambda_n > 0$ et que la série $\sum\limits_{n \in \mathbb{N}} \lambda_n (\, f \mid q_n \,)^2_w$ est convergente.
d) Soit $t \in \,]\alpha, \beta[$. On suppose que la série $\sum\limits_{n \in \mathbb{N}} \dfrac{q_n^2(t)}{\lambda_n}$ est convergente. Montrer que :

$$\left| \sum_{n=0}^{+\infty} (\, f \mid q_n \,)_w q_n(t) \right| \leqslant \Big(\sum_{n=0}^{+\infty} \lambda_n (\, f \mid q_n \,)^2_w \Big)^{1/2} \Big(\sum_{n=0}^{+\infty} \frac{q_n^2(t)}{\lambda_n} \Big)^{1/2}.$$

e) On suppose que $\alpha < \alpha' < \beta' < \beta$. Montrer que la série : $\sum\limits_{n=0}^{+\infty} \frac{q_n^2(t)}{\lambda_n}$ est uniformément convergente sur $[\alpha', \beta']$ (on raisonnera cas par cas, en considérant les cas de Jacobi, de Laguerre généralisé et d'Hermite). Conclure.

III.4.25 On note $w(t) = \mathrm{e}^{-t^2}$.
On rappelle que l'on a défini, en relation avec les polynômes d'Hermite, des opérateurs $\partial := \frac{1}{\sqrt{2}} \frac{\mathrm{d}}{\mathrm{d}t}$ et $\partial^* := -\frac{1}{\sqrt{2}} \frac{\mathrm{d}}{\mathrm{d}t} + \sqrt{2} t$ sur l'espace des fonctions de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$.
Si f est continue sur $\mathbb{R}$, on note, quand l'intégrale existe, $\mathbf{E}(f) = \frac{1}{\sqrt{\pi}} \displaystyle\int_{-\infty}^{+\infty} f(t) \, \mathrm{e}^{-t^2} \, \mathrm{d}t$.
On note, comme d'habitude, $L_w := \{ f \in \mathcal{C}(\mathbb{R}) | \; \mathbf{E}(f^2) < +\infty \}$.
Soit φ une fonction de classe $\mathcal{C}^\infty$ sur $\mathbb{R}$, appartenant à L_w, ainsi que toutes ses dérivées.

a) Montrer que, pour tout $n \in \mathbb{N}$, $\mathbf{E}(\partial^n \varphi) = 2^{-n/2} \mathbf{E}(\varphi H_n)$. Montrer que la série :

$$\sum_{n \in \mathbb{N}} \mathbf{E}(\partial^n \varphi) \frac{H_n}{2^{n/2} n!}$$

converge vers φ dans l'espace préhilbertien L_w.

b) Montrer, en utilisant le résultat de l'exercice III.4.24, ci-dessus, que la série de fonctions $\sum\limits_{n \in \mathbb{N}} \mathbf{E}(\partial^n \varphi) \frac{H_n}{2^{n/2} n!}$ converge uniformément sur tout segment de $\mathbb{R}$.

c) Soit $x \in \mathbb{R}$ fixé. On pose $\varphi(t) := e^{2tx - x^2}$. Montrer que $\varphi \in \mathcal{C}^\infty(\mathbb{R})$, et appartient à L_w, ainsi que toutes ses dérivées. Calculer $\partial^n \varphi$, puis $\mathbf{E}(\partial^n \varphi)$. Que peut-on en déduire ?

Exercices sur la section 4

III.4.26 $**$ Le but de l'exercice est de prouver la proposition 67 de la page 620. Nous reprenons les notations de cet énoncé.

a) Montrer que :

$$\sqrt{D_n D_{n-1}}\, q_n(t) = \int_{(\alpha,\beta)^n} \begin{vmatrix} \lambda_1^0 & \lambda_1^1 & \dots & \lambda_1^n \\ \lambda_2^1 & \lambda_2^2 & \dots & \lambda_2^{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_n^{n-1} & \lambda_n^n & \dots & \lambda_n^{2n-1} \\ 1 & t & \dots & t^n \end{vmatrix} w(\lambda_1)\dots w(\lambda_n)\ \mathrm{d}\lambda_1\dots\mathrm{d}\lambda_n \tag{74}$$

$$= \int_{(\alpha,\beta)^n} \lambda_1^0\lambda_2^1\dots\lambda_n^{n-1} \begin{vmatrix} 1 & \lambda_1 & \dots & \lambda_1^n \\ 1 & \lambda_2 & \dots & \lambda_2^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_n & \dots & \lambda_n^n \\ 1 & t & \dots & t^n \end{vmatrix} w(\lambda_1)\dots w(\lambda_n)\ \mathrm{d}\lambda_1\dots\mathrm{d}\lambda_n. \tag{75}$$

b) On remarque que, dans les formules ci-dessus, on peut remplacer $\lambda_1,\dots,\lambda_n$ par une permutation arbitraire $\lambda_{\sigma(1)},\dots,\lambda_{\sigma(n)}$ sans changer le résultat. En déduire que :

$$\sqrt{D_n D_{n-1}}\, q_n(t) = \frac{1}{n!}\int_{(\alpha,\beta)^n} \Big[\sum_{\sigma\in\mathfrak{S}_n} \varepsilon(\sigma)\lambda_{\sigma(1)}^0,\dots,\lambda_{\sigma(n)}^{n-1}\Big] \prod_{i=1}^n (t-\lambda_i)$$
$$\prod_{1\leqslant h<k\leqslant n} (\lambda_k-\lambda_h) w(\lambda_1)\dots w(\lambda_n)\ \mathrm{d}\lambda_1\dots\mathrm{d}\lambda_n.$$

On a noté $\mathfrak{S}_n$ le groupe des permutations de n éléments et $\varepsilon(\sigma)$ la signature de $\sigma\in\mathfrak{S}_n$ (on utilisera le calcul d'un déterminant de Vandermonde).

c) Montrer que les polynômes orthonormés (q_n) associés au poids w admettent la représentation intégrale :

$$q_n(t) = \frac{1}{n!\sqrt{D_n D_{n-1}}}\int_{(\alpha,\beta)^n} \prod_{i=1}^n (t-\lambda_i) \prod_{1\leqslant h<k\leqslant n} (\lambda_h-\lambda_k)^2 w(\lambda_1)\dots w(\lambda_n)\ \mathrm{d}\lambda_1\dots\mathrm{d}\lambda_n$$

(on utilisera à nouveau le calcul d'un déterminant de Vandermonde).

d) Montrer que les déterminants de Hankel admettent la représentation intégrale

$$D_{n-1} = \frac{1}{n!}\int_{(\alpha,\beta)^n} \prod_{1\leqslant h<k\leqslant n} (\lambda_h-\lambda_k)^2 w(\lambda_1)\dots w(\lambda_n)\ \mathrm{d}\lambda_1\dots\mathrm{d}\lambda_n$$

III.4.27 Montrer la formule de la proposition 65 de la page 618 par une vérification directe.

III.4.28 Le but de cet exercice est de montrer le théorème d'inversion de la transformée de Stieltjes (théorème 82 de la page 640), quand μ est la mesure associée à un poids w sur un intervalle $]\alpha,\beta[$. On note pour simplifier $S(z) := S(\mu)(z)$. On note F la fonction de répartition de w (elle est monotone sur $\mathbb{R}$).
Soient $c, d \in]\alpha,\beta[$, avec $c < d$, et $y > 0$.
a) Montrer que :

$$\int_{c+iy}^{d+iy} S(z)\,\mathrm{d}z = \int_\alpha^\beta \Big(\int_{c+iy}^{d+iy} \frac{\mathrm{d}z}{z-u}\Big) w(u)\,\mathrm{d}u$$

(les chemins d'intégration étant parallèles à l'axe des abcisses) .
b) On pose $z = x + iy$ $(x, y \in \mathbb{R})$. Montrer que :

$$\int_c^d \operatorname{Im} S(z)\,\mathrm{d}x = \int_\alpha^\beta \Big(\arctan\frac{u-d}{y} - \arctan\frac{u-c}{y}\Big) w(u)\,\mathrm{d}u.$$

c) Montrer que, lorsque y tend vers 0, $\displaystyle\int_\alpha^\beta \arctan\frac{u-d}{y} w(u)\,\mathrm{d}u$ tend vers $-\pi F(d) + \dfrac{\pi}{2}\mu_0$ et que $\displaystyle -\frac{1}{\pi}\int_c^d \operatorname{Im} S(z)\,\mathrm{d}x$ tend vers $F(d) - F(c) = \displaystyle\int_c^d w(u)\,\mathrm{d}u$.
d) On suppose maintenant que $c, d \in \mathbb{R}$, $c < d$. Calculer la limite de $\displaystyle\frac{1}{\pi}\int_c^d \operatorname{Im} S(z)\,\mathrm{d}x$, quand y tend vers zéro.

III.4.29 On note, pour tout $\varepsilon > 0$, et tout $t \in \mathbb{R}$, $p_\varepsilon(t) = \frac{1}{\pi}\frac{\varepsilon}{t^2+\varepsilon^2}\cdot$ Soient w un poids sur un intervalle $]\alpha,\beta[$ et μ la mesure associée, considérée comme mesure sur $\mathbb{R}$ $(w(t)\,\mathrm{d}t = \mathbf{1}_{]\alpha,\beta[} w(t)\,\mathrm{d}t)$.

a) Montrer que $\displaystyle\int_{-\infty}^{+\infty} p_\varepsilon(t)\,\mathrm{d}t = 1$.

b) Soit f une fonction continue bornée sur $\mathbb{R}$. On pose $\nu_\varepsilon(f) := \displaystyle\int_{-\infty}^{+\infty} f(t) p_\varepsilon(t)\,\mathrm{d}t$. Calculer la limite de $\nu_\varepsilon(f)$, quand ε tend vers 0.

c) Soit f une fonction continue bornée sur $\mathbb{R}$. Montrer que, pour tout $\varepsilon > 0$:

$$-\frac{1}{\pi}\int_{-\infty}^{+\infty} \operatorname{Im} S(\mu)(t+i\varepsilon) f(t)\,\mathrm{d}t = \int_{-\infty}^{+\infty} f_\varepsilon(t) w(t)\,\mathrm{d}t.$$

avec $f_\varepsilon(t) := \displaystyle\int_{-\infty}^{+\infty} p_\varepsilon(t-s) f(s)\,\mathrm{d}s$.

d) Montrer que, pour tout $t \in \mathbb{R}$, $f_\varepsilon(t)$ tend vers $f(t)$, quand ε tend vers 0. En déduire que

$$\int_{-\infty}^{+\infty} f(t) w(t)\,\mathrm{d}t = \int_\alpha^\beta f(t) w(t)\,\mathrm{d}t = -\lim_{\varepsilon\to 0}\frac{1}{\pi}\int_{-\infty}^{+\infty} \operatorname{Im} S(\mu)(t+i\varepsilon) f(t)\,\mathrm{d}t.$$

Comparer avec l'exercice précédent.

III.4.30 Montrer que dans la proposition 80 de la page 639 la convergence est uniforme sur tout fermé de $\mathbb{C} \setminus [\alpha,\beta]$. En déduire que dans le théorème 87 de la page 644, la convergence de la suite de (fonctions) fractions rationnelles (R_n/S_n) est uniforme sur tout fermé de $\mathbb{C} \setminus [\alpha,\beta]$.

On utilise dans les cinq exercices suivants les notations de la section 4.2.3.

III.4.31 On définit, sur $\mathbb{R}_-^*$, les fonctions $x \mapsto I_n(x) := \displaystyle\int_{-x}^{+\infty} \frac{\mathrm{e}^{-u}}{u^{n+1}}\,\mathrm{d}u$. On note :

$$f(x) := -\mathrm{e}^{-x} \int_{-x}^{+\infty} \frac{\mathrm{e}^{-u}}{u}\,\mathrm{d}u = -\mathrm{e}^{-x} I_0(x)$$

et l'on définit les fonctions polynomiales en $1/x$: $f_0(x) := 0$ et, pour $n \in \mathbb{N}^*$:

$$x \mapsto f_n(x) = \frac{1}{x} + \frac{1!}{x^2} + \cdots + \frac{(n-1)!}{x^n}.$$

a) Montrer que, pour $x < 0$, $f = f_n + (-1)^{n+1} n!\,\mathrm{e}^{-x} I_n$.

b) Montrer que, pour tout $x < 0$, $0 < I_n(x) < \frac{\mathrm{e}^x}{|x|^{n+1}}$. En déduire, pour la fonction f, l'existence de développements limités à un ordre arbitraire au voisinage de $-\infty$.

c) Donner un encadrement de $f(x)$ en utilisant les $f_n(x)$ (on distinguera les cas $n = 2p$ et $n = 2p+1$). Pour $x < 0$ fixé, comment peut-on optimiser cet encadrement ? Donner une valeur approchée de $f(-10)$.

III.4.32

Soit $n \in \mathbb{N}$. On définit, sur $\mathbb{R}_-^*$ les fonctions $x \mapsto J_n(x) := n! \displaystyle\int_{-x}^{+\infty} \frac{(u+x)^n \mathrm{e}^{-u}}{u^{n+1}}\,\mathrm{d}u$.

a) Montrer que $J_n = n! \sum\limits_{k=0}^{n} \binom{n}{k} x^k I_k$.

b) Montrer que, pour tout $n \in \mathbb{N}^*$, $xJ_n' = nJ_n - n^2 J_{n-1}$ et $J_n' = n(-J_{n-1} + J_{n-1}')$, puis que la suite $(J_n)_{n\in\mathbb{N}}$ vérifie la relation de récurrence des polynômes de Laguerre (*cf.* (54), page 600) :

$$J_{n+1} + (x - 2n - 1)J_n + n^2 J_{n-1} = 0.$$

III.4.33 Montrer, en utilisant l'exercice précédent, que, pour tout $n \in \mathbb{N}^*$, J_n est solution, sur $]-\infty, 0[$, de l'équation différentielle de Laguerre $xy'' + (1-x)y' + ny = 0$.

III.4.34 Soit $n \in \mathbb{N}^*$.

a) Montrer que $J_n = L_n I_0 - h_n \mathrm{e}^x$, où L_n est le $n+1^{\text{ème}}$ polynôme de Laguerre et h_n est un polynôme déterminé de manière unique. Montrer que le polynôme h_n est de degré $n-1$, que le coefficient de son terme de plus haut degré est $(-1)^{n+1}$, et que la suite (h_n) vérifie la relation de récurrence des polynômes de Laguerre : $h_{n+1} + (x - 2n - 1)h_n + n^2 h_{n-1} = 0$ $(n \geqslant 2)$;

b) Montrer que le polynôme h_n, défini en a), est la *partie entière* de la fraction rationnelle $-L_n f_n$, puis que la fraction rationnelle $-\frac{h_n}{L_n}$ est la $n^{\text{ème}}$ réduite de la fraction continue associée à la mesure de Laguerre μ.

III.4.35 Le but de cet exercice est de montrer, en utilisant les exercices ci-dessus, le théorème 89 de la page 646.

a) Montrer que, pour tout $x < 0$ et tout $n \in \mathbb{N}$, $0 < J_n(x) < -n!\dfrac{\mathrm{e}^x}{x}$ et $L_n(x) > -nn!x$.

b) Soit $x < 0$ fixé. Quelle est la limite de $\frac{J_n(x)}{L_n(x)}$, quand n tend vers $+\infty$?

En déduire que $\lim\limits_{n\to+\infty} \frac{h_n(x)}{L_n(x)} = -f(x)$. Montrer que $|f(x) + \frac{h_n(x)}{L_n(x)}| < \frac{n!}{|x|^{n+1}}$

c) Calculer $h_3(-10)/L_3(-10)$. Comparer avec l'exercice III.4.31, question c).

Probabilités

Partie IV

Probabilités et statistiques ont de très nombreuses applications : en biologie, médecine, physique, sciences politiques, finance… Elles restent malheureusement mal comprises. Qui sait définir l'expression « espérance de vie » prononcée, en aveugle, presque chaque jour dans les médias ? Qui comprend l'importance de l'échantillonnage dans les sondages ?

Ces deux disciplines sont nées au même moment : la statistique vers 1650, quand sont publiées des tables de mortalité (John Graunt, puis Halley en fin du siècle), théorisées par Euler (travaillant par exemple sur un problème de pensions des veuves) ; le calcul des probabilités (finies) en 1654 avec Pascal à Paris et Fermat à Toulouse (succédant à quelques essais : Pacioli au XVe siècle, Cardan au XVIe siècle, Galilée). Il s'agit au début d'une tentative d'approche scientifique du hasard présent dans les jeux, dont ceux de dés (az-zahr : dé à jouer en arabe).

Les probabilités finies ont été codifiées par Christian Huygens et Jacques Bernoulli. Les probabilités continues furent notamment l'œuvre de De Moivre, Lagrange, Laplace, Poisson, couronnée par Lebesgue (comme application de sa théorie de l'intégration de 1901), et aussi Borel et Lévy. Kolmogorov les a formalisées de manière décisive (1933). Le travail de Wiener sur le mouvement brownien (1923) est aussi une clé pour le développement moderne du sujet.

Peu à peu, les statistiques ont fait place à « la statistique », en partie autonome des probabilités, même si les deux ont toujours, comme à leur naissance, un aspect appliqué, important et performant. Leur spectre est maintenant très large avec une forte composante théorique. Ce domaine des mathématiques est lié à d'autres comme l'analyse (théorie de la mesure, équations aux dérivées partielles) ou la théorie des nombres. Des outils théoriques sophistiqués sont apparus dans la seconde partie du XXe siècle : équations différentielles stochastiques (Ito, Döblin), calcul de Malliavin. Ils ont d'importantes retombées applicatives, en finance par exemple (formule de Black-Scholes pour le prix des options).

Après les apprentissages dans un cadre essentiellement fini du cours de [L1], cette partie aborde les vrais concepts généraux, autour de la notion de variable aléatoire (discrète ou à densité), avec une présentation des principales lois de probabilités. Un dernier module initie à la vaste théorie des théorèmes limites, aux applications sans nombre.

Notions fondamentales sur les probabilités

IV.1

Ce module pose les bases de la théorie mathématique des probabilités. Il montre qu'une analyse mathématique et rigoureuse peut être appliquée à des phénomènes aléatoires. Ce module reprend certains des résultats du module de probabilités-statistique du livre [L1], consacré aux probabilités finies, mais il les présente dans un cadre plus général, puisque l'on considère ici le cas général où le nombre de résultats possibles peut être fini, infini dénombrable, ou infini non dénombrable.

1 Ensemble fondamental et événements

Lorsque le hasard entre en jeu (par exemple, lorsque l'on joue au loto), on ne peut pas prédire le résultat avec déterminisme. Cependant, une analyse est encore possible, dont le but est d'étudier la probabilité de réalisation d'un certain événement A (par exemple, « j'aurai les cinq bons numéros et le numéro chance au loto »). La probabilité de l'événement A est un nombre qui quantifie la vraisemblance de cet événement (par exemple, $0{,}000005\ \%\ \ldots$). L'étude mathématique prend en compte le fait que la réalisation de l'événement A n'est pas assurée, mais dépend du résultat d'une expérience aléatoire (par exemple, le tirage sans remise de 5 numéros parmi 49 et le tirage d'un numéro chance parmi 10).

1.1 Ensemble fondamental

L'analyse d'une expérience aléatoire commence par le recensement de l'ensemble de tous les résultats possibles de l'expérience. Cet ensemble non vide noté Ω est l'ensemble fondamental ou l'univers des possibles ; ses éléments ω sont appelés réalisations. L'ensemble fondamental Ω peut-être fini, infini dénombrable ou infini non dénombrable.

Exemple 1. **Jeu de pile ou face.** On lance une pièce. Deux résultats sont possibles, Pile et Face, de sorte que $\Omega = \{P, F\}$.

Exemple 2. **Lancer de deux dés.** On lance deux dés numérotés de 1 à 6. Le résultat est un couple (ω_1, ω_2) où $\omega_j \in [\![1, 6]\!]$ désigne le résultat du $j^{ème}$ dé. On a donc ici $\Omega = [\![1, 6]\!]^2$.

Exemple 3. **Battage de cartes.** On bat un jeu de 52 cartes numérotées de 1 à 52. Le nouvel ordre dans lequel sont rangées les cartes est caractérisé par une permutation ω des 52 premiers entiers, où $\omega(i)$ est la position de la $i^{ème}$ carte du paquet après battage. Ici l'ensemble fondamental Ω est donc l'ensemble des permutations de $[\![1, 52]\!]$.

Exemple 4. On lance une pièce jusqu'à obtenir Pile. L'espace fondamental est alors :

$$\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\} \cup \{\omega_\infty\} \ ,$$

où ω_i représente le résultat $FF \cdots FP$ où les $i-1$ premiers lancers donnent Face et le $i^{ème}$ Pile. ω_∞ est l'événement $FFFF \cdots$ où tous les lancers donnent des Faces.

Exemple 5. On considère la durée de vie d'un appareil.
L'espace fondamental est alors $\Omega = \mathbb{R}_+$, où $\omega \in \mathbb{R}_+$ représente l'instant où l'appareil tombe en panne.

1.2 La notion d'événement

Un événement A est défini intuitivement comme un fait qui peut arriver ou non, selon le résultat de l'expérience.
Dans l'exemple 3, on peut considérer les événements suivants :
a) la carte numéro i est la première du paquet après battage,
b) la carte numéro i est à sa place (i.e la $i^{ème}$ du paquet),
c) aucune carte n'est à sa place.
Dans l'exemple 5, on peut considérer l'événement :
a) l'appareil tombe en panne pendant la période de garantie.
Ces événements correspondent à un sous-ensemble, également noté A, de Ω.
Dans l'exemple 3,
a) $A_i := \{\omega \in \Omega \mid \omega(i) = 1\}$,
b) $B_i := \{\omega \in \Omega \mid \omega(i) = i\}$,
c) $C := \bigcap_{i \in [\![1,52]\!]} C_i$, où $C_i := \{\omega \in \Omega \mid \omega(i) \neq i\}$ est le complémentaire de B_i.
Dans l'exemple 5,
a) $A := [0, a]$ où $a > 0$ est la durée de la garantie.
Pour résumer, si ω est une réalisation et si $\omega \in A$ on dit que l'événement A s'est réalisé. L'ensemble vide $\varnothing$ est appelé événement impossible, car quelle que soit la réalisation ω, l'événement impossible ne se réalise pas ($\omega \notin \varnothing$). L'ensemble fondamental Ω est appelé événement certain, car quelle que soit la réalisation ω, l'événement certain se réalise ($\omega \in \Omega$).
Nous identifierons dorénavant événement et sous-ensemble de Ω lui correspondant. On utilisera simultanément le langage de la théorie des ensembles et celui des probabilités. Le tableau IV.1.1 est un dictionnaire qui donne la correspondance entre les notions les plus fréquemment utilisées. On rappelle également quelques relations usuelles sur les ensembles : soient I un ensemble d'indices, fini ou infini, $(A_i)_{i \in I}$ une famille de sous-ensembles de Ω

Théorie des ensembles	**Probabilités**
sous-ensemble A	événement A
ensemble vide $\varnothing$	événement impossible $\varnothing$
ensemble fondamental Ω	événement certain Ω
$A \subset B$	A entraîne B
intersection $A \cap B$	A et B
union $A \cup B$	A ou B
$A^c =$ complémentaire de A	$A^c =$ contraire de A
A et B sont disjoints, $A \cap B = \varnothing$	A et B sont des événements incompatibles
$(A_i)_{i\in I}$ est une partition de Ω, i.e. $A_i \cap A_j = \varnothing \; \forall i \neq j$ et $\bigcup_{i\in I} A_i = \Omega$	$(A_i)_{i\in I}$ est un système complet d'événements

TABLE IV.1.1 – Vocabulaire de la théorie des ensembles et des probabilités.

et $B \subset \Omega$. On a :

$$\text{Formules de De Morgan} \begin{cases} \left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c \,, \\ \left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c \,, \end{cases}$$

$$\text{Distributivité} \begin{cases} B \bigcap \left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} (B \cap A_i) \;, \\ B \bigcup \left(\bigcap_{i\in I} A_i\right) = \bigcap_{i\in I} (B \cup A_i) \;. \end{cases}$$

1.3 La notion de tribu

Pour permettre l'analyse d'une expérience aléatoire on doit considérer *a priori* une classe $\mathcal{A}$ d'événements. Dans les cas élémentaires comme dans les exemples 1-4 de la page 665 (ensemble fondamental fini ou dénombrable), cette classe $\mathcal{A}$ est prise la plupart du temps égale à l'ensemble de tous les sous-ensembles de Ω, noté $\mathcal{P}(\Omega)$, i.e. tout sous-ensemble peut être considéré comme événement. En particulier, tout A à un élément, $A = \{\omega\}$, est appelé événement élémentaire. Cette classe $\mathcal{P}(\Omega)$ satisfait les propriétés élémentaires suivantes : elle contient Ω et elle est stable par complémentation et par réunion finie ou dénombrable. Dans le cas d'un ensemble fondamental Ω infini non dénombrable, on ne va pas en général considérer toutes les parties de Ω comme des événements, car l'ensemble des parties de Ω est très gros, et contient des sous-ensembles qui ne nous intéressent pas forcément. Ainsi, dans l'exemple 5 de la page 666, on est seulement intéressé par des événements du type : « la durée de vie de l'appareil tombe dans un intervalle de temps donné ». On travaille alors avec une classe $\mathcal{A}$ d'événements dont on réclame qu'elle contienne les sous-ensembles qui nous intéressent, et aussi qu'elle satisfasse les propriétés élémentaires que l'on vient d'énumérer. Une telle classe $\mathcal{A} \subset \mathcal{P}(\Omega)$ est appelée tribu sur l'ensemble fondamental Ω.

Définition 1. Une famille $\mathcal{A}$ de parties de Ω est une *tribu* si :
(i) $\Omega \in \mathcal{A}$.
(ii) $A \in \mathcal{A} \implies A^c \in \mathcal{A}$.
(iii) $(A_i)_{i\in\mathbb{N}} \in \mathcal{A}^{\mathbb{N}} \implies \bigcup\limits_{i\in\mathbb{N}} A_i \in \mathcal{A}$.

Notons que $\varnothing = \Omega^c$ appartient à $\mathcal{A}$. De plus, puisque $\varnothing \in \mathcal{A}$, la propriété (iii) de stabilité par union dénombrable entraîne la stabilité par réunion finie. Elle entraîne aussi la stabilité par intersection dénombrable (d'après l'égalité $\left(\bigcap\limits_i A_i\right)^c = \bigcup\limits_i A_i^c$) et donc celle par intersection finie. La notion de tribu est centrale en théorie de la mesure, elle est évoquée dans le module d'intégration, en particulier dans la section 3 et la définition 30 de la page 322.

Exemples. (i) La plus petite tribu sur Ω est $\mathcal{A} := \{\varnothing, \Omega\}$.
(ii) Si $A \subset \Omega$, alors $\mathcal{A} := \{\varnothing, A, A^c, \Omega\}$ est une tribu.
(iii) L'ensemble $\mathcal{P}(\Omega)$ des parties de Ω est une tribu sur Ω. C'est la tribu naturelle pour les ensembles Ω finis ou dénombrables. C'est en fait toujours dans ce cadre que l'on se placera dans le cas Ω fini ou dénombrable.
(iv) Si $\Omega := \mathbb{R}$, alors la tribu de Borel $\mathcal{B}(\mathbb{R})$ est la plus petite tribu qui contienne les intervalles de $\mathbb{R}$. Son existence est établie dans la définition 31 de la page 322. On peut démontrer que $\mathcal{B}(\mathbb{R}) \subsetneq \mathcal{P}(\mathbb{R})$. Toutefois les seuls événements que nous manipulerons seront les intervalles de $\mathbb{R}$ (attention, $]a, +\infty[$ et $]-\infty, +\infty[$ sont des intervalles) ou des réunions finies ou dénombrables d'intervalles. Donc on prendra toujours comme tribu associée à $\mathbb{R}$ la tribu borélienne.
(v) Si $\Omega := I$ où $I := [a, b]$ est un intervalle de $\mathbb{R}$, la tribu de Borel $\mathcal{B}(I)$ est la plus petite tribu qui contienne les intervalles de I. On montre que $\mathcal{B}(I) = \{I \cap A \mid A \in \mathcal{B}(\mathbb{R})\}$.
(vi) Si $\Omega := \mathbb{R}^d$, alors la tribu de Borel $\mathcal{B}(\mathbb{R}^d)$ est la plus petite tribu qui contienne les pavés de $\mathbb{R}^d$ (i.e. les ensembles du type $]a_1, b_1[\times \cdots \times]a_d, b_d[$ où $a_i \in \mathbb{R} \cup \{-\infty\}$ et $b_i \in \mathbb{R} \cup \{+\infty\}$).
(vii) D'une manière générale, pour un ensemble d'indices I et pour une famille $(A_i)_{i\in I}$ de parties de Ω, il est impossible d'expliciter la tribu $\mathcal{A}$ engendrée par la famille $(A_i)_{i\in I}$, c'est-à-dire la plus petite tribu contenant les A_i, sauf dans le cas d'une famille finie, proposé dans l'exercice IV.1.2. Tout ce que l'on peut dire au niveau de ce cours, c'est que cette tribu $\mathcal{A}$ existe et est unique et que c'est une famille de parties de Ω obtenues, à partir de la famille $(A_i)_{i\in I}$, par l'application qui à une telle famille associe celles des réunions dénombrables de ses éléments et de leurs complémentaires. Mais, à la différence d'une construction par récurrence classique, il faut ici user d'une récurrence *transfinie* faisant intervenir la théorie des nombres ordinaux.

2 Probabilités

Pour justifier notre définition d'une probabilité, nous allons faire appel à notre intuition et discuter la signification usuelle de ce qu'est la probabilité d'un événement. Considérons un événement A pouvant se produire lors d'une certaine expérience aléatoire (par exemple, $A :=$« obtenir Pile » lors du lancer d'une pièce). Supposons que l'on puisse ré-

péter un grand nombre N de fois cette expérience aléatoire. Notons $N(A)$ le nombre de fois où l'événement A se produit. La fréquence empirique d'apparition de A, $N(A)/N$, est elle-même aléatoire. Mais notre expérience courante tend à nous faire penser que, lorsque le nombre N de répétitions de l'expérience tend vers l'infini, le rapport $N(A)/N$ se stabilise autour d'une valeur limite déterministe (on peut penser que $N(A)/N$ tend vers $1/2$ dans le cas où A est l'événement « obtenir Pile » lors du lancer d'une pièce). Cette limite est notre idée intuitive de la probabilité $\mathbb{P}(A)$ de l'événement A. Dans cette vision intuitive, si A est l'événement certain $A = \Omega$, alors $N(\Omega) = N$ et donc $\mathbb{P}(\Omega) = 1$. De plus, si les événements A et B sont incompatibles, $A \cap B = \varnothing$, on a $N(A \cup B) = N(A) + N(B)$ et donc $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$: une probabilité doit être une fonction additive d'ensembles. La définition mathématique des probabilités reprend les propriétés évoquées ci-dessus, et les pose comme axiomes :

Définition 2. Soient Ω un ensemble non vide et $\mathcal{A}$ une tribu sur Ω. Une probabilité est une fonction $\mathbb{P} : \mathcal{A} \to [0,1]$ telle que :
(i) $\mathbb{P}(\Omega) = 1$.
(ii) Si $(A_i)_{i\in\mathbb{N}}$ est une famille dénombrable d'événements incompatibles (c'est-à-dire vérifiant $A_i \cap A_j = \varnothing \ \forall i \neq j$), alors :

$$\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)\,.$$

La propriété (ii) est appelée σ-additivité, ou additivité dénombrable. Le triplet $(\Omega, \mathcal{A}, \mathbb{P})$ est appelé espace de probabilité.

2.1 Propriétés élémentaires d'une probabilité

On énonce dans la proposition quelques propriétés simples mais fondamentales d'une probabilité.

Proposition 1. Soit $\mathbb{P}$ une probabilité sur $(\Omega, \mathcal{A})$.
1. $\mathbb{P}(\varnothing) = 0$.
2. Si $n \geqslant 1$ et $(A_i)_{i\in[\![0,n]\!]}$ est une famille finie d'événements incompatibles, alors :

$$\mathbb{P}\left(\bigcup_{i=0}^{n} A_i\right) = \sum_{i=0}^{n} \mathbb{P}(A_i)\,.$$

3. Si $A \in \mathcal{A}$, alors $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.
4. Si $A_1, A_2 \in \mathcal{A}$ et $A_1 \subset A_2$, alors $\mathbb{P}(A_1) \leqslant \mathbb{P}(A_2)$.
5. Si $A_1, A_2 \in \mathcal{A}$, alors $\mathbb{P}(A_1 \cup A_2) + \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$.
6. Si $(A_i)_{i\in\mathbb{N}}$ est une famille d'événements, alors $\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) \leqslant \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)$.

Démonstration. Notons $\alpha := \mathbb{P}(\varnothing)$. D'une part, en prenant $A_i := \varnothing$ pour tout $i \in \mathbb{N}$, on a bien une famille de parties deux à deux disjointes car $\varnothing = \varnothing \cap \varnothing$. De plus, leur

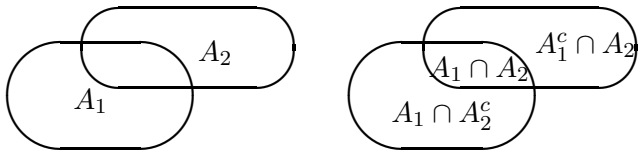


FIGURE IV.1.1 – Décomposition de $A_1 \cup A_2$ en une union disjointe de trois événements $A_1 \cap A_2^c$, $A_1 \cap A_2$ et $A_1^c \cap A_2$.

union donne $\varnothing$, donc la propriété (ii) de σ-additivité donne $\mathbb{P}(\varnothing) = \sum_{i\in\mathbb{N}} \mathbb{P}(\varnothing)$, c'est-à-dire $\alpha = \sum_{i\in\mathbb{N}} \alpha$, ce qui montre que $\alpha = 0$ et prouve la propriété 1.

La propriété (ii) de la définition implique la propriété 2 d'additivité finie : si l'on a affaire à une famille finie de $n+1$ parties deux à deux disjointes $(A_i)_{i\in[\![0,n]\!]}$, alors en complétant cette famille par des ensembles vides $A_i = \varnothing$, $i \geqslant n+1$, on obtient une famille de parties deux à deux disjointes, donc :

$$\mathbb{P}\Big(\bigcup_{i=0}^{n} A_i\Big) = \mathbb{P}\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i) = \sum_{i=0}^{n} \mathbb{P}(A_i) + \sum_{i=n+1}^{+\infty} \mathbb{P}(\varnothing) = \sum_{i=0}^{n} \mathbb{P}(A_i)\,.$$

La propriété 3 résulte de la décomposition $\Omega = A \cup A^c$ valable pour tout $A \in \mathcal{A}$. En appliquant la propriété d'additivité finie, on obtient $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.

La propriété 4 résulte de la décomposition $A_2 = A_1 \cup (A_2 \cap A_1^c)$ valable si $A_1 \subset A_2$. Encore une fois, A_1 et $A_2 \cap A_1^c$ sont disjoints, donc on obtient $\mathbb{P}(A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \cap A_1^c)$. Comme $\mathbb{P}$ est une probabilité, le dernier terme est un réel positif ou nul ce qui donne $\mathbb{P}(A_2) \geqslant \mathbb{P}(A_1)$.

La propriété 5 s'obtient en considérant les événements $A_1 \cap A_2^c$, $A_1 \cap A_2$ et $A_1^c \cap A_2$. Ces trois événements sont deux à deux disjoints et satisfont :

$$A_1 \cup A_2 = (A_1 \cap A_2^c) \cup (A_1 \cap A_2) \cup (A_1^c \cap A_2)\,.$$

On peut s'aider d'une dessin pour s'en rendre compte (voir figure IV.1.1). Donc :

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1 \cap A_2^c) + \mathbb{P}(A_1 \cap A_2) + \mathbb{P}(A_1^c \cap A_2)\,. \tag{1}$$

D'autre part, les événements $A_1^c \cap A_2$ et $A_1 \cap A_2$ sont deux à deux disjoints et leur réunion est A_2. Donc :

$$\mathbb{P}(A_2) = \mathbb{P}(A_1^c \cap A_2) + \mathbb{P}(A_1 \cap A_2)\,.$$

De même, en échangeant les rôles de A_1 et A_2 :

$$\mathbb{P}(A_1) = \mathbb{P}(A_2^c \cap A_1) + \mathbb{P}(A_1 \cap A_2)\,.$$

En substituant les deux dernières égalités dans (1), on obtient :

$$\mathbb{P}(A_1 \cup A_2) + \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)\,.$$

Démontrons la propriété 6. Pour tout $\omega \in \bigcup_{n\in\mathbb{N}} A_n$ désignons par $N(\omega)$ le plus petit indice n

pour lequel $\omega \in A_n$ et définissons :

$$B_n := \{\omega \mid N(\omega) = n\} = A_n \cap \left(\bigcup_{m=0}^{n-1} A_m\right)^c .$$

Comme les A_m sont des événements de la tribu $\mathcal{A}$ pour $m \in [\![0, n]\!]$, on a aussi $B_n \in \mathcal{A}$. De plus, les B_n sont deux à deux disjoints, et leur union forme $\bigcup_{n\in\mathbb{N}} A_n$. Donc les B_n forment une partition de $\bigcup_{n\in\mathbb{N}} A_n$. D'après la σ-additivité de $\mathbb{P}$ on a :

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(B_n) .$$

On applique alors la propriété 4 en utilisant le fait que $B_n \subset A_n$. On obtient ainsi le résultat cherché. ■

Une probabilité $\mathbb{P}$ satisfait aussi une certaine propriété de monotonie. On commence par préciser ce que l'on entend par suite croissante pour une suite d'événements. On dit qu'une suite d'événements $(A_n)_{n\in\mathbb{N}}$ est croissante [respectivement décroissante] si $n \leqslant m \Rightarrow A_n \subset A_m$ [respectivement $n \geqslant m \Rightarrow A_n \subset A_m$].
Pour une suite croissante [respectivement décroissante], $\lim_{n\to+\infty} A_n$ est définie par $\bigcup_{n\in\mathbb{N}} A_n$ [respectivement $\bigcap_{n\in\mathbb{N}} A_n$].

Proposition 2.
Si $(A_n)_{n\in\mathbb{N}}$ est une suite croissante [respectivement décroissante] d'événements, alors :

$$\mathbb{P}(\lim_{n\to+\infty} A_n) = \lim_{n\to+\infty} \mathbb{P}(A_n) .$$

Démonstration.
Pour démontrer la proposition dans le cas croissant introduisons les événements $B_0 := A_0$ et $B_n := A_n \cap A_{n-1}^c$ qui sont deux à deux disjoints et tels que $A_n = \bigcup_{m=0}^{n} B_m$. On a donc :

$$\mathbb{P}(A_n) = \sum_{m=0}^{n} \mathbb{P}(B_m) . \tag{2}$$

D'une part, comme $B_n \subset A_n$, on a $\bigcup_{n\in\mathbb{N}} B_n \subset \bigcup_{n\in\mathbb{N}} A_n$.

D'autre part, pour tout $m \in \mathbb{N}$, $A_m = \bigcup_{n=0}^{m} B_n \subset \bigcup_{n\in\mathbb{N}} B_n$. Comme c'est vrai pour tout m, on a $\bigcup_{m\in\mathbb{N}} A_m \subset \bigcup_{n\in\mathbb{N}} B_n$. On a donc au final $\lim_n A_n = \bigcup_{n\in\mathbb{N}} A_n = \bigcup_{n\in\mathbb{N}} B_n$. Par la propriété de σ-additivité de $\mathbb{P}$:

$$\mathbb{P}(\lim_{n\to+\infty} A_n) = \mathbb{P}\left(\bigcup_{m=0}^{+\infty} B_m\right) = \sum_{m=0}^{+\infty} \mathbb{P}(B_m) .$$

Comme la série $\sum_n \mathbb{P}(B_n)$ est convergente :

$$\mathbb{P}(\lim_{n\to+\infty} A_n) = \lim_{n\to+\infty} \sum_{m=0}^{n} \mathbb{P}(B_m)\,.$$

En utilisant l'équation (2), on obtient le résultat cherché :

$$\mathbb{P}(\lim_{n\to+\infty} A_n) = \lim_{n\to+\infty} \mathbb{P}(A_n)\,.$$

Le cas décroissant peut se déduire du cas croissant par passage aux complémentaires. ■

2.2 Probabilité uniforme sur un ensemble fini

Soient Ω un ensemble fini non vide, et $\mathcal{A} := \mathcal{P}(\Omega)$. On note $\operatorname{card}(A)$ le cardinal d'un sous-ensemble A de Ω. La formule :

$$\mathbb{P}(A) := \frac{\operatorname{card}(A)}{\operatorname{card}(\Omega)}\,,$$

définit une probabilité sur $(\Omega, \mathcal{A})$, appelée la probabilité uniforme sur l'ensemble fini Ω. On a en particulier $\mathbb{P}(\{\omega\}) = 1/\operatorname{card}(\Omega)$. Comme on l'a montré dans le livre [L1], la probabilité uniforme est la seule probabilité sur un ensemble fini qui satisfasse la propriété d'équiprobabilité (équiprobabilité de tous les résultats). Le calcul des probabilités se réduit alors à des dénombrements. Nous rappelons ici quelques notions élémentaires de dénombrement vues dans le livre [L1].

1) Le nombre de permutations de $[\![1, n]\!]$ (ou de bijections d'un ensemble à n éléments) est $n!$. Par convention, on pose $0! = 1$.

2) Le nombre d'arrangements de k éléments dans un ensemble à n éléments avec $k \leqslant n$ (ou d'injections de $[\![1, k]\!]$ dans $[\![1, n]\!]$) est $A_n^k = n!/(n-k)!$.

3) Le nombre de sous-ensembles à k éléments dans un ensemble contenant n éléments avec $k \leqslant n$ est $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$.

Exemple. **Dates d'anniversaire.** On considère une classe de n élèves. On cherche la probabilité pour qu'au moins deux élèves soient nés le même jour.

Pour simplifier, on suppose qu'il n'y pas d'année bissextile. On définit l'espace fondamental $\Omega := [\![1, 365]\!]^n$ avec $\omega = (\omega_1, \ldots, \omega_n)$ où ω_i est le jour de naissance de l'élève i. La tribu naturelle associée est $\mathcal{A} = \mathcal{P}(\Omega)$. On choisit la probabilité uniforme sur Ω. On note A l'événement « au moins 2 élèves sont nés le même jour ». On peut maintenant raisonner :

$$\begin{aligned}
\mathbb{P}(A) &= 1 - \mathbb{P}(A^c) = 1 - \mathbb{P}(\text{« tous les élèves sont nés des jours différents »}) \\
&= 1 - \mathbb{P}(\{\omega \mid \omega_i \neq \omega_j, \forall i \neq j\}) = 1 - \frac{\operatorname{card}(\{\omega \mid \omega_i \neq \omega_j, \forall i \neq j\})}{365^n} \\
&= 1 - \frac{\operatorname{card}(\{\text{injections de } [\![1, n]\!] \text{ dans } [\![1, 365]\!]\})}{365^n} \\
&= 1 - \frac{365!}{(365-n)!365^n} = p_n\,.
\end{aligned}$$

On obtient les valeurs numériques suivantes :

$$p_{20} \simeq 0{,}411\,, \quad p_{23} \simeq 0{,}507\,, \quad p_{30} \simeq 0{,}706\,, \quad p_{50} \simeq 0{,}970\,, \quad p_{366} = 1\,.$$

Pour obtenir ces valeurs numériques, il faut utiliser une bonne calculatrice, ou alors raisonner un peu plus. En effet, $1 - p_n$ s'écrit aussi :

$$1-p_n = \frac{365}{365}\frac{365-1}{365}\cdots\frac{365-n+1}{365} = \prod_{k=1}^{n-1}\left(1-\frac{k}{365}\right) = \exp\left(\sum_{k=1}^{n-1}\ln\left(1-\frac{k}{365}\right)\right)\,.$$

Si n n'est pas grand, une évaluation approximative de cette somme se fait en remplaçant $\ln(1-x)$ par $-x$ et en utilisant la somme d'une progression arithmétique :

$$\sum_{k=1}^{n-1} k = \frac{n(n-1)}{2}\,,$$

qui conduit à l'approximation $1-p_n \approx \exp(-n^2/730)$. Pour voir par exemple pour quel n on a $p_n \sim 1/2$ on prend $n \approx \sqrt{730\ln 2} \sim 23$. Pour un calcul plus rigoureux, on peut utiliser l'encadrement valable pour tout $0 < x < 1$:

$$-x - \frac{x^2}{2(1-x)} < \ln(1-x) < -x - \frac{x^2}{2}\cdot$$

Pour obtenir ces majorations, on part du développement en série entière :

$$\ln(1-x) = -\sum_{k=1}^{+\infty}\frac{x^k}{k}\,,$$

qui montre immédiatement la majoration de droite, alors que la majoration de gauche s'obtient par :

$$\ln(1-x) > -x - \frac{1}{2}\sum_{k=2}^{+\infty} x^k = -x - \frac{x^2}{2(1-x)}\cdot$$

On a aussi besoin de la somme des premiers carrés :

$$\sum_{k=1}^{n-1} k^2 = \frac{n(2n-1)(n-1)}{6}\,,$$

qui s'établit par récurrence. Si $x \leqslant (n-1)/365$, alors $-1/(1-x) \geqslant -365/(365-n+1)$. D'où l'encadrement rigoureux :

$$\begin{aligned}-\frac{n(n-1)(2n-1)}{6}\frac{1}{2\times 365^2}\frac{365}{365-n+1} &< \ln(1-p_n) + \frac{n(n-1)}{2}\frac{1}{365}\\ &< -\frac{n(n-1)(2n-1)}{6}\frac{1}{2\times 365^2}\,.\end{aligned}$$

Considérons maintenant le problème voisin suivant : on cherche la probabilité de l'événement $B =$ « au moins un élève est né le même jour que le professeur ». En notant x le jour

de naissance du professeur, on a :

$\mathbb{P}(B) = 1 - \mathbb{P}$ (« tous les élèves sont nés un jour différent de x »)

$$= 1 - \mathbb{P}\left(\{\omega \mid \omega_i \neq x, \forall i\}\right) = 1 - \frac{\operatorname{card}\left(\{\omega \mid \omega_i \neq x, \forall i\}\right)}{365^n} = 1 - \left(\frac{364}{365}\right)^n = q_n\,.$$

On obtient les valeurs numériques suivantes :

$$q_{20} \simeq 0{,}053\,, \quad q_{23} \simeq 0{,}061\,, \quad q_{30} \simeq 0{,}079\,, \quad q_{50} \simeq 0{,}128\,, \quad q_{366} \simeq 0{,}634\,.$$

Ici aussi, on peut obtenir une expression approchée de q_n en remarquant que :

$$\ln(1 - q_n) = n \ln\left(\frac{364}{365}\right) = n \ln\left(1 - \frac{1}{365}\right) \simeq -\frac{n}{365}\,,$$

si bien que $q_n \simeq 1 - \exp(-n/365)$.

Remarquez que les valeurs de p_n et q_n sont très différentes, bien que les énoncés des deux problèmes soient proches. Dès qu'une classe ou une assemblée contient plus de 23 personnes, il y a plus d'une chance sur deux pour que deux personnes soient nées le même jour !

2.3 Probabilités sur un ensemble dénombrable

Soit Ω un ensemble fini ou infini dénombrable. Par définition, Ω est en bijection avec $\{1, \ldots, n\}$ (s'il est fini de cardinal n) ou avec $\mathbb{N}^*$ (s'il est infini dénombrable), et l'on peut numéroter ses éléments par des entiers. La proposition qui suit montre qu'une probabilité $\mathbb{P}$ sur $\big(\Omega, \mathcal{P}(\Omega)\big)$ est entièrement caractérisée par les valeurs des probabilités des singletons $\mathbb{P}(\{\omega\})$, $\omega \in \Omega$.

Proposition 3. Soit $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$ un ensemble fini ou dénombrable.
(i) Toute suite de réels positifs $(p_i)_{i \geqslant 1}$, telle que $\sum\limits_{i \geqslant 1} p_i = 1$, définit une probabilité $\mathbb{P}$ sur $\mathcal{A} = \mathcal{P}(\Omega)$:

$$\mathbb{P}(A) = \sum_{i \mid \omega_i \in A} p_i\,. \tag{3}$$

La probabilité $\mathbb{P}$ vérifie $\mathbb{P}(\{\omega_i\}) = p_i$ pour tout $i \geqslant 1$.
(ii) Réciproquement, toute probabilité $\mathbb{P}$ sur $\big(\Omega, \mathcal{P}(\Omega)\big)$ est du type précédent, avec $p_i = \mathbb{P}(\{\omega_i\})$.

Démonstration. Regardons le (i). Soit $\mathbb{P}$ définie par (3). Pour tout $A \in \mathcal{P}(\Omega)$, on a $\mathbb{P}(A) \in [0, 1]$. Par ailleurs, pour tout $i \geqslant 1$, $\mathbb{P}(\{\omega_i\}) = \sum\limits_{j \mid \omega_j = \omega_i} p_j = p_i$. Il faut montrer que $\mathbb{P}$ vérifie les conditions de la définition 2 d'une probabilité (page 669). La première condition est facile car $\mathbb{P}(\Omega) = \sum\limits_{i \geqslant 1} p_i = 1$. Vérifions maintenant la seconde condition (la σ-additivité de $\mathbb{P}$). Soit $(A_i)_{i \in \mathbb{N}}$ une suite d'événements deux à deux disjoints. Alors :

$$\mathbb{P}\Big(\bigcup_{i \in \mathbb{N}} A_i\Big) = \sum_{j \mid \omega_j \in \bigcup\limits_{i \in \mathbb{N}} A_i} p_j.$$

Comme on a affaire à une série à termes positifs, le théorème de Fubini pour les séries à termes positifs (*cf.* le livre [L2], module « Suites et séries de fonctions ») nous autorise à effectuer cette somme dans l'ordre que nous préférons, donc :

$$\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \sum_{j|\omega_j\in A_i} p_j = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)\,.$$

Montrons la réciproque (ii). Soit $\mathbb{P}$ une probabilité sur $\big(\Omega, \mathcal{P}(\Omega)\big)$. Soit $A \in \mathcal{A}$. Les événements $(\{\omega\})_{\omega\in A}$ sont disjoints deux à deux, donc la σ-additivité de $\mathbb{P}$ implique que :

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{\omega\in A} \{\omega\})\right) = \sum_{\omega\in A} \mathbb{P}(\{\omega\})\,. \qquad \blacksquare$$

Exemple. Reprenons l'exemple 4 de la page 666 : on lance une pièce jusqu'à obtenir Pile. Si l'on note p (avec $0 < p < 1$) la probabilité d'obtenir Pile lors d'un lancer (on généralise ici le cas naturel $p = 1/2$), alors on a :

$$\mathbb{P}(\{\omega_i\}) = (1-p)^{i-1}p, \quad i \geqslant 1\,.$$

Par la propriété d'additivité dénombrable, l'événement $A =$ « on finit par obtenir Pile » est de probabilité 1 :

$$\mathbb{P}(A) = \sum_{i\in\mathbb{N}^*} \mathbb{P}(\{\omega_i\}) = \sum_{i\in\mathbb{N}^*} (1-p)^{i-1}p = 1\,.$$

Remarque. Il n'existe pas de probabilité qui satisfasse la propriété d'équiprobabilité sur un ensemble fondamental infini dénombrable Ω. En effet, si une telle probabilité existait, il faudrait que la probabilité d'un singleton $\{\omega\}$, $\omega \in \Omega$, soit indépendante de ω. En d'autres mots, il faudrait qu'il existe $p \in [0,1]$ tel que $\mathbb{P}(\{\omega\}) = p\ \forall\omega \in \Omega$. Mais alors, comme $\mathbb{P}$ est σ-additive, on devrait avoir $\mathbb{P}(\Omega) = \sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = \sum_{\omega\in\Omega} p$. Si $p = 0$, alors $\mathbb{P}(\Omega) = 0$, ce qui est impossible. Si $p > 0$, alors $\mathbb{P}(\Omega) = +\infty$, ce qui est aussi impossible. Une probabilité satisfaisant la propriété d'équiprobabilité sur un ensemble fondamental infini dénombrable ne peut donc pas exister. On va voir dans le prochain paragraphe une notion d'équiprobabilité convenable pour des ensembles infinis particuliers.

2.4 Probabilités uniformes sur $\mathbb{R}$ (ou $\mathbb{R}^d$)

Nous avons déjà rencontré l'idée d'équiprobabilité dans le cas d'ensembles finis. Elle se traduit par le fait que la probabilité de chaque événement élémentaire ou singleton est la même. Cela suffit à caractériser entièrement la probabilité uniforme par la proposition 3. Dans le cas de $\mathbb{R}$, la situation est plus délicate. Cherchons ce que pourrait être une probabilité uniforme sur un intervalle I de $\mathbb{R}$, par exemple $I = [0,1]$. Réclamer l'égalité des probabilités des singletons, comme d'ailleurs réclamer n'importe quelle distribution des probabilités des singletons, ne suffit pas à déterminer une probabilité, car beaucoup d'événements (sous-ensembles de I) ne sont pas dénombrables, et ne sont donc pas unions

dénombrables de singletons. Une bonne représentation de l'idée d'équiprobabilité est de réclamer que « la probabilité que le résultat tombe dans un sous-intervalle est proportionnelle à la longueur du sous-intervalle, et ne dépend pas de sa position » $\big($par exemple, $\mathbb{P}([0,1/2]) = \mathbb{P}([1/4,3/4]) = \mathbb{P}([1/2,1])\big)$. La théorie des probabilités se trouve alors naturellement liée avec la théorie de la mesure et de l'intégration. Nous utiliserons la proposition suivante (démontrée dans le corollaire 149 de la page 325 du module d'intégration).

Proposition 4 (Existence et unicité de la mesure de Lebesgue).
Il existe une unique fonction à valeurs positives définie sur la tribu des boréliens $\mathcal{B}(\mathbb{R})$, notée $\lambda : \mathcal{B}(\mathbb{R}) \to [0,+\infty]$, qui est σ-additive et qui se réduit à la longueur sur les intervalles : $\lambda\left(]a,b[\right) = b-a$, pour tous $a<b$.
On appelle λ la mesure de Lebesgue sur $\mathbb{R}$.

La tribu des boréliens $\mathcal{B}(\mathbb{R})$ a déjà été introduite page 668, c'est la plus petite tribu qui contienne les intervalles.

Notez que $\lambda([a,b]) = \lambda(]a,b[)$ pour tous $a<b$, car $\lambda([a,b]) = \lim\limits_{n\to+\infty} \lambda(]a-1/n, b+1/n[)$ par la propriété de stabilité par limite décroissante (comme dans la proposition 2). On a donc aussi $\lambda(\{a\}) = 0$ pour tout $a \in \mathbb{R}$.

Ainsi, si $I = [0,1]$, alors il existe une unique probabilité uniforme sur $\big([0,1], \mathcal{B}([0,1])\big)$ (c'est-à-dire qui est telle que « la probabilité que le résultat tombe dans un sous-intervalle est proportionnelle à la longueur du sous-intervalle »). Cette probabilité est la mesure de Lebesgue $\mathbb{P}(\cdot) := \lambda(\cdot \cap I)$. De même, si $I := [a,b]$, l'unique probabilité uniforme sur $\big(I, \mathcal{B}(I)\big)$ est $\mathbb{P}(\cdot) := \lambda(\cdot \cap I)/\lambda(I)$ avec $\lambda(I) = b-a$.

Le résultat précédent se généralise sur $\mathbb{R}^d$, $d \geqslant 1$.

Proposition 5. Pour tout entier $d \geqslant 1$ fixé, il existe une unique fonction à valeurs positives définie sur la tribu des boréliens $\mathcal{B}(\mathbb{R}^d)$, notée $\lambda : \mathcal{B}(\mathbb{R}^d) \to [0,+\infty]$, qui est σ-additive et qui se réduit au volume sur les pavés :

$$\lambda\big(]a_1,b_1[\times\cdots\times]a_d,b_d[\big) = \prod_{i=1}^{d}(b_i-a_i)\,,$$

où $a_i < b_i$ pour $i \in [\![1,d]\!]$. On appelle λ la mesure de Lebesgue sur $\mathbb{R}^d$.

La tribu $\mathcal{B}(\mathbb{R}^d)$ sur $\mathbb{R}^d$ est la plus petite tribu contenant les pavés (i.e., les produits d'intervalles), elle a été introduite page 668. Si $V \in \mathcal{B}(\mathbb{R}^d)$ est un domaine de $\mathbb{R}^d$ de mesure de

Lebesgue $\lambda(V) \in]0, +\infty[$, alors la probabilité uniforme sur le domaine V est définie par :

$$\mathbb{P}_V(A) = \frac{\lambda(A \cap V)}{\lambda(V)}, \quad A \in \mathcal{B}(\mathbb{R}^d).$$

Remarque. La tribu $\mathcal{B}(\mathbb{R})$ [respectivement $\mathcal{B}(\mathbb{R}^d)$] est la tribu borélienne sur $\mathbb{R}$ (respectivement sur $\mathbb{R}^d$), définie comme la plus petite tribu qui contient tous les intervalles de $\mathbb{R}$ (respectivement tous les pavés de $\mathbb{R}^d$). D'une part, on peut montrer que $\mathcal{B}(\mathbb{R}) \subsetneq \mathcal{P}(\mathbb{R})$ [respectivement $\mathcal{B}(\mathbb{R}^d) \subsetneq \mathcal{P}(\mathbb{R}^d)$], mais construire explicitement une partie de $\mathbb{R}$ (respectivement $\mathbb{R}^d$) qui ne soit pas borélienne n'est pas facile (et réclame l'axiome du choix). D'autre part, la tribu borélienne contient tous les ouverts et les fermés de $\mathbb{R}$ (respectivement de $\mathbb{R}^d$), car les ouverts de $\mathbb{R}$ (respectivement de $\mathbb{R}^d$) sont des unions dénombrables d'intervalles ouverts (respectivement de pavés), et les fermés sont des complémentaires d'ouverts. Enfin, la partie de la proposition 4 concernant l'existence d'une mesure σ-additive qui se réduit à la longueur sur les intervalles est fausse si l'on cherche à définir une telle mesure sur $\mathcal{P}(\mathbb{R})$ au lieu de $\mathcal{B}(\mathbb{R})$: c'est la raison principale pour laquelle on travaille avec l'espace $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ et non avec $\big(\mathbb{R}, \mathcal{P}(\mathbb{R})\big)$. On peut cependant étendre la mesure de Lebesgue à la tribu de Lebesgue, qui est strictement plus grande que la tribu des boréliens et strictement incluse dans $\mathcal{P}(\mathbb{R})$, comme discuté dans la section 3.2.3 de la page 325 du module d'intégration.

2.5 Probabilités de réunions d'ensembles : règle d'inclusion-exclusion

On sait calculer la probabilité d'une union d'événements deux à deux disjoints : c'est simplement la somme des probabilités des événements. On s'intéresse ici au calcul de la probabilité d'une union d'événements quelconques. La règle d'inclusion-exclusion s'applique à une suite $(A_i)_{i \in [\![1,n]\!]}$ d'événements pour lesquels on connaît *a priori* les probabilités des conjonctions d'événements. On peut ainsi calculer la probabilité de $A = A_1 \cup A_2 \cup \cdots \cup A_n$, i.e. la probabilité que « au moins l'un des A_i est réalisé ».

Proposition 6 (Formule d'inclusion-exclusion).
Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité et $(A_i)_{i \in [\![1,n]\!]}$ une suite finie d'événements. La formule d'inclusion-exclusion s'écrit :
1) Dans le cas de deux événements : $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$.
2) Dans le cas de trois événements :

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3) = {} & \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) \\ & - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3) \\ & + \mathbb{P}(A_1 \cap A_2 \cap A_3).\end{aligned}$$

3) Dans le cas général de n événements :

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right). \tag{4}$$

La relation (4) s'appelle formule d'inclusion-exclusion car c'est exactement ce que l'on fait. Pour calculer $\mathbb{P}(\bigcup_i A_i)$, on commence par calculer la somme $\sum_i \mathbb{P}(A_i)$. Mais on surestime ainsi le résultat, car on a compté plusieurs fois les intersections. On enlève alors la somme des probabilités des intersections d'ordre 2 : $\sum_{i\neq j} \mathbb{P}(A_i \cap A_j)$. Cette fois-ci, on sous-estime le résultat, car on a enlevé plusieurs fois les intersections d'ordre 3. Il faut alors les remettre, etc.

Démonstration. Le 1) a déjà été montré dans la proposition 1. On va montrer le cas général 3) par récurrence sur n. Supposons la formule montrée au rang n. Soit une suite d'événements $(A_i)_{i\in[\![1,n+1]\!]}$. En appliquant le 1) au couple $\bigcup_{i=1}^{n} A_i$ et A_{n+1}, on obtient :

$$\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathbb{P}\left(A_{n+1}\right) - \mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \bigcap A_{n+1}\right) .$$

En appliquant l'hypothèse de récurrence au premier terme du membre de droite :

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) .$$

En appliquant l'hypothèse de récurrence au troisième terme du membre de droite :

$$\begin{aligned}\mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \bigcap A_{n+1}\right) &= \mathbb{P}\left(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})\right) \qquad \text{par distributivité} \\ &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k} \cap A_{n+1}\right) .\end{aligned}$$

En regroupant :

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) \\ &\quad + \sum_{k=0}^{n} (-1)^{k} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k} \cap A_{n+1}\right) . \qquad (5)\end{aligned}$$

Or en cassant le membre de droite de l'équation (4) au rang $n+1$ en deux (avec d'un côté les ensembles d'indices qui contiennent $n+1$, et de l'autre côté ceux qui ne le contiennent pas), on obtient :

$$\begin{aligned}&\sum_{k=1}^{n+1} (-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n+1} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) \\ &\qquad = \sum_{k=1}^{n+1} (-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) \\ &\qquad\quad + \sum_{k=1}^{n+1} (-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_{k-1}\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_{k-1}} \cap A_{n+1}\right) .\end{aligned}$$

Dans le premier terme du membre de droite la somme s'arrête en fait à $k = n$, car il y a au plus n entiers distincts entre 1 et n. Dans le second terme, on opère le changement d'indice $k' = k - 1$. On retrouve alors l'expression (5).
On verra dans le module IV.2, page 721, une démonstration plus élégante et directe de cette proposition. ■

Application. Le problème des rencontres ou des « matchings » (nombre de dérangements). N hommes sont invités à une réception. Chacun donne son chapeau au vestiaire. Les chapeaux sont mélangés et à la fin de la réception chaque invité en choisit un au hasard. Quelle est la probabilité qu'aucun des hommes ne choisisse son propre chapeau ? Quelle est la probabilité que k hommes exactement sélectionnent leur propre chapeau ? Du point de vue mathématique ce problème est équivalent au tirage d'une permutation aléatoire σ de l'ensemble $[\![1, N]\!]$. Un point fixe i d'une permutation $\sigma = \big(\sigma(1), \sigma(2), \ldots, \sigma(N)\big)$ est défini par la condition $\sigma(i) = i$. Une permutation sans point fixe, i.e. telle que $\sigma(i) \neq i$ pour tout $i \in [\![1, N]\!]$, est appelée dérangement.

Exemple. On présente ci-dessous trois permutations de $[\![1, 6]\!]$ dont la première est un dérangement mais pas les deux suivantes :

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 1 & 6 & 3 & 5 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 & \mathit{5} & 6 \\ 2 & 3 & 6 & 1 & \mathit{5} & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & \mathit{2} & 3 & 4 & \mathit{5} & \mathit{6} \\ 4 & \mathit{2} & 1 & 3 & \mathit{5} & \mathit{6} \end{pmatrix}$$

un dérangement — un point fixe — trois points fixes

L'ensemble fondamental Ω est l'ensemble de toutes les permutations de $[\![1, N]\!]$, donc on a $\operatorname{card}(\Omega) = N!$. En faisant l'hypothèse d'équiprobabilité, la probabilité d'un événement A est donnée par $\mathbb{P}(A) := \operatorname{card}(A)/\operatorname{card}(\Omega)$. Pour la première partie on calcule la probabilité du complémentaire de l'événement « au moins un homme prend son propre chapeau ». Appelons A_i l'événement « le $i^{\text{ème}}$ homme choisit son propre chapeau » (i.e. $\sigma(i) = i$). Donc :

$$A := A_1 \cup A_2 \cup \cdots \cup A_N$$

est l'événement « au moins un homme choisit son propre chapeau ».
On vérifie que l'on peut construire une bijection Ψ de A_i dans l'ensemble des permutations de $\{1, \ldots, i-1, i+1, \ldots, N\}$ en posant $\Psi(\sigma) = (\sigma(1), \ldots, \sigma(i-1), \sigma(i+1), \ldots, \sigma(N))$. Cela montre que $\operatorname{card}(A_i) = (N-1)!$ et donc $\mathbb{P}(A_i) = (N-1)!/N! = 1/N$. De même la probabilité que deux hommes, par exemple 1 et 2, choisissent leur propre chapeau est $\mathbb{P}(A_1 \cap A_2) = (N-2)!/N!$ car on peut construire une bijection de $A_1 \cap A_2$ dans l'ensemble des permutations de $\{3, \ldots, N\}$. En généralisant cet argument, on trouve que la probabilité de l'intersection de k événements, par exemple $A_1 \cap A_2 \cap \cdots \cap A_k$, est donnée par :

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_k) = \frac{(N-k)!}{N!} \cdot$$

Il y a $\binom{N}{k}$ termes dans la $k^{\text{ème}}$ somme de la formule d'inclusion-exclusion correspondant

aux $\binom{N}{k}$ combinaisons possibles de k indices. Donc :

$$\mathbb{P}(A) = \sum_{k=1}^{N} \binom{N}{k} (-1)^{k+1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_k) = \sum_{k=1}^{N} \frac{(-1)^{k+1}}{k!} \cdot$$

On désigne par $P_N(k)$ la probabilité que k des hommes exactement sélectionnent leur propre chapeau. On vient de démontrer que :

$$P_N(0) = 1 - \mathbb{P}(A) = 1 - \sum_{k=1}^{N} \frac{(-1)^{k+1}}{k!} = \sum_{j=0}^{N} \frac{(-1)^j}{j!} \cdot$$

Il y a $\binom{N}{k}$ possibilités pour que k des hommes exactement sélectionnent leur propre chapeau. Alors pour les autres $(N-k)$ hommes il y a $(N-k)!P_{N-k}(0)$ possibilités qu'aucun de ces hommes ne choisisse son propre chapeau. Par conséquent :

$$P_N(k) = \frac{\binom{N}{k}(N-k)!P_{N-k}(0)}{N!} = \frac{P_{N-k}(0)}{k!} = \sum_{j=0}^{N-k} \frac{(-1)^j}{j!k!} \cdot$$

Notons que $P_N(0) \underset{N\to+\infty}{\longrightarrow} \exp(-1)$ (c'est la série exponentielle $\exp(-x)$ pour $x = 1$). Plus généralement, pour chaque k fixé, on a $P_N(k) \to \exp(-1)/k!$. En conclusion, nous avons démontré un théorème limite (on en verra d'autres plus généraux plus tard) pour la distribution de matchings : la distribution donnée par les $P_N(k)$ converge vers la distribution dite de Poisson de paramètre 1 (voir la définition 8 à la page 729).

3 Probabilités conditionnelles

3.1 Intuition et définition

Cette notion s'introduit chaque fois que l'on acquiert une information partielle sur le résultat d'une expérience aléatoire. Pour la dégager, procédons comme page 668 où nous avons défini les probabilités par analogie avec les fréquences empiriques d'apparition. Considérons une expérience aléatoire décrite par un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ (par exemple, le battage de cartes de l'exemple 2 de la page 666). On cherche à évaluer la probabilité qu'un certain événement A se réalise (par exemple, la dernière carte est un as) sachant qu'un événement B s'est réalisé (par exemple, la première carte est un as). Pour avoir une estimée de cette probabilité conditionnelle, on va mener N répétitions de l'expérience aléatoire dans des « conditions identiques et indépendantes », puis on ne va considérer que les $N(B)$ répétitions de l'expérience qui ont réalisé B. On peut alors penser que la fréquence $N(A \cap B)/N(B)$ de réalisation de l'événement A parmi les $N(B)$ répétitions de l'expérience qui ont réalisé B fournit une estimation de la probabilité conditionnelle de A sachant B. Puisque $N(A \cap B)/N(B)$ est le rapport de $N(A \cap B)/N$ par $N(B)/N$, la probabilité conditionnelle de A sachant B apparaît ici comme le rapport de $\mathbb{P}(A \cap B)$ par $\mathbb{P}(B)$. Nous adopterons ce rapport comme définition d'une probabilité conditionnelle.

Définition 3. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. La *probabilité d'un événement A conditionnelle en l'événement B*, notée $\mathbb{P}(A \mid B)$, est définie par la formule :

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

pourvu que $\mathbb{P}(B) > 0$.

Notez que la barre « $\mid$ » située entre deux événements doit se lire « conditionnellement à » ou « sachant ».
Fixons un événement B dont la probabilité est supposée strictement positive. La fonction définie sur $\mathcal{A}$ par $A \mapsto \mathbb{P}(A \mid B)$ est elle-même une probabilité sur $(\Omega, \mathcal{A})$, comme le montre la proposition suivante.

Proposition 7. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. Soit $B \in \mathcal{A}$ un événement de probabilité non nulle $\mathbb{P}(B) > 0$. La fonction définie sur $\mathcal{A}$ par :

$$A \mapsto \mathbb{P}(A \mid B), \quad A \in \mathcal{A}$$

est une probabilité sur $(\Omega, \mathcal{A})$.

Démonstration. Il faut prouver que la fonction $\begin{array}{ccc} \mathcal{A} & \longrightarrow & \mathbb{R} \\ A & \longmapsto & \mathbb{P}(A \mid B) \end{array}$ vérifie les conditions nécessaires pour être une une probabilité. La première condition de la définition 2 de la page 669 explique pourquoi nous avons divisé par $\mathbb{P}(B)$. En effet :

$$\mathbb{P}(\Omega \mid B) = \frac{\mathbb{P}(B \cap \Omega)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1\,.$$

Si maintenant nous avons une suite d'événements $(A_i)_{i\in\mathbb{N}} \in \Omega^{\mathbb{N}}$ deux à deux disjoints, alors les événements $(B \cap A_i)_{i\in\mathbb{N}}$ sont deux à deux disjoints, et donc :

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i \mid B\right) &= \frac{\mathbb{P}\left(B \bigcap (\bigcup_{i\in\mathbb{N}} A_i)\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} (B \cap A_i)\right)}{\mathbb{P}(B)} \\ &= \frac{\sum_{i\in\mathbb{N}} \mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} = \sum_{i\in\mathbb{N}} \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i \mid B)\,,\end{aligned}$$

ce qui montre que $\mathbb{P}(\cdot \mid B)$ vérifie la propriété de σ-additivité. ∎

Exemple. Considérons la probabilité uniforme $\mathbb{P}_u$ définie sur un espace Ω fini muni de la tribu $\mathcal{P}(\Omega)$ par :

$$\mathbb{P}_u(A) := \frac{\operatorname{card}(A)}{\operatorname{card}(\Omega)}\,.$$

Si l'événement B n'est pas impossible ($B \neq \varnothing$), sa probabilité $\mathbb{P}_u(B)$ est strictement positive et pour tout événement A :

$$\mathbb{P}_u(A \mid B) = \frac{\mathbb{P}_u(A \cap B)}{\mathbb{P}_u(B)} = \frac{\operatorname{card}(A \cap B)}{\operatorname{card}(B)}\,.$$

Si $A \subset B$, cette formule se réduit à :

$$\mathbb{P}_u(A \mid B) = \frac{\operatorname{card}(A)}{\operatorname{card}(B)}\cdot$$

Cela montre que, pour les événements A tels que $A \subset B$, la probabilité conditionnelle $\mathbb{P}_u(\cdot \mid B)$ est égale à la probabilité uniforme sur B.

Il n'y a pas de relation d'ordre simple entre la probabilité conditionnelle d'un événement et sa probabilité simple, comme le montre l'exercice suivant.

Exercice 1.
On jette deux dés. Considérons les événements $A :=$« la somme est 5 », $B :=$« le premier dé donne 3 » et $C :=$« le premier dé donne au moins 3 ». Calculer $\mathbb{P}(A)$, $\mathbb{P}(A \mid B)$ et $\mathbb{P}(A \mid C)$.
Solution.
On a $\mathbb{P}(A) = 4/36 = 1/9$, $\mathbb{P}(B) = 1/6$ et $\mathbb{P}(A \cap B) = 1/36$, donc $\mathbb{P}(A \mid B) = 1/6$. La probabilité conditionnelle de A sachant B est plus grande que la probabilité de A sans autre information : $\mathbb{P}(A|B) > \mathbb{P}(A)$. D'autre part :

$$\mathbb{P}(C) = 4/6 = 2/3 \qquad \text{et} \qquad \mathbb{P}(A \cap C) = 2/36 = 1/18,$$

on trouve donc $\mathbb{P}(A \mid C) = 1/12$. La probabilité conditionnelle de A sachant C est plus petite que la probabilité de A sans autre information : $\mathbb{P}(A|C) < \mathbb{P}(A)$.

3.2 Formule de Bayes

La formule de définition des probabilités conditionnelles est souvent utile sous sa forme multiplicative :

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \text{ si } \mathbb{P}(B) > 0\,. \tag{6}$$

La proposition élémentaire suivante est fondamentale.
On rappelle qu'une suite B_i d'événements (suite finie, $i \in [\![1, n]\!]$, ou infinie $i \geqslant 1$) est appelée partition de Ω si $B_i \cap B_j = \varnothing$ $\forall i \neq j$ et si $\bigcup_i B_i = \Omega$. Chaque $\omega \in \Omega$ appartient alors à exactement un des B_i.

Proposition 8 (Formule des probabilités totales).
Pour tous événements A et B (avec $0 < \mathbb{P}(B) < 1$) :

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)\,.$$

Plus généralement, si $(B_i)_{i\in I}$ est une partition finie ou dénombrable de Ω, avec $\mathbb{P}(B_i) \in]0, 1[$ $\forall i \in I$, alors :

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i)\,.$$

Démonstration. Puisque les $A \cap B_i$ sont deux à deux disjoints avec $A = \bigcup_{i\in I}(A \cap B_i)$:

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A \cap B_i) = \sum_{i\in I} \mathbb{P}(A|B_i)\mathbb{P}(B_i)\,,$$

d'après la définition 3 et l'équation (6). ∎

Exemple. On dispose de deux pièces. L'une est honnête, l'autre a deux piles. On choisit une pièce au hasard et on la lance trois fois. Quelle est la probabilité d'obtenir trois piles ? Pour résoudre ce problème on conditionne par la pièce choisie, i.e. on applique la formule des probabilités totales pour les événements $A :=$ « on tire trois piles », $B_1 :=$ « la pièce est honnête » et $B_2 :=$ « la pièce a deux piles ». On a :

$$\mathbb{P}(A) = \mathbb{P}(A\,|\,B_1)\mathbb{P}(B_1) + \mathbb{P}(A\,|\,B_2)\mathbb{P}(B_2) = \frac{1}{8} \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{9}{16}\cdot$$

Exercice 2.

M. et Mme Personne ont deux enfants, garçons ou filles, les quatre configurations possibles sur le sexe des enfants étant équiprobables.

1. Quelle est la probabilité que les deux enfants Personne soient des garçons ?
   (a) sans autre information.
   (b) sachant que l'aîné est un garçon.
   (c) sachant que l'un des enfants au moins est un garçon.
2. Vous téléphonez chez les Personne en l'absence des parents. C'est un garçon qui répond. Quelle est la probabilité que les deux enfants Personne soient des garçons ? Ici, on ajoute que, dans les familles comprenant une fille et un garçon, c'est la fille qui répond au téléphone avec probabilité p.

Solution.

1. L'espace fondamental est $\Omega := \{F, G\}^2$. Une réalisation est une paire $\omega = (\omega_1, \omega_2)$ où ω_1 est le sexe de l'aîné et ω_2 est le sexe du cadet. L'espace fondamental est muni de la probabilité uniforme. On note A l'événement « les deux enfants sont des garçons ». On a bien sûr $A = \{(GG)\}$.
   (a) $\mathbb{P}(A) = \operatorname{card}(A)/\operatorname{card}(\Omega) = 1/4$.
   (b) Soit B l'événement « l'aîné est un garçon ». On a $B = \{(GG), (GF)\}$. Donc :
   $$\mathbb{P}(A\,|\,B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\operatorname{card}(A \cap B)}{\operatorname{card}(B)} = \frac{1}{2}\cdot$$
   (c) Soit C l'événement « un des enfants au moins est un garçon ».
   On a $C = \{(GG), (GF), (FG)\}$. Donc :
   $$\mathbb{P}(A\,|\,C) = \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)} = \frac{\operatorname{card}(A \cap C)}{\operatorname{card}(C)} = \frac{1}{3}\cdot$$
2. Pour être rigoureux, il faut changer la définition de l'espace fondamental et prendre par exemple $\Omega := \{F, G\}^2 \times \{1, 2\}$. Le dernier élément du triplet ω représente l'indice de l'enfant qui répond au téléphone. Ici, l'espace fondamental n'est pas muni de la probabilité uniforme ! On note D l'événement « c'est un garçon qui répond ». D'après la formule des probabilités totales :
   $$\begin{aligned}\mathbb{P}(D) &= \mathbb{P}(D\,|\,\{(FF)\})\mathbb{P}(\{(FF)\}) + \mathbb{P}(D\,|\,\{(FG)\})\mathbb{P}(\{(FG)\}) \\ &\quad + \mathbb{P}(D\,|\,\{(GF)\})\mathbb{P}(\{(GF)\}) + \mathbb{P}(D\,|\,\{(GG)\})\mathbb{P}(\{(GG)\}) \\ &= 0 \times \frac{1}{4} + (1-p) \times \frac{1}{4} + (1-p) \times \frac{1}{4} + 1 \times \frac{1}{4} = \frac{3-2p}{4},\end{aligned}$$
   et l'on a de même : $\mathbb{P}(A \cap D) = \mathbb{P}(D \cap \{(GG)\}) = \mathbb{P}(D\,|\,\{(GG)\})\mathbb{P}(\{(GG)\}) = \dfrac{1}{4}\cdot$

Donc :

$$\mathbb{P}(A\,|\,D) = \frac{\mathbb{P}(A \cap D)}{\mathbb{P}(D)} = \frac{1}{3-2p}\,.$$

Notez que :

Si $p = 1$, alors $\mathbb{P}(A\,|\,D) = 1$. En effet, dans ce cas, si un garçon répond, on est sûr qu'il n'y avait pas de fille à la maison, sinon elle aurait répondu.

Si $p = 0$, alors $\mathbb{P}(A \mid D) = 1/3$. En effet, dans ce cas, si un garçon répond, on a comme information qu'il y a au moins un garçon dans la famille, et l'on retrouve le résultat du 1(c).

Proposition 9 (Formule de Bayes).
Pour tous événements A et B avec $\mathbb{P}(B) \in]0,1[$ et $\mathbb{P}(A) > 0$:

$$\mathbb{P}(B\,|\,A) = \frac{\mathbb{P}(A\,|\,B)\mathbb{P}(B)}{\mathbb{P}(A\,|\,B)\mathbb{P}(B) + \mathbb{P}(A\,|\,B^c)\mathbb{P}(B^c)}\,.$$

Plus généralement, si $(B_i)_{i\in I}$ est une partition finie ou dénombrable de Ω avec $\mathbb{P}(B_i) \in]0,1[\ \forall i \in I$, et si A est un événement avec $\mathbb{P}(A) > 0$, alors, pour tout $k \in I$:

$$\mathbb{P}(B_k\,|\,A) = \frac{\mathbb{P}(A\,|\,B_k)\mathbb{P}(B_k)}{\sum\limits_{i\in I}\mathbb{P}(A\,|\,B_i)\mathbb{P}(B_i)}\,.$$

Démonstration. On écrit la définition d'une probabilité conditionnelle :

$$\mathbb{P}(B_k\,|\,A) = \frac{\mathbb{P}(B_k \cap A)}{\mathbb{P}(A)}\,.$$

Le numérateur peut se réécrire $\mathbb{P}(B_k \cap A) = \mathbb{P}(B_k \mid A)\mathbb{P}(A)$ d'après l'équation (6). On applique enfin la formule des probabilités totales au dénominateur. ■

Exercice 3.

Un laboratoire propose un test de dépistage de la maladie de la vache folle. La notice précise la qualité du test : lorsque le test est appliqué sur une vache malade, le test est positif dans 99,8 %des cas. Lorsque le test est appliqué sur une vache saine, le test est négatif dans 99,6 % des cas. On sait d'autre part qu'il y a une vache malade sur 100 000. Peut-on avoir confiance dans ce test ? Pour cela, on cherche la réponse à la question : si le test est positif, quelle est la probabilité que la vache soit malade ?

Solution. On note M l'événement « la vache testée est malade » et T^+ l'événement « le test est positif ». On cherche $\mathbb{P}(M\,|\,T^+)$:

$$\begin{aligned}\mathbb{P}(M\,|\,T^+) &= \frac{\mathbb{P}(M \cap T^+)}{\mathbb{P}(T^+)} = \frac{\mathbb{P}(T^+\,|\,M)\mathbb{P}(M)}{\mathbb{P}(T^+\,|\,M)\mathbb{P}(M) + \mathbb{P}(T^+\,|\,M^c)\mathbb{P}(M^c)}\\ &= \frac{0{,}998\ 10^{-5}}{0{,}998\ 10^{-5}\ +\ (1-0{,}996)\ (1-10^{-5})}\,,\end{aligned}$$

soit $\mathbb{P}(M\,|\,T^+) = 0{,}25\ \%$.

Ce résultat peut sembler étonnant, car les informations indiquées sur la notice tendent à montrer que le test est très bon. Or, si l'on prend la décision de tuer les vaches qui sont positives au test, le résultat précédent montre que l'on va tuer de nombreuses vaches pour rien. Cependant, le test n'est pas à jeter. Considérons maintenant la question : si le test est négatif, quelle est la probabilité que la vache soit malade ? Notez que ce cas est catastrophique : cela correspond à laisser passer une vache malade dans la chaîne de distribution alimentaire.
On note toujours M l'événement « la vache testée est malade » et T^- l'événement « le test est négatif ». On a bien sûr $T_- = T_+^c$. On cherche $\mathbb{P}(M\,|\,T^-)$:

$$\begin{aligned}\mathbb{P}(M\,|\,T^-) &= \frac{\mathbb{P}(M\cap T^-)}{\mathbb{P}(T^-)} = \frac{\mathbb{P}(T^-\,|\,M)\mathbb{P}(M)}{\mathbb{P}(T^-\,|\,M)\mathbb{P}(M)+\mathbb{P}(T^-\,|\,M^c)\mathbb{P}(M^c)}\\ &= \frac{(1-0{,}998)\;10^{-5}}{(1-0{,}998)\;10^{-5}\;+0{,}996\;(1-10^{-5})}\,,\end{aligned}$$

soit $\mathbb{P}(M\,|\,T^-) \simeq 2\;10^{-6}$ %. Ainsi, à cause du test, on va tuer des vaches pour rien, mais grâce au test, la probabilité de laisser passer une vache malade dans la chaîne de distribution alimentaire est très faible.

3.3 Conditionnement multiple

Nous avons déjà vu que la probabilité de la conjonction des deux événements peut être calculée par conditionnement sur un des événements. L'itération de cette technique est appelée conditionnement multiple.

Proposition 10 (théorème du conditionnement multiple).
Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité et $(A_i)_{i\in[\![1,n]\!]}$ une suite finie d'événements. La formule du conditionnement multiple s'écrit :
1. Dans le cas de deux événements, si A_1 est de probabilité non nulle :

$$\mathbb{P}(A_1\cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2\,|\,A_1)\,.$$

2. Dans le cas de trois événements, si $A_1\cap A_2$ est de probabilité non nulle :

$$\mathbb{P}(A_1\cap A_2\cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2\,|\,A_1)\mathbb{P}(A_3\,|\,A_1\cap A_2)\,.$$

3. Dans le cas général, si $A_1\cap\cdots\cap A_{n-1}$ est de probabilité non nulle :

$$\mathbb{P}(A_1\cap A_2\cap\cdots\cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2\,|\,A_1)\cdots\mathbb{P}(A_n\,|\,A_1\cap A_2\cap\cdots\cap A_{n-1})\,.$$

Démonstration. Notons d'abord que, si $A_1\cap\cdots\cap A_{n-1}$ est de probabilité non nulle, alors tous les événements $A_1\cap\cdots\cap A_k$, $k\in[\![1,n-1]\!]$, sont de probabilité non nulle d'après le point 4. de la proposition 1 de la page 669.
Le 1. est une application de la définition 3 de la probabilité conditionnelle. Montrons le cas général par récurrence sur n. On applique la définition de la probabilité conditionnelle pour $\mathbb{P}(A\,|\,B)$ avec $A = A_n$ et $B = A_1\cap A_2\cap\cdots\cap A_{n-1}$:

$$\mathbb{P}(A_1\cap A_2\cap\cdots\cap A_n) = \mathbb{P}(A_1\cap A_2\cap\cdots\cap A_{n-1})\mathbb{P}(A_n\,|\,A_1\cap A_2\cap\cdots\cap A_{n-1})\,.$$

Il suffit d'appliquer l'hypothèse de récurrence au premier terme du membre de droite. ∎

Exercice 4.
Considérons une urne contenant six boules blanches et quatre boules rouges. On tire successivement trois boules sans remise. Quelle est la probabilité de la suite « blanc, blanc, rouge » ?
Solution. Pour répondre à cette question, nous désignons les événements $B_1 :=$ « la première boule est blanche », $B_2 :=$ « la deuxième boule est blanche » et $B_3 :=$ « la troisième boule est rouge ». Alors :

$$\mathbb{P}(B_1 \cap B_2 \cap B_3) = \mathbb{P}(B_1)\mathbb{P}(B_2 \,|\, B_1)\mathbb{P}(B_3 \,|\, B_1 \cap B_2) = \frac{6}{10} \times \frac{5}{9} \times \frac{4}{8} = \frac{1}{6}\cdot$$

Remarquons qu'il faut conditionner dans le bon ordre. Par exemple, la formule :

$$\mathbb{P}(B_1 \cap B_2 \cap B_3) = \mathbb{P}(B_3)\mathbb{P}(B_2 \,|\, B_3)\mathbb{P}(B_1 \,|\, B_2 \cap B_3)$$

est juste mais elle ne sert à rien en pratique.

4 Indépendance

4.1 Indépendance de deux événements

Nous allons introduire la notion d'indépendance. Intuitivement, on parle d'indépendance lorsque la réalisation d'un événement A n'a pas d'influence sur la probabilité de réalisation d'un autre événement B. Prenons l'exemple du lancer de deux pièces. Soient $A :=$ « la première donne pile » et $B :=$ « la seconde donne pile ». A et B sont intuitivement indépendants, dans le sens où le résultat du lancer de la première pièce n'influe pas sur le résultat du second lancer. D'après la définition d'une probabilité conditionnelle, on a $\mathbb{P}(B \mid A) = \mathbb{P}(B \mid A^c) = \mathbb{P}(B)$. Cette relation est vérifiée si, et seulement si, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Cette dernière formule, qui a un sens même quand l'une des probabilités est nulle, est celle que nous adopterons comme définition.

Définition 4. Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité et A, B des événements. A et B sont indépendants si :

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\,.$$

Remarques.
1) Il ne faut pas confondre les notions d'événements indépendants et d'événements incompatibles. Si deux événements sont incompatibles, alors la réalisation de l'un d'entre eux influe beaucoup sur la probabilité de réalisation de l'autre, puisqu'en fait elle l'annule. En fait, si deux événements A et B sont à la fois indépendants et incompatibles, alors $\mathbb{P}(A \cap B) = 0$ car $A \cap B = \varnothing$. Donc $0 = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ ce qui montre que A ou B est de probabilité nulle.
2) La relation d'indépendance est commutative, mais pas transitive ! Contre-exemple : Si A et B sont indépendants, alors B et A sont indépendants, mais A et A ne sont pas (en général) indépendants, car alors on aurait $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$ qui ne peut être vérifié que si $\mathbb{P}(A) = 0$ ou 1.

L'exemple précédent du lancer de deux pièces est assez intuitif. Cependant, l'intuition peut être mise en défaut, dans le sens où deux événements peuvent sembler indépendants, alors qu'ils ne le sont pas, ou vice-versa. L'exercice qui suit illustre cet avertissement.

Exercice 5.

Une famille a n enfants où $n \geqslant 2$. Nous étudions les événements $A :=$ « la famille a des enfants des deux sexes » et $B :=$ « la famille a au plus une fille ». Montrer que A et B sont deux événements indépendants si, et seulement si, $n = 3$.

Solution. En faisant l'hypothèse d'équiprobabilité, on trouve successivement pour $\mathbb{P}(A \cap B)$, $\mathbb{P}(A)$ et $\mathbb{P}(B)$:

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}\left(\text{« la famille a 1 fille et } n-1 \text{ garçons »}\right) = n/2^n\,,\\
\mathbb{P}(A) &= 1 - \mathbb{P}\left(\text{« les enfants de la famille sont du même sexe »}\right)\\
&= 1 - \mathbb{P}\left(\text{« la famille a } n \text{ garçons »}\right) - \mathbb{P}\left(\text{« la famille a } n \text{ filles »}\right)\\
&= 1 - 2/2^n\,,\\
\mathbb{P}(B) &= \mathbb{P}\left(\text{« la famille a } n \text{ garçons »}\right) + \mathbb{P}\left(\text{« la famille a 1 fille et } n-1 \text{ garçons »}\right)\\
&= (n+1)/2^n\,.
\end{aligned}$$

Les événements A et B sont indépendants si, et seulement si :

$$n2^{-n} = (1 - 2^{-n+1})(n+1)2^{-n}\,.$$

Une étude de la fonction $x \mapsto x - (1 - 2^{-x+1})(x+1)$ montre qu'elle admet un unique zéro sur $\mathbb{R}_+$ en $x = 3$ ce qui montre le résultat.

La notion d'indépendance est liée à la probabilité $\mathbb{P}$ en vigueur. Deux événements A et B peuvent être indépendants sous une certaine probabilité $\mathbb{P}$, et ne plus l'être sous une autre. On a discuté ce point dans le livre [L1]. En particulier, si C est un troisième événement de probabilité non nulle, alors le fait que A et B soient indépendants, i.e. $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, n'implique pas que A et B soient conditionnellement indépendants, i.e. on n'a pas forcément $\mathbb{P}(A \cap B \,|\, C) = \mathbb{P}(A \,|\, C)\mathbb{P}(A \,|\, C)$, comme le montre le calcul suivant.

Exercice 6.

On jette deux dés. Considérons les événements $A :=$ « la somme est 7 », $B :=$ « le premier dé donne 4 » et $C :=$ « la second dé donne 3 ».

1. Montrer que A et B sont indépendants.
2. Montrer que, sachant C, A et B ne sont pas conditionnellement indépendants.

Solution. L'ensemble fondamental est bien entendu :

$$\Omega := \{(k_1, k_2) \mid k_j \in [\![1, 6]\!] \text{ pour } j = 1, 2\} = [\![1, 6]\!]^2\,.$$

Les événements A, B et C sont les sous-ensembles de Ω donnés par :

$$\begin{aligned}
A &= \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}\,,\\
B &= \{(4,6), (4,5), (4,4), (4,3), (4,2), (4,1)\}\,,\\
C &= \{(1,3), (2,3), (3,3), (4,3), (5,3), (6,3)\}\,.
\end{aligned}$$

Ainsi $\operatorname{card}(A) = 6$, $\operatorname{card}(B) = 6$ et $\operatorname{card}(C) = 6$. En faisant l'hypothèse d'équiprobabilité, on calcule aisément $\mathbb{P}(A) = 1/6$, $\mathbb{P}(B) = 1/6$, $\mathbb{P}(C) = 1/6$. De plus $\operatorname{card}(A \cap B) = 1$, $\operatorname{card}(A \cap C) = 1$, $\operatorname{card}(B \cap C) = 1$, et $\operatorname{card}(A \cap B \cap C) = 1$. Donc :

$$\mathbb{P}(A \cap B) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = \mathbb{P}(A)\mathbb{P}(B)\,,$$

ce qui montre que A et B sont indépendants. D'autre part :

$$\begin{aligned}
\mathbb{P}(A \cap B \,|\, C) &= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} = \frac{\operatorname{card}(A \cap B \cap C)}{\operatorname{card}(C)} = \frac{1}{6}\,,\\
\mathbb{P}(A \,|\, C) &= \frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)} = \frac{\operatorname{card}(A \cap C)}{\operatorname{card}(C)} = \frac{1}{6}\,,\\
\mathbb{P}(B \,|\, C) &= \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)} = \frac{\operatorname{card}(B \cap C)}{\operatorname{card}(C)} = \frac{1}{6}\,.
\end{aligned}$$

On vérifie alors que :

$$\mathbb{P}(A \cap B \,|\, C) = \frac{1}{6} \neq \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = \mathbb{P}(A \,|\, C)\mathbb{P}(B \,|\, C)\,,$$

ce qui montre que A et B ne sont pas conditionnellement indépendants.

4.2 Indépendance de plusieurs événements

On prolonge la notion d'indépendance de deux événements au cas des suites d'événements.

Définition 5. On dit que les événements $(A_i)_{1 \in [\![1,n]\!]}$ sont indépendants deux à deux si pour tout $(i,j) \in [\![1,n]\!]^2$ et $i \neq j$, A_i et A_j sont indépendants.

L'exemple suivant montre que cette définition est trop faible pour prolonger la notion d'indépendance aux suites d'événements.

Exemple. Considérons les dates d'anniversaire de trois personnes et les événements $A :=$ « 1 et 2 ont le même anniversaire », $B :=$ « 1 et 3 ont le même anniversaire » et $C :=$ « 2 et 3 ont le même anniversaire ». Évidemment $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/365$ et l'on a aussi $\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = 1/365^2$. Donc A, B, et C sont indépendants deux à deux. Par contre, si deux événements sont réalisés, par exemple A et B, le troisième événement C est sûrement réalisé, car les trois personnes ont toutes le même anniversaire. Par conséquent C n'est pas indépendant de la conjonction de A et B :

$$\mathbb{P}((A \cap B) \cap C) = \mathbb{P}(A \cap B) = \frac{1}{365^2} \neq \frac{1}{365^3} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \mathbb{P}(A \cap B)\mathbb{P}(C)\,.$$

Pour prolonger notre notion d'indépendance aux cas de trois événements on devra imposer une condition supplémentaire non comprise dans la définition de l'indépendance deux à deux, à savoir :

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)\,.$$

Pour une suite de n événements on a la définition suivante.

Définition 6. On dit que les événements $(A_i)_{1\in[\![1,n]\!]}$ sont *indépendants* (ou *mutuellement indépendants*) si pour tout $k \in [\![2,n]\!]$ et pour toute sous-famille d'indices à k éléments $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$, on a :

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) .$$

On a déjà vu que n ($n > 2$) événements peuvent être indépendants deux à deux sans être indépendants mutuellement. Inversement, si n événements sont mutuellement indépendants, alors ils sont indépendants deux à deux.
Enfin, le nombre de conditions imposées dans la définition 6 est égal à :

$$\binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n} = 2^n - \binom{n}{1} - \binom{n}{0} = 2^n - n - 1 .$$

Proposition 11.
Soit $(A_i)_{i\in[\![1,n]\!]}$ une suite d'événements mutuellement indépendants.
1) Pour chaque $i \in [\![1,n]\!]$, posons $B_i = A_i$ ou $B_i = A_i^c$. Alors $(B_i)_{i\in[\![1,n]\!]}$ est une suite d'événements mutuellement indépendants.
2) Toute sous-famille de $(A_i)_{i\in[\![1,n]\!]}$ est constituée d'événements mutuellement indépendants.

Démonstration. Considérons le 1). On va montrer par récurrence sur $k = [\![0,n]\!]$ l'assertion :
(H_k) Pour tout sous-ensemble d'indices $J' \subset [\![1,n]\!]$, pour tout sous-ensemble d'indices $J \subset [\![1,n]\!]$ avec $J' \cap J = \varnothing$ et $\operatorname{card}(J) = k$, on a :

$$\mathbb{P}\left(\Big(\bigcap_{j\in J} A_j^c\Big) \cap \Big(\bigcap_{j\in J'} A_j\Big)\right) = \prod_{j\in J} \mathbb{P}(A_j^c) \prod_{j\in J'} \mathbb{P}(A_j) .$$

(H_0) est immédiat car $(A_j)_{j\in[\![1,n]\!]}$ est une famille mutuellement indépendante.

Supposons (H_k), avec $k \in [\![0,n-1]\!]$.
Soient deux sous-ensembles d'indices J et J' avec $J' \cap J = \varnothing$ et $\operatorname{card}(J) = k+1$. On découpe J en deux : $J = J_k \cup \{j_{k+1}\}$, avec $\operatorname{card}(J_k) = k$. On pose :

$$C_k = \Big(\bigcap_{j\in J_k} A_j^c\Big) \bigcap \Big(\bigcap_{j\in J'} A_j\Big).$$

Comme $C_k \cap A_{j_{k+1}}^c$ et $C_k \cap A_{j_{k+1}}$ forment une partition de C_k, on a :

$$\mathbb{P}(C_k \cap A_{j_{k+1}}^c) = \mathbb{P}(C_k) - \mathbb{P}(C_k \cap A_{j_{k+1}}) . \tag{7}$$

C_k est une intersection d'un certain nombre d'événements A_i et de k événements A_i^c. De même pour $C_k \cap A_{j_{k+1}}$. On peut donc appliquer l'hypothèse de récurrence aux deux termes

du membre de droite de (7) :

$$\mathbb{P}(C_k \cap A^c_{j_{k+1}}) = \prod_{j\in J_k} \mathbb{P}(A^c_j) \prod_{j\in J'} \mathbb{P}(A_j) - \prod_{j\in J_k} \mathbb{P}(A^c_j) \prod_{j\in J'} \mathbb{P}(A_j)\mathbb{P}(A_{j_{k+1}})\,. \tag{8}$$

Or on a :

$$\Big(\bigcap_{j\in J} A^c_j\Big) \bigcap \Big(\bigcap_{j\in J'} A_j\Big) = C_k \cap A^c_{j_{k+1}}\,.$$

En regroupant les termes du membre de droite de (8) :

$$\mathbb{P}\left(\Big(\bigcap_{j\in J} A^c_j\Big) \bigcap \Big(\bigcap_{j\in J'} A_j\Big)\right) = \prod_{j\in J_k} \mathbb{P}(A^c_j)(1-\mathbb{P}(A_{j_{k+1}})) \prod_{j\in J'} \mathbb{P}(A_j)\,,$$

et comme $1-\mathbb{P}(A_{j_{k+1}}) = \mathbb{P}(A^c_{j_{k+1}})$ et $J = J_k \cup \{j_{k+1}\}$ la récurrence est démontrée. Cela prouve le 1).

Le 2) est trivial car toute sous-suite d'une sous-famille est une sous-suite de la famille. ■

Enfin, on généralise la définition d'indépendance pour une suite infinie d'événements.

Définition 7. Soit $(A_i)_{i\in\mathbb{N}}$ une suite d'événements définis sur un même espace de probabilité. On dit que les événements sont indépendants si, pour toute sous-famille d'indices finie $I \subset \mathbb{N}$, la famille $(A_i)_{i\in I}$ est indépendante au sens de la définition 6.

4.3 Construction d'un espace de probabilité

Considérons la situation suivante : on lance n fois une pièce de monnaie mal équilibrée. L'espace des résultats possibles du jeu est $\Omega := \{P, F\}^n$, où P signifie « pile » et F « face ». Notons P_k l'événement « le $k^{ème}$ lancer donne pile » et F_k l'événement « le $k^{ème}$ lancer donne face ». On montre ci-dessous qu'il existe une unique probabilité sur $\big(\Omega, \mathcal{P}(\Omega)\big)$ telle que :

1. pour tout k la probabilité de P_k est égale à p et celle de F_k est égale à $1-p$ avec $p \in [0,1]$.
2. les résultats des différents lancers sont indépendants.

Cette probabilité est donnée par :

$$\mathbb{P}(\{\omega\}) = p^{\pi(\omega)}(1-p)^{n-\pi(\omega)}\,. \tag{9}$$

où $\pi(\omega) :=$ le nombre de piles dans ω. On verra dans le module suivant le schéma de Bernoulli, qui permet d'obtenir un résultat analogue dans le cas d'une suite de lancers infinie dénombrable.

Démonstration. Un élément de Ω est un n-uplet $\omega := (\omega_1, \ldots, \omega_n)$ où ω_i est P ou F. On note pour tout ω :

$$\Pi(\omega) := \{i \in [\![1,n]\!] \mid \omega_i = P\}\,, \tag{10}$$

i.e. les indices i pour lesquels $\omega_i = P$. On a alors :

$$\{\omega\} = \Big(\bigcap_{i\in\Pi(\omega)} P_i\Big) \bigcap \Big(\bigcap_{i\in\Pi(\omega)^c} F_i\Big)\,.$$

où l'on a noté $\Pi(\omega)^c$ le complémentaire de $\Pi(\omega)$ dans $[\![1,n]\!]$, i.e. les indices i pour lesquels $\omega_i = F$. On introduit aussi :

$$\pi(\omega) := \operatorname{card}\big(\Pi(\omega)\big)\,. \tag{11}$$

Supposons qu'il existe une probabilité $\mathbb{P}$ sur $\big(\Omega, \mathcal{P}(\Omega)\big)$ vérifiant les deux hypothèses 1 et 2. L'hypothèse 2 (indépendance) implique que :

$$\mathbb{P}(\{\omega\}) = \prod_{i\in\Pi(\omega)} \mathbb{P}(P_i) \times \prod_{i\in\Pi(\omega)^c} \mathbb{P}(F_i)\,,$$

et l'hypothèse 1 impose alors (9). Cela montre qu'une probabilité $\mathbb{P}$ vérifiant 1 et 2 vérifie nécessairement (9), et d'après la proposition 3(ii) de la page 674, la probabilité $\mathbb{P}$ est ainsi entièrement caractérisée. Cela prouve l'unicité.

Pour que $\mathbb{P}$ définie par (9) réponde à la question, il faut d'abord montrer que $\mathbb{P}$ est une probabilité. En vertu de la proposition 3(i) de la page 674, il faut vérifier que $\sum\limits_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = 1$. Or :

$$\sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = \sum_{i=0}^{n} \sum_{\omega|\pi(\omega)=i} \mathbb{P}(\{\omega\}) = \sum_{i=0}^{n} p^i(1-p)^{n-i} \operatorname{card}\left(\{\omega \mid \pi(\omega) = i\}\right)\,.$$

$\operatorname{card}\left(\{\omega \mid \pi(\omega) = i\}\right)$ est le nombre de sous-ensembles de i éléments que l'on peut choisir dans $[\![1,n]\!]$, donc ce nombre est $\binom{n}{i}$. On obtient alors :

$$\sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = \sum_{i=0}^{n} \binom{n}{i} p^i(1-p)^{n-i}\,,$$

et d'après la formule du binôme de Newton, cette somme vaut 1. Donc $\mathbb{P}$ définie par (9) est une probabilité.

Soit $k \in [\![1,n]\!]$. La probabilité de l'événement P_k est :

$$\mathbb{P}(P_k) = \sum_{\omega|\omega_k=P} \mathbb{P}(\{\omega\}) = \sum_{i=0}^{n} p^i(1-p)^{n-i} \operatorname{card}\left(\{\omega \mid \pi(\omega) = i,\, \omega_k = P\}\right)\,.$$

Regardons l'ensemble $\{\omega \mid \pi(\omega) = i,\, \omega_k = P\}$. Comme on impose que $\omega_k = P$, il reste $n-1$ coordonnées ω_i libres, et l'on doit choisir parmi ces $n-1$ coordonnées les $i-1$ qui sont égales à P. Donc le cardinal de cet ensemble est $\binom{n-1}{i-1}$ et :

$$\mathbb{P}(P_k) = \sum_{i=1}^{n} p^i(1-p)^{n-i}\binom{n-1}{i-1} = p\sum_{j=0}^{n-1} p^j(1-p)^{n-1-j}\binom{n-1}{j} = p\,,$$

ce qui montre que $\mathbb{P}$ vérifie l'hypothèse 1.

Considérons maintenant un sous-ensemble d'indices $J \subset [\![1,n]\!]$ et calculons la probabilité de la conjonction d'événements $\bigcap_{k\in J} P_k$:

$$\begin{aligned}\mathbb{P}\Big(\bigcap_{k\in J} P_k\Big) &= \sum_{\omega|\omega_k=P\,\forall k\in J} \mathbb{P}(\{\omega\}) \\ &= \sum_{i=0}^{n} p^i(1-p)^{n-i}\,\mathrm{card}\,\left(\{\omega \mid \pi(\omega)=i,\, \omega_k=P\ \forall k\in J\}\right)\,.\end{aligned}$$

En notant $j = \mathrm{card}(J)$, le cardinal de l'ensemble qui intervient ici est nul pour $i < j$ et vaut $\binom{n-j}{i-j}$ pour $j \leqslant i \leqslant n$. Donc :

$$\mathbb{P}\Big(\bigcap_{k\in J} P_k\Big) = \sum_{i=j}^{n} p^i(1-p)^{n-i}\binom{n-j}{i-j} = p^j\sum_{l=0}^{n-j} p^l(1-p)^{n-j-l}\binom{n-j}{l} = p^j = \prod_{k\in J}\mathbb{P}(P_k)\,.$$

Cela montre que $(P_i)_{i\in[\![1,n]\!]}$ est une famille d'événements mutuellement indépendants sous la probabilité $\mathbb{P}$. En utilisant la proposition 11, on obtient que $\mathbb{P}$ vérifie l'hypothèse 2. ■

4.4 Probabilité de réunions d'événements indépendants

Soit $(A_i)_{i\in[\![1,n]\!]}$ une suite d'événements mutuellement indépendants. Il y a une formule simple pour calculer la probabilité de $A_1 \cup A_2 \cup \cdots \cup A_n$.

Proposition 12. Si $(A_i)_{i\in[\![1,n]\!]}$ est une suite d'événements mutuellement indépendants, alors :

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = 1 - [1-\mathbb{P}(A_1)][1-\mathbb{P}(A_2)]\cdots[1-\mathbb{P}(A_n)]\,.$$

Démonstration. On passe au complémentaire de $A_1 \cup A_2 \cup \cdots \cup A_n$ et l'on applique la règle de De Morgan $(\bigcup_i A_i)^c = \bigcap_i A_i^c$. Par la proposition 11 les événements A_i^c sont mutuellement indépendants. Donc :

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) &= 1 - \mathbb{P}((A_1 \cup A_2 \cup \cdots \cup A_n)^c) \\ &= 1 - \mathbb{P}(A_1^c \cap A_2^c \cap \cdots \cap A_n^c) \\ &= 1 - \mathbb{P}(A_1^c)\mathbb{P}(A_2^c)\cdots\mathbb{P}(A_n^c)\,,\end{aligned}$$

et l'on conclut en utilisant le fait que $\mathbb{P}(A_i^c) = 1 - \mathbb{P}(A_i)$. ■

Exercice 7.

On suppose qu'une personne est soumise n fois à un risque p d'accident où $0 < p < 1$. Quelle est la probabilité d'avoir au moins un accident ?

Solution. Pour répondre à cette question on pose A_i = « la $i^{ème}$ expérience provoque un accident » et l'on suppose que les A_i sont mutuellement indépendants. Avec $p = \mathbb{P}(A_i)$ on a :

$$\mathbb{P}(\text{« au moins un accident »}) = \mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = 1 - (1-p)^n\,.$$

- Lorsque n tend vers $+\infty$ et p est fixé, cette probabilité tend vers 1 car $1-p \in]0,1[$.
- Lorsque n tend vers $+\infty$ et $p = \lambda/n$ pour un λ fixé, alors :
$$\mathbb{P}(\text{« au moins un accident »}) = 1 - (1-\lambda/n)^n,$$
qui tend vers $1-\exp(-\lambda)$ quand $n \to +\infty$. Donc la probabilité qu'il n'arrive aucun accident est environ :
$$\mathbb{P}(\text{« pas d'accident »}) \simeq \exp(-\lambda)\,,$$
si n est très grand.

Tennis. Deux joueurs, que nous appellerons Arthur et Bernard, se livrent à une partie de tennis. On considère les règles usuelles du tennis lors d'un jeu : le score de chaque joueur passe par 0, 15, 30, 40 à chaque point remporté ; si un des joueurs dépasse 40 alors que l'autre est en dessous de 30, il est déclaré gagnant du jeu ; si les deux joueurs atteignent 40/40, on applique la règle d'égalité-avantage qui veut qu'un joueur est déclaré gagnant du jeu aussitôt qu'il acquiert deux points d'avance. La probabilité de gagner un point par Arthur est p. Tous les points sont gagnés ou perdus de manière indépendante.
On suppose que les deux joueurs sont arrivés à égalité (40/40 ou au-delà). Quelle est la probabilité qu'Arthur remporte le jeu ?
On note R l'événement « Arthur gagne le jeu » et, pour $i \geqslant 1$, W_i est l'événement « Arthur gagne le $i^{ème}$ point » (à partir de l'instant où les deux joueurs sont à égalité). On utilise la formule des probabilités totales avec le système complet formé de $W_1 \cap W_2$, $W_1 \cap W_2^c$, $W_1^c \cap W_2$ et $W_1^c \cap W_2^c$:
$$\begin{aligned}\mathbb{P}(R) &= \mathbb{P}(R\,|\,W_1 \cap W_2)\mathbb{P}(W_1 \cap W_2) + \mathbb{P}(R\,|\,W_1 \cap W_2^c)\mathbb{P}(W_1 \cap W_2^c)\\ &\quad + \mathbb{P}(R\,|\,W_1^c \cap W_2)\mathbb{P}(W_1^c \cap W_2) + \mathbb{P}(R\,|\,W_1^c \cap W_2^c)\mathbb{P}(W_1^c \cap W_2^c)\,.\end{aligned}$$
On calcule les quatre probabilités conditionnelles. Si Arthur gagne les deux premiers points, il gagne le jeu, donc :
$$\mathbb{P}(R\,|\,W_1 \cap W_2) = 1\,.$$
Si Arthur gagne le premier point et perd le second, on revient à égalité, donc :
$$\mathbb{P}(R\,|\,W_1 \cap W_2^c) = \mathbb{P}(R)\,.$$
De même si Arthur perd le premier point et gagne le second :
$$\mathbb{P}(R\,|\,W_1 \cap W_2^c) = \mathbb{P}(R)\,.$$
Enfin, si Arthur perd les deux premiers points, alors il perd le jeu, donc :
$$\mathbb{P}(R\,|\,W_1^c \cap W_2^c) = 0\,.$$
Au total :
$$\mathbb{P}(R) = 1 \times \mathbb{P}(W_1 \cap W_2) + \mathbb{P}(R)\mathbb{P}(W_1 \cap W_2^c) + \mathbb{P}(R)\mathbb{P}(W_1^c \cap W_2) + 0 \times \mathbb{P}(W_1^c \cap W_2^c)\,.$$
Les événements W_1 et W_2 sont indépendants. En utilisant la proposition 11, on peut calculer les quatre probabilités impliquant W_1 et W_2, et l'on obtient :
$$\mathbb{P}(R) = p^2 + 2\mathbb{P}(R)p(1-p)\,.$$

Il n'y a plus qu'à résoudre cette équation linéaire :

$$\mathbb{P}(R) = \frac{p^2}{1 - 2p(1-p)} \cdot \tag{12}$$

Tennis (suite). On reprend l'exemple précédent et l'on considère le jeu à son début 0/0. Quelle est la probabilité qu'Arthur gagne le jeu ?
On note G l'événement « Arthur gagne le jeu », A_k l'événement « Arthur gagne le jeu et Bernard gagne k points » et A_e l'événement « Arthur gagne le jeu après égalité 40/40 ». D'après les règles du tennis, le quadruplet $\{A_0, A_1, A_2, A_e\}$ est une partition de l'événement G, on a donc :

$$\mathbb{P}(G) = \mathbb{P}(A_0) + \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_e)\,.$$

On calcule les quatre probabilités :
1) A_0 arrive seulement si Arthur gagne les quatre premiers points :

$$\mathbb{P}(A_0) = p^4\,.$$

2) A_1 arrive si le jeu se déroule en cinq points et si Bernard en gagne un sur les quatre premiers. Il y a $\binom{4}{1} = 4$ possibilités pour Bernard de gagner ce point, donc :

$$\mathbb{P}(A_1) = 4(1-p)p^4\,.$$

2) A_2 arrive si le jeu se déroule en six points et si Bernard en gagne deux avant le sixième et dernier point. Il y a $\binom{5}{2} = 10$ possibilités pour Bernard de gagner ces deux points, donc :

$$\mathbb{P}(A_2) = 10(1-p)^2p^4\,.$$

3) On note E l'événement « les deux joueurs arrivent à 40/40 ». On a :

$$\mathbb{P}(A_e) = \mathbb{P}(A_e \,|\, E)\mathbb{P}(E)\,.$$

L'événement E se produit si Arthur et Bernard gagnent chacun trois points sur les six premiers points du jeu :

$$\mathbb{P}(E) = \binom{6}{3}p^3(1-p)^3 = 20p^3(1-p)^3\,.$$

La probabilité conditionnelle $\mathbb{P}(A_e \,|\, E)$ a été calculée dans l'exemple précédent. On obtient alors :

$$\mathbb{P}(A_e) = \frac{20p^5(1-p)^3}{1-2p(1-p)}\,,$$

et au total :

$$\mathbb{P}(G) = p^4 + 4(1-p)p^4 + 10(1-p)^2p^4 + \frac{20p^5(1-p)^3}{1-2p(1-p)}\,.$$

On peut tracer cette probabilité en fonction de p (figure IV.1.2 de la page ci-contre). On constate qu'elle présente une transition assez raide de valeurs proches de 0 lorsque $p < 1/2$ à des valeurs proches de 1 lorsque $p > 1/2$. Par exemple, si Arthur est légèrement meilleur que Bernard et a $60\ \%$ [respectivement $70\ \%$] de chances de gagner un point, alors sa

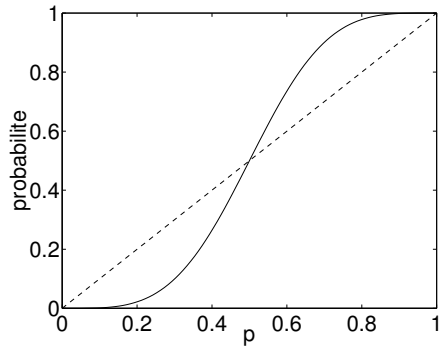


FIGURE IV.1.2 – Probabilité de gagner un jeu
en fonction de la probabilité de gagner un point.

probabilité de gagner un jeu est de $74\ \%$ [respectivement $90\ \%$]. C'est tout l'objectif des règles du tennis : pouvoir distinguer les bons joueurs des mauvais joueurs. Cet effet est encore renforcé par la structure « jeu-set-match ».

Exemple. On considère une pièce biaisée qui a une probabilité p de donner pile et $1-p$ de donner face.
1) On lance la pièce jusqu'à obtenir un pile. On note A_n l'événement « n lancers sont nécessaires pour obtenir un pile ». Trouver la probabilité de l'événement A_n et vérifier que l'événement $A :=$ « on obtient un pile un jour » est de probabilité 1.
On a déjà obtenu ce résultat, mais on va le dériver d'une manière systématique qui peut être adaptée à des situations plus complexes (voir la suite). On note P l'événement « on obtient pile au premier lancer ». On utilise la formule des probabilités totales :

$$\begin{aligned}\mathbb{P}(A_n) &= \mathbb{P}(A_n \,|\, P)\mathbb{P}(P) + \mathbb{P}(A_n \,|\, P^c)\mathbb{P}(P^c)\\ &= \begin{cases} p\,, & \text{si } n = 1\,,\\ 0 + (1-p)\mathbb{P}(A_{n-1})\,, & \text{si } n \geqslant 2\,,\end{cases}\end{aligned}$$

et donc les probabilités $p_n = \mathbb{P}(A_n)$ satisfont la relation de récurrence :

$$p_n = (1-p)p_{n-1}\,,\ \ n \geqslant 2\,, \qquad p_1 = p\,.$$

Cette relation se résout immédiatement et donne $p_n = (1-p)^{n-1}p$ pour tout $n \geqslant 1$. Comme $A = \bigcup_{n\geqslant 1} A_n$ et que les événements A_n sont deux à deux disjoints, on vérifie que :

$$\mathbb{P}(A) = \sum_{n=1}^{+\infty} \mathbb{P}(A_n) = \sum_{n=1}^{+\infty} (1-p)^{n-1}p = 1\,.$$

2) On lance la pièce jusqu'à obtenir deux piles de suite. On note B_n l'événement « n lancers sont nécessaires pour obtenir deux piles de suite ». Trouver la probabilité de l'événement B_n et vérifier que l'événement $B :=$ « on obtient une suite de deux piles un jour » est de probabilité 1.
La stratégie est la même qu'au 1). Il faut trouver un système complet qui permette d'appliquer la formule des probabilités totales pour obtenir une relation de récurrence. Un bon

système complet est ici $\{F, PF, PP\}$ où F est l'événement « on obtient face au premier lancer », PF est l'événement « on obtient pile au premier lancer et face au second » et PP est l'événement « on obtient pile aux deux premiers lancers ». On peut alors écrire que :

$$\begin{aligned}\mathbb{P}(B_n) &= \mathbb{P}(B_n\,|\,F)\mathbb{P}(F) + \mathbb{P}(B_n\,|\,PF)\mathbb{P}(PF) + \mathbb{P}(B_n\,|\,PP)\mathbb{P}(PP)\\ &= \begin{cases} 0\,, & \text{si } n = 1\,,\\ p^2\,, & \text{si } n = 2\,,\\ \mathbb{P}(B_{n-1})(1-p) + \mathbb{P}(B_{n-2})p(1-p) + 0\,, & \text{si } n \geqslant 3\,,\end{cases}\end{aligned}$$

et donc les probabilités $q_n := \mathbb{P}(B_n)$ satisfont la relation de récurrence :

$$\begin{aligned}&q_n = (1-p)q_{n-1} + p(1-p)q_{n-2}\,, \quad n \geqslant 3\,,\\ &q_1 = 0\,, \quad q_2 = p^2\,.\end{aligned}$$

Il arrive souvent que l'on ait à résoudre de telles relations de récurrence en probabilité. On discute la méthode de résolution générale dans le complément de la page 701. On commence par déterminer les racines du polynôme associé :

$$P(X) = X^2 - (1-p)X - p(1-p)\,,$$

qui sont explicitement :

$$\lambda_1 = \frac{(1-p) - \sqrt{(1-p)(1+3p)}}{2}\,, \quad \lambda_2 = \frac{(1-p) + \sqrt{(1-p)(1+3p)}}{2}\,.$$

Ces deux racines sont distinctes si $p \in [0, 1[$. La probabilité q_n est une combinaison linéaire des deux suites géométriques de raisons λ_1 et λ_2 :

$$q_n = c_1\lambda_1^{n-1} + c_2\lambda_2^{n-1}\,, \quad n \geqslant 1\,,$$

avec les constantes c_1 et c_2 déterminées par les conditions $q_1 = 0$ and $q_2 = p^2$:

$$c_1 = -\frac{p^2}{\sqrt{(1-p)(1+3p)}}\,, \quad c_2 = \frac{p^2}{\sqrt{(1-p)(1+3p)}}\,.$$

Enfin, on peut calculer la probabilité de l'événement B :

$$\mathbb{P}(B) = \sum_{n=2}^{+\infty} q_n = \frac{c_1\lambda_1}{1-\lambda_1} + \frac{c_2\lambda_2}{1-\lambda_2}\,,$$

qui vaut bien 1 après calcul et réduction au même dénominateur.

5 Problèmes et paradoxes

5.1 Un exemple classique : la ruine du joueur

Arthur arrive dans un casino avec un certaine somme $a_0 \in \mathbb{N}^*$ en poche. Il décide de jouer à un certain jeu, la roulette par exemple. À chaque partie, il mise la même somme égale à $k \in \mathbb{N}^*$. S'il gagne la partie, il récupère sa mise doublée. S'il perd la partie, il perd sa mise. La probabilité qu'a Arthur de gagner une partie est $p \in]0, 1]$, et les parties sont gagnées

ou perdues de manière indépendante. Arthur se fixe comme objectif de gagner $a_0 + b_0$, $b_0 \in \mathbb{N}^*$, et de repartir avec cette somme lorsqu'il aura atteint cet objectif. La grande question est alors : est-ce qu'Arthur atteindra cet objectif avant d'être ruiné ? En effet, si la fortune d'Arthur atteint 0, alors il ne peut plus jouer et il rentre bredouille chez lui (on dit qu'il est ruiné, d'où le nom du problème).

On va d'abord considérer pour simplifier que la mise est de $k = 1$. On introduit $P(a)$ la probabilité qu'Arthur atteigne son objectif sachant qu'il a la somme a en poche. On introduit une partition de l'espace de probabilités déterminée par le fait qu'Arthur gagne ou non la première partie. S'il gagne la première partie, alors Arthur se retrouve avec $a+1$ en poche et la probabilité conditionnelle qu'il atteigne son objectif est $P(a+1)$. S'il perd la première partie, alors Arthur se retrouve avec $a-1$ en poche et la probabilité conditionnelle qu'il atteigne son objectif est $P(a-1)$. En utilisant la formule des probabilités totales, on obtient :

$$P(a) = P(a+1)p + P(a-1)(1-p)\,.$$

De plus, on sait que si Arthur se retrouve avec 0 en poche, alors il est ruiné et il ne peut plus atteindre son objectif :

$$P(0) = 0\,.$$

Enfin, si Arthur se retrouve avec $a_0 + b_0$ en poche, alors Arthur a atteint son objectif et donc :

$$P(a_0 + b_0) = 1\,.$$

Il faut donc résoudre une équation aux différences d'ordre 2 :

$$P(a+2) = \frac{1}{p}P(a+1) - \frac{1-p}{p}P(a)\,,$$

avec deux conditions aux bords en $a = 0$ et en $a = a_0 + b_0$. On recherche la valeur de $P(a_0)$ correspondant à la fortune initiale d'Arthur.

On considère d'abord ici que $p \neq 1/2$.

On étudie le polynôme du second degré $X^2 - (1/p)X + (1-p)/p$ qui possède deux racines distinctes $\lambda_1 = 1$ et $\lambda_2 = (1-p)/p \neq 1$. On sait d'après le complément de la page 701 que la solution générale de l'équation aux différences est une combinaison linéaire des fonctions $a \mapsto \lambda_1^a$ et $a \mapsto \lambda_2^a$. Les deux coefficients de cette combinaison linéaire sont fixés par les conditions aux bords et l'on obtient :

$$P(a) = \frac{1-\lambda_2^a}{1-\lambda_2^{a_0+b_0}}\,.$$

Dans le cas où $p = 1/2$, le polynôme du second degré $X^2 - 2X + 1$ possède une racine double $\lambda_1 = 1$, donc la solution est une fonction de la forme $c_1 + c_2 a$. Les deux coefficients c_1 et c_2 sont fixés par les conditions aux bords et l'on obtient :

$$P(a) = \frac{a}{a_0 + b_0}\,.$$

Application 1 : jeu de roulette idéal $p = 1/2$, avec $a_0 = b_0 = 100$.

On a alors $P(a_0) = 1/2 = 50\ \%$.

Application 2 : jeu de roulette réel $p = 18/37$ (il y a un zéro dans une roulette de casino, et le croupier considère que zéro n'est ni pair, ni impair...), avec $a_0 = b_0 = 100$. On a alors $\lambda_2 = (1-p)/p = 19/18$ et donc :

$$P(a_0) = \frac{1-(19/18)^{100}}{1-(19/18)^{200}} = \frac{1}{1+(19/18)^{100}} \simeq 0{,}45\ \%\,.$$

Autrement dit, vous pouvez parier votre chemise qu'Arthur va être ruiné.
Considérons maintenant le cas où Arthur mise la somme k à chaque partie. Pour ne pas compliquer le problème avec des parties entières de nombres réels, on va supposer que a_0 et b_0 sont des multiples de k. En reprenant les mêmes arguments (il suffit de diviser tous les nombres par k pour se retrouver dans la situation précédente où $k = 1$), on trouve dans le cas $p = 1/2$ le même résultat : la probabilité qu'Arthur atteigne son objectif est :

$$P_k(a_0) = \frac{a_0}{a_0 + b_0}\,.$$

Par contre, dans le cas $p \neq 1/2$, on trouve que cette probabilité vaut :

$$P_k(a_0) = \frac{1-\lambda_2^{a_0/k}}{1-\lambda_2^{(a_0+b_0)/k}}\,.$$

Application 3 : jeu de roulette réel $p = 18/37$, avec $a_0 = b_0 = 100$. On a alors $P_1(a_0) \simeq 0{,}45\ \%$, $P_2(a_0) \simeq 6{,}28\ \%$, $P_{10}(a_0) \simeq 36{,}80\ \%$, et $P_{100}(a_0) \simeq 48{,}65\ \%$.
Ainsi, dans un jeu qui, comme la roulette, vous est défavorable, même légèrement, il faut jouer le tout pour le tout en un coup si vous voulez maximiser vos chances de gain. Évidemment, c'est moins drôle...

5.2 Paradoxes

Ce paragraphe recense un certain nombre de problèmes qui, sous une forme ou une autre, reviennent périodiquement. On peut appliquer sur ces problèmes des raisonnements intuitifs très simples, mais qui conduisent à des réponses absurdes. On va considérer ici ces problèmes et les raisonnements faux qui leur sont souvent appliqués, puis on va montrer pourquoi ces raisonnements sont faux, et enfin on donnera une solution juste.
Paradoxe de Galton (1894). Vous lancez trois pièces de monnaie. Comme il n'y a que deux résultats possibles, pile ou face, on est sûr qu'au moins deux pièces donnent le même résultat. La troisième pièce a une chance sur deux d'être pile ou face, donc, au final, il y a une chance sur deux que les trois pièces donnent le même résultat.
Ce raisonnement ne peut pas être vrai. La probabilité que les trois pièces soient égales est $\mathbb{P}(\text{ trois piles }) + \mathbb{P}(\text{ trois faces }) = 1/8 + 1/8 = 1/4$. Alors qu'est-ce qui cloche dans le raisonnement précédent ?
L'erreur réside essentiellement dans la construction de l'espace de probabilité sous-jacent. La « troisième » pièce n'est pas identifiée initialement dans l'espace de probabilité $\Omega := \{F, P\}^3$, elle est déterminée par les deux autres. La probabilité que la « troisième » pièce soit pile ou face est en fait une probabilité conditionnelle. On peut faire un calcul juste de la manière suivante. La probabilité de l'événement $A :=$ « deux pièces au

moins donnent pile » est :

$$\mathbb{P}(A) = \mathbb{P}(\text{ trois piles }) + \mathbb{P}(\text{ deux piles, un face }) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}\cdot$$

De même, la probabilité de l'événement $B :=$ « deux pièces au moins donnent face » $= A^c$ est :

$$\mathbb{P}(B) = \mathbb{P}(\text{ trois faces }) + \mathbb{P}(\text{ deux faces, un pile }) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}\cdot$$

On obtient donc :

$$\begin{aligned}
\mathbb{P}(\text{ la troisième pièce est } P \,|\, A) &= \frac{\mathbb{P}(\text{ trois piles })}{\mathbb{P}(A)} = \frac{1}{4}\,,\\
\mathbb{P}(\text{ la troisième pièce est } F \,|\, A) &= 1 - \frac{1}{4} = \frac{3}{4}\,,\\
\mathbb{P}(\text{ la troisième pièce est } F \,|\, B) &= \frac{\mathbb{P}(\text{ trois faces })}{\mathbb{P}(B)} = \frac{1}{4}\,,\\
\mathbb{P}(\text{ la troisième pièce est } P \,|\, B) &= 1 - \frac{1}{4} = \frac{3}{4}\cdot
\end{aligned}$$

Donc on voit qu'il est faux de dire que « la troisième pièce a une chance sur deux d'être pile ou face ».

De la difficulté d'interpréter les résultats d'un test. Un test médical a pour but de comparer deux méthodes de traitement des calculs rénaux (des petites accrétions minérales qui se développent dans les reins) : la méthode M_C est la chirurgie, la méthode M_L est la lithotritie (destruction des calculs par un appareil à ultra-sons). 700 patients ont été traités, 350 par la méthode M_C, 350 par la méthode M_L. 273 patients traités par la méthode M_C ont guéri, 289 patients traités par la méthode M_L ont guéri. En notant R l'événement « le patient guérit », on a donc :

$$\begin{aligned}
\operatorname{card}(M_C \cap R) = 273, \quad \operatorname{card}(M_C \cap R^c) = 77,\\
\operatorname{card}(M_L \cap R) = 289, \quad \operatorname{card}(M_L \cap R^c) = 61.
\end{aligned}$$

Le taux de réussite (R) selon la méthode de traitement est donc :

$$\begin{aligned}
\mathbb{P}(R|\,M_C) &= \frac{\mathbb{P}(M_C \cap R)}{\mathbb{P}(M_C)} = \frac{273}{350} \simeq 78\,\%\,,\\
\mathbb{P}(R|\,M_L) &= \frac{\mathbb{P}(M_L \cap R)}{\mathbb{P}(M_L)} = \frac{289}{350} \simeq 83\,\%\,.
\end{aligned}$$

La chirurgie semble donc moins performante. Cependant, pour chaque intervention on a mesuré la taille des calculs retirés et on les classe en deux catégories : les gros calculs (C_G) ont un diamètre supérieur à 2 cm, les petits calculs (C_P) ont un diamètre inférieur à 2 cm. En classant les patients par taille de calculs, par type d'intervention, et par le succès de l'intervention, on obtient :

	méthode par chirurgie M_C	méthode par lithotritie M_L
gros calculs C_G	191 réussites 71 échecs	60 réussites 26 échecs
petits calculs C_P	82 réussites 6 échecs	229 réussites 35 échecs

et donc les probabilités conditionnelles de succès sont :

	méthode par chirurgie M_C	méthode par lithotritie M_L
gros calculs C_G	$\frac{191}{191+71} \simeq 73\ \%$	$\frac{60}{60+26} \simeq 70\ \%$
petits calculs C_P	$\frac{82}{82+6} \simeq 93\ \%$	$\frac{229}{229+35} \simeq 87\ \%$

Probabilité conditionnelle de succès $\mathbb{P}(R \,|\, M_. \cap C_.)$

On voit donc que, que ce soit pour des patients ayant des petits calculs ou des patients ayant des gros calculs, la chirurgie a de meilleurs résultats. En revanche, quand on considère les mêmes données en ne distinguant pas les groupes, on trouve que la chirurgie a de moins bons résultats. Ce n'est pas contradictoire, et cet apparent paradoxe est souvent rencontré. La vérité est que, lors de ce test, les patients envoyés en chirurgie étaient les patients que les docteurs estimaient les plus difficiles. On avait donc une surreprésentation de patients avec de gros calculs parmi les 350 patients envoyés en chirurgie par rapport à ceux envoyés en lithotritie.

Monty Hall (du nom d'un présentateur de télévision). Lors d'un jeu télévisé, le candidat fait face à trois portes fermées. Une voiture se trouve derrière l'une d'entre elles, et rien derrière les deux autres. Le candidat désigne une des portes, sans l'ouvrir. Le présentateur, qui sait où se trouve la voiture, ouvre alors une des deux autres portes derrière laquelle il n'y a rien. Il offre alors au candidat la possibilité de changer son choix. Le candidat a-t-il intérêt à changer ?

La règle (non dite) qui régit le comportement du présentateur est la suivante : si le candidat a initialement choisi une mauvaise porte, alors le présentateur ouvre la seconde porte derrière laquelle il n'y a rien. Si le candidat a initialement choisi la bonne porte, alors le présentateur choisit au hasard une des deux portes derrière laquelle il n'y a rien.

On note P l'événement « le premier choix du candidat est bon » et G l'événement « le candidat ouvre finalement la bonne porte ». La formule des probabilités totales donne :

$$\mathbb{P}(G) = \mathbb{P}(G \,|\, P)\mathbb{P}(P) + \mathbb{P}(G \,|\, P^c)\mathbb{P}(P^c)\,.$$

Le choix initial du candidat a une chance sur trois d'être le bon, donc $\mathbb{P}(P) = 1/3$ et $\mathbb{P}(P^c) = 2/3$. Les deux probabilités conditionnelles dépendent de la stratégie adoptée par le candidat. On calcule dans la suite la probabilité $\mathbb{P}(G)$ obtenue avec les deux stratégies.

Stratégie 1 : le candidat change son choix initial. Si le candidat avait choisi initialement la bonne porte, alors il perd et donc $\mathbb{P}(G \,|\, P) = 0$. Si le candidat avait choisi initialement une

mauvaise porte, alors en changeant il prend forcément la bonne porte car le présentateur lui a désigné une mauvaise ; donc $\mathbb{P}(G \,|\, P^c) = 1$. Au total, sous la stratégie 1, on a $\mathbb{P}(G) = 2/3$.
Stratégie 2 : le candidat ne change pas son choix initial. On a alors $\mathbb{P}(G) = 1/3$.
On double donc ses chances de gain en changeant d'avis !

6 Complément : équations aux différences

À plusieurs reprises dans ce module, nous avons rencontré la situation où la quantité cherchée $p(n)$ dépend d'un certain paramètre $n \in \mathbb{N}$ et est solution d'une équation aux différences du type :

$$p(n+k) = \sum_{i=0}^{k-1} a_i p(n+i)\,, \quad n \geqslant 0\,, \tag{13}$$

avec des conditions initiales prescrites pour les k premières valeurs :

$$p(i) = p_i\,, \quad i = 0, \ldots, k-1\,. \tag{14}$$

L'existence et l'unicité de la solution d'un tel système est une conséquence d'un résultat général concernant les suites définies par récurrence (*cf.* le théorème 2 de la page 66 du livre [L1]).
Le cas des équations d'ordre 1 se traite très facilement. Dans ce cas $p(n)$ satisfait $p(n+1) = a_0 p(n)$. On a donc affaire à une suite géométrique et l'expression de $p(n)$ est $p(n) = p_0 a_0^n$.
Dans le cas général, l'analyse d'une équation aux différences d'ordre k commence par l'étude du polynôme P de degré k construit sur l'équation (13) et défini par :

$$P(X) = X^k - a_{k-1}X^{k-1} - \cdots - a_1 X - a_0\,.$$

Si $\lambda \in \mathbb{C}$ est une racine du polynôme P, alors $p(n) = \lambda^n$ est solution de l'équation aux différences (13), car :

$$0 = \lambda^n P(\lambda) = \lambda^{n+k} - \sum_{i=0}^{k-1} a_i \lambda^{n+1} = p(n+k) - \sum_{i=0}^{k-1} a_i p(n+i)\,.$$

L'idée est alors de trouver une combinaison linéaire de solutions de cette forme qui satisfasse les conditions initiales (14).
Considérons l'équation aux différences d'ordre 2 :

$$p(n+2) = a_1 p(n+1) + a_0 p(n)\,, \quad n \geqslant 0\,, \tag{15}$$

avec les conditions initiales :

$$p(0) = p_0\,, \quad p(1) = p_1\,. \tag{16}$$

Le polynôme du second degré $X^2 - a_1 X - a_0$ admet deux racines distinctes λ_1, λ_2 dans $\mathbb{C}$ si $a_1^2 + 4a_0 \neq 0$ et une racine double $\lambda_1 = a_1/2$ si $a_1^2 + 4a_0 = 0$. Dans le premier cas ($a_1^2 + 4a_0 \neq 0$), on trouve que $c_1\lambda_1^n + c_2\lambda_2^n$ satisfait le problème (15-16) si l'on choisit c_1 et c_2 tels que $c_1 + c_2 = p_0$ et $c_1\lambda_1 + c_2\lambda_2 = p_1$ (le déterminant de ce système linéaire est $\lambda_1 - \lambda_2 \neq 0$ donc le couple (c_1, c_2) existe et est unique).

Dans le second cas ($a_1^2+4a_0=0$), on trouve que $(c_1n+c_2)\lambda_1^n$ satisfait le problème (15-16) si l'on choisit c_1 et c_2 tels que $c_2=p_0$ et $(c_1+c_2)\lambda_1=p_1$ (ce qui est possible dès que $\lambda_1\neq 0$).
Pour résumer, on a la proposition suivante :

Proposition 13. Soient $a_0, a_1 \in \mathbb{R}$, non tous deux nuls, et soient $p_0, p_1 \in \mathbb{R}$. Il existe une unique solution $\big(p(n)\big)_{n\in\mathbb{N}}$ à l'équation aux différences (15) avec les conditions initiales (16).
Si $a_1^2+4a_0\neq 0$, alors :

$$p(n)=c_1\lambda_1^n+c_2\lambda_2^n\,,$$

avec $\begin{cases} \lambda_1 = \dfrac{a_1+\sqrt{a_1^2+4a_0}}{2}\,, \qquad \lambda_2=\dfrac{a_1-\sqrt{a_1^2+4a_0}}{2}\,, \\ c_1 = \dfrac{\lambda_2p_0-p_1}{\lambda_2-\lambda_1}\,, \qquad c_2=\dfrac{p_1-\lambda_1p_0}{\lambda_2-\lambda_1}\,. \end{cases}$

Si $a_1^2+4a_0=0$, alors :

$$p(n)=(c_1n+c_2)\lambda_1^n\,,$$

avec $\lambda_1=a_1/2$, $c_1=p_1/\lambda_1-p_0$, et $c_2=p_0$.

Exercice type corrigé

IV.1.0 On jette trois dés. Calculer :

1. la probabilité d'avoir exactement un 6.
2. la probabilité d'obtenir au moins un 6.
3. la probabilité d'obtenir au moins deux faces identiques.

Préciser l'espace de probabilité $(\Omega,\mathcal{A},\mathbb{P})$ choisi et à quelle partie de Ω correspond chaque événement.
Solution. L'ensemble fondamental est $\Omega:=[\![1,6]\!]^3$, dont le cardinal est $6^3=216$. On munit cet ensemble de la tribu $\mathcal{P}(\Omega)$ et de la probabilité uniforme.

1. On note A l'événement « on obtient exactement un six », c'est-à-dire $A=A_1\cup A_2\cup A_3$, où :

$$\begin{aligned} A_j &:= \text{« le } j^{\text{ème}} \text{ dé donne un six, les deux autres sont différents de six »} \\ &= \big\{\omega=(\omega_1,\omega_2,\omega_3)\in\Omega \mid \omega_j=6, \omega_i\neq 6, i\in\{1,2,3\}\setminus\{j\}\big\}. \end{aligned}$$

   On a donc $\operatorname{card}(A_j)=1\times 5\times 5=25$, d'où :

$$\mathbb{P}(A_j)=\frac{\operatorname{card}(A_j)}{\operatorname{card}(\Omega)}=\frac{25}{216}\,,$$

   et comme A est union disjointe des A_j :

$$\mathbb{P}(A)=\mathbb{P}(A_1)+\mathbb{P}(A_2)+\mathbb{P}(A_3)=3\times\frac{25}{216}=\frac{25}{72}\simeq 34{,}72\,\%\,.$$

2. On note B l'événement « on obtient au moins un six ».
   Il est plus facile de raisonner avec le complémentaire $B^c =$ « on n'obtient aucun six ». On a en effet $B^c = \{\omega \in \Omega \mid \omega_j \neq 6, j = 1, 2, 3\}$. Donc $\operatorname{card}(B^c) = 5^3 = 125$ et :
$$\mathbb{P}(B) = 1 - \mathbb{P}(B^c) = 1 - \frac{\operatorname{card}(B^c)}{\operatorname{card}(\Omega)} = 1 - \frac{125}{216} = \frac{91}{216} \simeq 42{,}13\,\%\,.$$
3. On note C l'événement « on obtient au moins deux faces identiques ».
   Il est ici aussi plus facile de raisonner avec le complémentaire (en fait, c'est généralement le cas pour des événements qui font intervenir l'expression « au moins »). Ici $C^c =$ « les trois faces sont distinctes » $= \{\omega = (\omega_1, \omega_2, \omega_3) \in \Omega \mid \omega_j \text{ tous distincts}\}$. Dénombrer C^c est aisé, $\operatorname{card}(C^c) = 6 \times 5 \times 4 = 120$, donc :
$$\mathbb{P}(C) = 1 - \mathbb{P}(C^c) = 1 - \frac{\operatorname{card}(C^c)}{\operatorname{card}(\Omega)} = 1 - \frac{120}{216} = \frac{96}{216} \simeq 44{,}44\,\%\,.$$

Pour aller plus loin

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

IV.1.1 Soient $A_1, \ldots, A_n$ des événements d'un espace Ω muni d'une tribu $\mathcal{A}$. Donner l'écriture ensembliste des événements suivants :

1. Au moins un des A_i est réalisé.
2. Aucun des A_i n'est réalisé.

Expliciter ces événements lorsque $\Omega = \mathbb{R}$ et $A_i = [1/i, 1 + 1/i]$.

IV.1.2 Soit Ω un ensemble non vide.

1. Expliciter les éléments de la tribu engendrée par deux parties A et B de Ω.
2. Expliciter les éléments de la tribu engendrée par une famille finie $(A_1, A_2, \ldots, A_n)$ de parties de Ω.

IV.1.3 Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité, A et B des événements avec $\mathbb{P}(B) > 0$. Montrer que :

1. Si $\mathbb{P}(A) = 1$ alors $\mathbb{P}(B \,|\, A) = \mathbb{P}(B)$ et $\mathbb{P}(A \,|\, B) = 1$.
2. Si $\mathbb{P}(A) = 0$ alors $\mathbb{P}(A \,|\, B) = 0$.

IV.1.4 On considère trois cartes : une avec les deux faces rouges, une avec les deux faces blanches, et une avec une face rouge et une face blanche. On tire une carte au hasard, on expose une face au hasard : elle est rouge.
Quelle est la probabilité que la face cachée soit blanche (on construira d'abord l'espace de probabilité adéquat) ?

IV.1.5 Dans une usine, deux chaînes A et B fabriquent le même composant de manière indépendante. Les composants, issus de la chaîne A [respectivement, B] sont défectueux avec probabilité 2 % [respectivement 4 %]. La chaîne A [respectivement, B] produit 30 [respectivement, 20] composants par jour. On choisit au hasard un composant produit dans la journée.

1. Avec quelle probabilité est-il défectueux ?
2. S'il est défectueux, quelle est la probabilité qu'il ait été produit par la chaîne B ?

IV.1.6 On cherche un parapluie qui avec la probabilité $p/7$ se trouve dans l'un quelconque des sept étages d'un immeuble $(0 \leqslant p \leqslant 1)$. On a exploré en vain les six premiers étages. Quelle est la probabilité que le parapluie se trouve au septième étage ?

IV.1.7 Considérons une urne contenant quatre boules blanches et huit boules rouges. Quelle est la probabilité de la suite « blanc, blanc, rouge » si l'on tire les boules sans remise ?

IV.1.8 On considère n urnes $(n \geqslant 1)$, numérotées de 1 à n. L'urne numérotée k contient k boules blanches et $n-k$ boules noires. On choisit au hasard une urne puis une boule dans cette urne. Quelle est la probabilité d'obtenir une boule blanche ?

IV.1.9 On considère n individus $I_1, \ldots, I_n$. I_1 reçoit une information sous la forme de « oui » ou « non », la transmet à I_2, et ainsi de suite jusqu'à I_n, qui l'annonce au monde. Chaque individu transmet ce qu'il a entendu avec la probabilité p $(0 \leqslant p \leqslant 1)$, et transmet le contraire avec probabilité $1-p$. Les comportements des n individus sont supposés indépendants. Calculer la probabilité p_n que l'information soit fidèlement transmise. Que se passe-t-il quand n tend vers l'infini ?

IV.1.10 Un buveur impénitent décide d'essayer de ne plus boire. S'il ne boit pas un jour donné, alors la probabilité qu'il ne boive pas le lendemain est de $0{,}4$. S'il succombe à la tentation un jour, alors le remords fait que la probabilité qu'il ne boive pas le lendemain monte à $0{,}8$.

1. Quelle est la probabilité que ce buveur ne boive pas le $n^{ème}$ jour ?
2. Que se passe-t-il lorsque n tend vers $+\infty$?

IV.1.11 On classe les gérants de portefeuilles de valeurs boursières en deux catégories : ceux qui sont bien informés et ceux qui ne le sont pas. Lorsqu'un gérant bien informé achète une valeur boursière pour son client, la probabilité que le cours de celle-ci monte est de $0{,}8$; dans le cas d'un gérant mal informé, le cours descend avec une probabilité $0{,}6$. Dans l'annuaire professionnel, un gérant sur 10 est bien informé. Un client choisit au hasard un gérant dans l'annuaire et lui demande d'acheter une valeur :

1. Quelle est la probabilité que le cours de la valeur de l'action monte ?
2. Sachant que le cours de la valeur est monté, quelle est la probabilité que le gérant choisi soit mal informé ?
3. Le gérant achète 4 valeurs V_1, V_2, V_3 et V_4 pour le client. On suppose que les cours des valeurs fluctuent de manière indépendante. Calculer la probabilité que les cours de V_2 et V_4 montent et ceux de V_1 et V_3 descendent. Calculer la probabilité qu'il y ait exactement deux valeurs qui aient monté.

IV.1.12 Alphonse et Bernard tirent au pistolet sur une cible suivant les règles suivantes :

- ils tirent chacun leur tour. Le premier qui atteint la cible a gagné ;
- lorsqu'il tire, Alphonse atteint la cible avec la probabilité a $(0 < a < 1)$ et il la rate avec la probabilité $\bar{a} = 1 - a$;
- lorsqu'il tire, Bernard atteint la cible avec la probabilité b $(0 < b < 1)$ et il la rate avec la probabilité $\bar{b} = 1 - b$;
- Alphonse tire le premier.

Ainsi, Alphonse (respectivement Bernard) n'effectue que des tirs de rang impair (respectivement pair).
On considère, pour tout entier $n \geqslant 1$, les événements A_{2n-1} := « Alphonse gagne à l'issue du tir n° $2n-1$ », B_{2n} := « Bernard gagne à l'issue du tir n° $2n$ », A := « Alphonse gagne » et B := « Bernard gagne ».

1. Calculer, en fonction de a et b les probabilités des événements A_1, B_2 et A_3. Plus généralement, calculer $\mathbb{P}(A_{2n-1})$ et $\mathbb{P}(B_{2n})$.
2. En déduire les probabilités $\mathbb{P}(A)$ et $\mathbb{P}(B)$. Vérifier la relation $\mathbb{P}(A) + \mathbb{P}(B) = 1$.
3. On suppose que $a = 1/3$. Quelle est la valeur de b telle que la partie soit équilibrée (i.e. telle que $\mathbb{P}(A) = \mathbb{P}(B) = 1/2$) ?

IV.1.13 Un examen comporte 15 questions. Chaque question admet trois réponses possibles. Les étudiants répondent à chaque question indépendamment. On suppose que les étudiants ayant préparé l'examen sont en proportion 70 % et répondent correctement à une question avec une probabilité de $0{,}8$. Les 30 % d'étudiants restant choisissent les réponses au hasard. Il faut au moins 8 bonnes réponses pour réussir l'examen.

1. Quelle est la proportion d'étudiants réussissant l'examen ?
2. Si un étudiant échoue, quelle est la probabilité qu'il ait préparé l'examen ?

Variables aléatoires discrètes

IV.2

1 Lois et variables aléatoires discrètes

1.1 Définitions

Dans une expérience aléatoire, on s'intéresse souvent à des quantités particulières dépendant du résultat ω, plutôt qu'au résultat dans son intégralité. Ainsi dans les jeux de hasard, les joueurs se préoccupent plus de leurs gains que de la suite des jeux qui les ont entraînés. Une telle quantité est appelée variable aléatoire (en abrégé, v.a.), c'est donc une fonction $X : \omega \mapsto X(\omega)$ définie sur l'espace fondamental Ω. Les variables aléatoires discrètes sont les v.a. à valeurs dans un espace E discret, c'est-à-dire un ensemble fini ou infini dénombrable. La définition exacte réclame aussi que l'événement « X prend une certaine valeur » est bien un événement de la tribu associée à l'espace de probabilité.

Définition 1. Une *variable aléatoire discrète* sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ est une application $X : \Omega \to E$ à valeurs dans un espace discret E telle que :

$$\text{pour tout } A \in \mathcal{P}(E), \quad \{X \in A\} \in \mathcal{A}\,. \tag{1}$$

La *loi* de X est la fonction $P_X : \mathcal{P}(E) \to [0,1]$ définie par :

$$P_X(A) := \mathbb{P}(X \in A), \text{ pour tout } A \in \mathcal{P}(E)\,. \tag{2}$$

Il est d'usage de noter par des majuscules $(X, Y, \ldots)$ les v.a., réservant $x, y, \ldots$ pour désigner des valeurs déterministes. Dans (1) on a utilisé la notation abrégée :

$$\{X \in A\} = \{\omega \in \Omega \mid X(\omega) \in A\}\,,$$

utilisée couramment en probabilité, et dans (2) la notation :

$$\mathbb{P}(X \in A) = \mathbb{P}(\{X \in A\})\,.$$

Remarque. Si Ω est fini ou dénombrable, alors en choisissant $\mathcal{A} := \mathcal{P}(\Omega)$, on est assuré que (1) est vérifié. Dans ce cadre, toute application $X : \Omega \to E$ est une variable aléatoire. Mais une variable aléatoire discrète peut très bien être définie sur un espace Ω infini non dénombrable. Dans ce cas, il est nécessaire de réclamer que les sous-ensembles de Ω de la forme $\{X \in A\}$ soient des événements de la tribu. Par la propriété de stabilité d'une tribu par union dénombrable, pour montrer (1) il suffit de montrer que les sous-ensembles $\{X = x\}$ pour $x \in E$ sont des événements de la tribu.

Exemple. Considérons un lancer de dé. L'espace fondamental est $\Omega = [\![1,6]\!]$. Le résultat du lancer d'un dé est ici la variable aléatoire $X(\omega) = \omega$, à valeurs dans $[\![1,6]\!]$.

Exemple. Considérons le lancer de deux dés. L'espace fondamental est $\Omega = [\![1,6]\!]^2$. Le résultat total est la somme des deux lancers, il est donné par la v.a. S définie par :

$$\begin{array}{rrcl} S: & \Omega & \longrightarrow & [\![2,12]\!] \\ & \omega = (\omega_1, \omega_2) & \longmapsto & S(\omega) = \omega_1 + \omega_2. \end{array}$$

Cette v.a. prend ses valeurs dans l'ensemble des entiers $[\![2,12]\!]$.

La proposition suivante justifie l'appellation loi de probabilité pour la fonction d'ensembles P_X.

Proposition 1.
1) La loi P_X d'une v.a. discrète X est caractérisée par la donnée des nombres $\{p_X(x),\ x \in E\}$ définis par :

$$p_X(x) := P_X(\{x\}) \text{ pour tout } x \in E\,. \tag{3}$$

Pour retrouver la loi on somme comme suit :

$$P_X(A) = \sum_{x \in A} p_X(x) \text{ pour tout } A \in \mathcal{P}(E)\,.$$

2) La fonction P_X est une probabilité sur $\big(E, \mathcal{P}(E)\big)$.

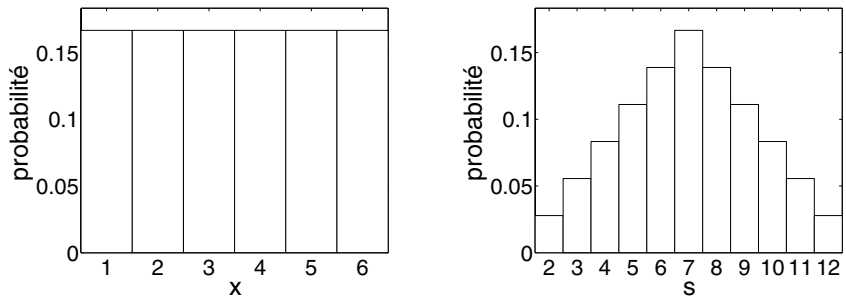
Démonstration. Soit $A \in \mathcal{P}(E)$. On peut décomposer A comme une union dénombrable d'événements élémentaires deux à deux disjoints $A = \bigcup_{x \in A} \{X = x\}$. La propriété de σ-additivité vérifiée par la probabilité $\mathbb{P}$ donne :

$$P_X(A) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} p_X(x)\,,$$

ce qui montre le 1).
Pour montrer le 2) il faut montrer que P_X vérifie les conditions requises dans la définition d'une probabilité (définition 2 de la page 669).
(i) $P_X(E) = \sum_{x \in E} p_X(x) = \sum_{x \in E} \mathbb{P}(X = x)$.
Or la suite des événements $(\{X = x\} \mid x \in E)$, est une partition de Ω, donc $P_X(E) = 1$.
(ii) Soit $(A_i)_{i \in \mathbb{N}}$ une suite d'événements incompatibles (ou sous-ensembles deux à deux disjoints) de $\mathcal{P}(E)$. Alors :

$$P_X\Big(\bigcup_{i \in \mathbb{N}} A_i\Big) = \sum_{x \in \bigcup_{i \in \mathbb{N}} A_i} p_X(x) = \sum_{i \in \mathbb{N}} \sum_{x \in A_i} p_X(x) = \sum_{i \in \mathbb{N}} P_X(A_i)\,.$$

Pour écrire la deuxième égalité on a appliqué le théorème de Fubini pour les séries à termes positifs (*cf.* le livre [L2], module « Suites et séries de fonctions ») qui nous autorise à effectuer la somme d'une série à termes positifs dans un ordre arbitraire. Ainsi P_X vérifie toutes les conditions requises dans la définition d'une probabilité sur $\big(E, \mathcal{P}(E)\big)$. ∎



(a) **(b)**

FIGURE IV.2.1 – Histogramme de la loi d'un lancer de dé (a) et de la loi de la somme du lancer de deux dés (b).

1.2 Histogrammes

Une loi de probabilité sur un espace E fini ou infini dénombrable est caractérisée par la donnée des nombres $p_X(x)$, $x \in E$. Dans le cas courant où E est un sous-ensemble de l'ensemble des entiers $\mathbb{Z}$, la loi peut être représentée par un histogramme, c'est-à-dire un ensemble de bâtons de largeur unité, centrés en $x \in E$ et de hauteur $p_X(x)$.

Exemple. On reprend l'exemple du lancer de dé. L'espace fondamental est $\Omega = [\![1,6]\!]$ muni de la probabilité uniforme. Le résultat du lancer d'un dé est la variable aléatoire $X(\omega) = \omega$, à valeurs dans $[\![1,6]\!]$, et l'on a donc :

$$p_X(1) = \frac{1}{6}\,, \quad p_X(2) = \frac{1}{6}\,, \quad \ldots, \quad p_X(6) = \frac{1}{6}\,\cdot$$

On représente cette loi par un histogramme sur la figure IV.2.1a.

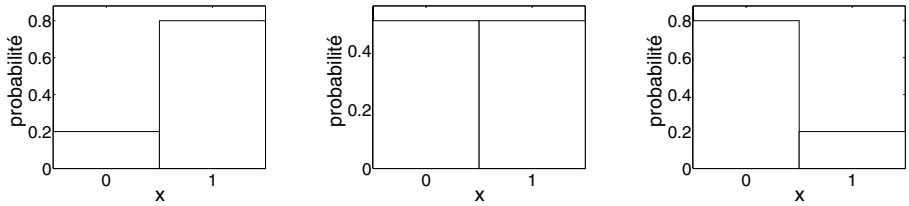
Exemple. Considérons le lancer de deux dés. L'espace fondamental est $\Omega = [\![1,6]\!]^2$ muni de la probabilité uniforme. Le résultat total est la somme des deux lancers, $S(\omega) = \omega_1 + \omega_2$, où $\omega = (\omega_1, \omega_2)$. Cette v.a. prend ses valeurs dans l'ensemble des entiers $[\![2,12]\!]$, et l'on obtient par des opérations de dénombrement élémentaires :

$$p_S(2) = \frac{1}{36}\,, \quad p_S(3) = \frac{2}{36}\,, \quad \ldots, \quad p_S(7) = \frac{6}{36}\,,$$
$$p_S(8) = \frac{5}{36}\,, \quad \ldots, \quad p_S(11) = \frac{2}{36}\,, \quad p_S(12) = \frac{1}{36}\,\cdot$$

On représente cette loi par un histogramme sur la figure IV.2.1b.

L'aire de chaque bâton d'un histogramme situé au dessus d'un certain entier x est donc égale à la probabilité de l'événement élémentaire $\{X = x\}$. Si l'on cherche la probabilité de l'événement $X \in [a,b]$, avec $a, b \in \mathbb{Z}$, alors il suffit de mesurer l'aire des bâtons compris entre a et b, puisque :

$$\mathbb{P}(X \in [a,b]) = \sum_{x=a}^{b} p_X(x)\,.$$



(a) $p = 0{,}8$ **(b)** $p = 0{,}5$ **(c)** $p = 0{,}2$

FIGURE IV.2.2 – Histogramme de la loi de Bernoulli de paramètre p.

2 Quelques lois usuelles

2.1 Loi de Bernoulli

Définition 2 (Loi de Bernoulli). Soit $p \in [0,1]$. La *loi de Bernoulli de paramètre* p est la probabilité P sur $\big(\{0,1\}, \mathcal{P}(\{0,1\})\big)$ déterminée par :

$$P(\{0\}) := 1 - p\,, \quad P(\{1\}) := p\,.$$

On la note $\mathcal{B}(p)$.

On représente les histogrammes de la loi de Bernoulli pour différentes valeurs du paramètre p sur la figure IV.2.2.
Toute v.a. discrète ne prenant pour valeurs que 0 ou 1 suit une loi de Bernoulli, comme le montre la proposition suivante.

Proposition 2. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité quelconque.
Soit $X : \Omega \to \{0,1\}$ une v.a. discrète ne prenant pour valeurs que 0 ou 1. Si l'on note $p := \mathbb{P}(X = 1)$, alors la loi de X est la loi de Bernoulli de paramètre p.

Démonstration.

La loi de X est caractérisée par les nombres $p_X(x) := \mathbb{P}(X = x)$, $x \in \{0,1\}$. On a $p_X(1) = p$ et, comme il s'agit d'une loi de probabilité, on a $p_X(0) = 1 - p_X(1) = 1 - p$. La loi de X est donc bien une loi de Bernoulli de paramètre p. ∎

On dit que X est une v.a. de Bernoulli de paramètre p. En abrégé, on écrit $X \sim \mathcal{B}(p)$ pour « X suit la loi $\mathcal{B}(p)$ ».
Un premier exemple de v.a. de Bernoulli est le suivant (jeu de Pile ou Face). Soient $\Omega := \{0,1\}$ et $\mathbb{P}$ la probabilité $\mathbb{P}(\{0\}) := 1 - p$, $\mathbb{P}(\{1\}) := p$, où $p \in [0,1]$ est un paramètre fixé. La fonction $X(\omega) := \omega$ est une v.a. de Bernoulli de paramètre p.
Les v.a. de Bernoulli apparaissent donc dans le jeu de Pile ou Face (« Pile » $:= \{X = 1\}$, « Face » $:= \{X = 0\}$), mais aussi de façon naturelle dans les espaces de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ les plus généraux. En effet, considérons un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$

arbitraire. Pour tout événement $A \in \mathcal{A}$, on définit la fonction indicatrice $\mathbf{1}_A : \Omega \to \mathbb{R}$ par :

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{si} \quad \omega \in A \\ 0 & \text{si} \quad \omega \notin A. \end{cases}$$

Il s'agit bien d'une v.a. discrète, car $\{\mathbf{1}_A = 1\} = A \in \mathcal{A}$, et $\{\mathbf{1}_A = 0\} = A^c \in \mathcal{A}$. Comme la v.a. $\mathbf{1}_A : \Omega \to \{0, 1\}$ ne prend pour valeurs que 0 ou 1, et comme $\mathbb{P}(\mathbf{1}_A = 1) = \mathbb{P}(A)$, la fonction indicatrice $\mathbf{1}_A$ est une v.a. de Bernoulli de paramètre $p = \mathbb{P}(A)$.

2.2 Loi binomiale et nombre de succès

Il existe de nombreuses situations où la question posée tourne autour du nombre de succès dans une suite d'expériences de Bernoulli. Considérons l'exemple de n lancers d'une pièce, avec une pièce qui tombe sur Pile [respectivement Face] avec probabilité p [respectivement $1 - p$] ($p \in [0, 1]$). On considère le nombre total S de succès (Pile) lors du jeu. Ici, comme le montre le raisonnement de la page 690, l'ensemble fondamental est $\Omega = \{F, P\}^n$, et la probabilité sur $\big(\Omega, \mathcal{P}(\Omega)\big)$ est $\mathbb{P}(\{\omega\}) = p^k(1-p)^{n-k}$ si ω comporte exactement k Piles et $n - k$ Faces. On note A_k le sous-ensemble de Ω constitué des éléments qui contiennent exactement k Piles. On introduit la variable aléatoire :

$$S = \sum_{k=0}^{n} k \mathbf{1}_{A_k} ,$$

c'est-à-dire que $S(\omega) =$ nombre de Piles contenus dans ω.

S prend ses valeurs dans $[\![0, n]\!]$, et l'on remarque que pour $k \in [\![0, n]\!]$, $\binom{n}{k}$ éléments ω de Ω comportent un total de k Piles. Puisque chacun de ces ω a même probabilité, on obtient :

$$\mathbb{P}(S = k) = \sum_{\omega \in A_k} \mathbb{P}(\{\omega\}) = \binom{n}{k} p^k (1-p)^{n-k} . \tag{4}$$

La loi de S est donc la loi du nombre de succès (Pile) dans un jeu de Pile ou Face, mais elle intervient couramment dans des modèles plus complexes, qui peuvent tous se reformuler comme la détermination du nombre de succès lors d'une suite de n expériences.

Définition 3 (Loi binomiale). Soient $n \in \mathbb{N}^*$ et $p \in [0, 1]$. On appelle *loi binomiale de paramètres* (n, p) la probabilité sur l'espace $([\![0, n]\!], \mathcal{P}([\![0, n]\!]))$ déterminée par :

$$P(\{k\}) := \binom{n}{k} p^k (1-p)^{n-k} , \qquad k \in [\![0, n]\!] . \tag{5}$$

En appliquant la formule du binôme on retrouve bien que la somme des probabilités élémentaires de la loi binomiale vaut 1 :

$$\sum_{k=0}^{n} P(\{k\}) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1 .$$

Le nom de la loi vient de cette formule. On note cette loi $\mathcal{B}(n, p)$.

V.a. binomiale. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. Soit $X : \Omega \to [\![0, n]\!]$ une v.a. discrète. On dit que X est une v.a. binomiale de paramètres (n, p), si elle a pour loi la loi binomiale de paramètres (n, p). En abrégé, on écrit $X \sim \mathcal{B}(n, p)$ pour « X suit la loi $\mathcal{B}(n, p)$ ».
Notez que, pour $n = 1$, on retrouve la loi de Bernoulli : $\mathcal{B}(1, p) = \mathcal{B}(p)$.

Proposition 3. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité sur lequel est définie une suite de n événements indépendants, notés $(A_i)_{i \in [\![1,n]\!]}$, qui possèdent la même probabilité de réalisation $p \in [0, 1]$. On note S le nombre d'événements A_i réalisés :

$$S := \operatorname{card}\left(i \in [\![1, n]\!] \mid A_i \text{ est réalisé}\right) = \sum_{i=1}^{n} \mathbf{1}_{A_i} .$$

Alors S suit la loi binomiale de paramètres (n, p).

Cette proposition recouvre l'exemple précédent du lancer de n pièces, où $\Omega = \{F, P\}^n$ et $A_i = \{\omega \mid \omega_i = P\}$.

Démonstration. La démonstration précédente avec la suite de tirages de pile ou face peut être directement appliquée à ce cadre plus général :

$$\mathbb{P}(S = k) = \sum_{\{i_1,\ldots,i_k\} \subset [\![1,n]\!]} \mathbb{P}\left(\left(\bigcap_{i \in \{i_1,\ldots,i_k\}} A_i\right) \cap \left(\bigcap_{i \in [\![1,n]\!] \setminus \{i_1,\ldots,i_k\}} A_i^c\right)\right) ,$$

où la somme est prise sur tous les sous-ensembles $\{i_1, \ldots, i_k\}$ à k éléments pris dans $[\![1, n]\!]$. Par indépendance des A_i, et en utilisant la proposition 11 de la page 689, on trouve que, pour tout sous-ensemble $\{i_1, \ldots, i_k\}$:

$$\begin{aligned}\mathbb{P}\left(\left(\bigcap_{i \in \{i_1,\ldots,i_k\}} A_i\right) \cap \left(\bigcap_{i \in [\![1,n]\!] \setminus \{i_1,\ldots,i_k\}} A_i^c\right)\right) &= \prod_{i \in \{i_1,\ldots,i_k\}} \mathbb{P}(A_i) \prod_{i \in [\![1,n]\!] \setminus \{i_1,\ldots,i_k\}} \mathbb{P}(A_i^c) \\ &= p^k (1-p)^{n-k} .\end{aligned}$$

Comme il y a $\binom{n}{k}$ sous-ensembles $\{i_1, \ldots, i_k\}$ à k éléments pris dans $[\![1, n]\!]$, on obtient finalement le résultat cherché. ∎

On peut donner une autre démonstration de la formule (4), moins combinatoire, et faisant appel à des probabilités conditionnelles. Reprenons l'exemple de n lancers de pièce, où l'on cherche le nombre total de Piles obtenus. On note $A_{n,k}$ l'événement « on obtient exactement k Piles sur les n lancers », et $p_{n,k} := \mathbb{P}(A_{n,k})$. On introduit aussi l'événement $B :=$ « le premier lancer donne un Pile ». D'après la formule des probabilités totales :

$$p_{n,k} = \mathbb{P}(A_{n,k} \,|\, B)\mathbb{P}(B) + \mathbb{P}(A_{n,k} \,|\, B^c)\mathbb{P}(B^c) .$$

Sachant B, l'événement $A_{n,k}$ se produit s'il y a exactement $k-1$ Piles sur les $n-1$ derniers lancers, donc $\mathbb{P}(A_{n,k} \,|\, B) = p_{n-1,k-1}$. De même, sachant B^c, l'événement $A_{n,k}$ se produit s'il y a exactement k Piles sur les $n-1$ derniers lancers, donc $\mathbb{P}(A_{n,k} \,|\, B^c) = p_{n-1,k}$. On

obtient donc :

$$p_{n,k} = p_{n-1,k-1}p + p_{n-1,k}(1-p)\,.$$

En tenant compte du fait que $\mathbb{P}(A_{n,0}) = (1-p)^n$ et $\mathbb{P}(A_{n,n}) = p^n$, c'est-à-dire $p_{n,0} = (1-p)^n$ et $p_{n,n} = p^n$, on obtient une formule de récurrence en n donnant de proche en proche les valeurs de $(p_{n,k})_{0\leqslant k\leqslant n}$, qui est très similaire à la formule définissant le remplissage du triangle de Pascal (voir le livre [L1]). Pour compléter cette analogie, dans le cas $p \in \,]0,1[$, notons :

$$q_{n,k} := \frac{p_{n,k}}{p^k(1-p)^{n-k}}\,.$$

La suite de nombres $(q_{n,k})_{0\leqslant k\leqslant n<+\infty}$ satisfait alors :

$$q_{n,k} = q_{n-1,k-1} + q_{n-1,k}\,,$$

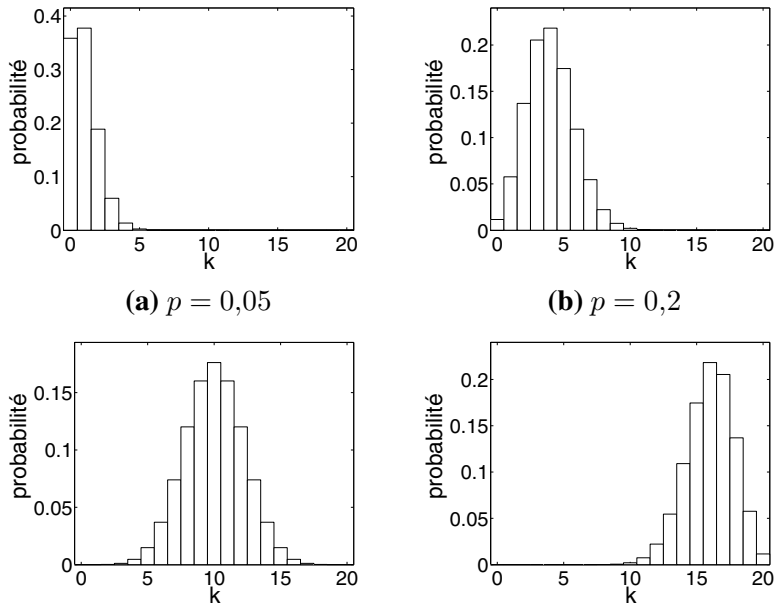
avec les conditions $q_{n,0} = q_{n,n} = 1$. Il s'agit cette fois-ci exactement de la relation du triangle de Pascal, et cela montre donc que $q_{n,k} = \binom{n}{k}$.
On peut donner quelques autres propriétés importantes de la loi binomiale. En particulier, il existe une relation simple entre $P(\{k+1\})$ et $P(\{k\})$, valable pour tout $k \in [\![0, n-1]\!]$:

$$\begin{aligned}
P(\{k+1\}) &= \binom{n}{k+1}p^{k+1}(1-p)^{n-k-1} = \frac{n!}{(k+1)!(n-k-1)!}p^{k+1}(1-p)^{n-k-1}\\
&= \frac{n-k}{k+1}\frac{n!}{k!(n-k)!}\frac{p}{1-p}p^k(1-p)^{n-k} = \frac{n-k}{k+1}\frac{p}{1-p}\binom{n}{k}p^k(1-p)^{n-k}\\
&= \frac{n-k}{k+1}\frac{p}{1-p}P(\{k\})\,.
\end{aligned}$$

Cette relation de récurrence partant de $P(\{0\}) = (1-p)^n$ est utile d'un point de vue pratique pour calculer les valeurs de $P(\{k\})$ dans le cas où n est grand, car le calcul de la formule donnée par (5) fait intervenir des facteurs factoriels qui peuvent être extrêmement grands. D'autre part, cette relation est utile pour déterminer la forme de la loi. On a :

$$\frac{P(\{k+1\})}{P(\{k\})} = \frac{n-k}{k+1}\frac{p}{1-p}\,,$$

qui est supérieur à 1 si $k < (n+1)p - 1$, ce qui montre que les valeurs des probabilités $P(\{k\})$ sont croissantes jusqu'à ce point. Le rapport est inférieur à 1 si $k > (n+1)p-1$, ce qui montre que les valeurs des probabilités $P(\{k\})$ sont décroissantes à partir de ce point. Dans le détail, notons $k_{max} = [(n+1)p]$(où $[x]$ désigne la partie entière de x, c'est-à-dire le plus grand entier inférieur ou égal à x). Le maximum de l'application $k \mapsto P(\{k\})$ est atteint uniquement en k_{max} si $(n+1)p$ n'est pas entier, et simultanément en $k_{max} - 1$ et en k_{max} si $(n+1)p$ est entier. La forme de la loi est visible lorsque l'on trace son histogramme. On représente les histogrammes de la loi binomiale pour $n = 20$ et pour différentes valeurs du paramètre p sur la figure IV.2.3 de la page suivante.



(a) $p = 0{,}05$ **(b)** $p = 0{,}2$

(c) $p = 0{,}5$ **(d)** $p = 0{,}8$

FIGURE IV.2.3 – Histogramme de la loi binomiale de paramètres (n, p) avec $n = 20$.

2.3 Temps d'attente et loi géométrique

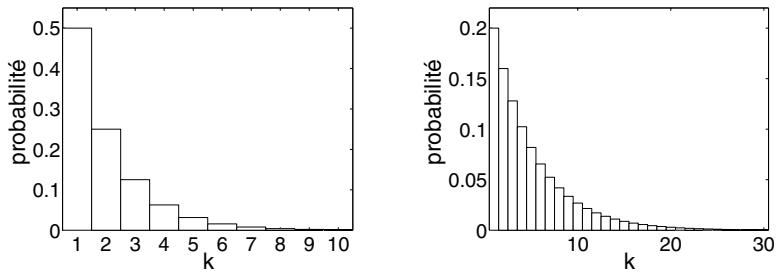
Une autre question couramment rencontrée en théorie des probabilités concerne des problèmes de temps d'attente. Reprenons l'exemple de la page 666 où l'on lance une pièce jusqu'à obtenir Pile. On utilise une pièce qui, à chaque lancer, donne Pile avec probabilité p, $p \in]0, 1]$. On considère l'instant $T \geqslant 1$ du premier succès (Pile). Ici l'ensemble fondamental est $\Omega := \{\omega_1, \omega_2, \ldots\} \cup \{\omega_\infty\}$ où ω_i représente le résultat $FF \cdots FP$ où les $i - 1$ premiers lancers ont donné Face et le $i^{ème}$ Pile. ω_∞ est l'événement $FFFF \cdots$ où tous les lancers donnent des Faces. La probabilité associée est $\mathbb{P}(\{\omega_i\}) := (1-p)^{i-1}p$ et $\mathbb{P}(\{\omega_\infty\}) := 0$, comme le montre la discussion sur l'exemple de la page 675. La v.a. T est définie par :

$$T := \sum_{1 \leqslant j \leqslant +\infty} j \mathbf{1}_{\{\omega_j\}},$$

c'est-à-dire que $T(\omega) = j$ si $\omega = \omega_j$, et sa loi est donc :

$$\mathbb{P}(T = i) = (1-p)^{i-1}p, \;\; i \geqslant 1,$$

avec $(1-p)^\infty = 0$ si $i = +\infty$. Cette loi se rencontre dans tous les problèmes de temps d'attente. C'est la loi géométrique de paramètre p.



(a) $p = 0{,}5$ **(b)** $p = 0{,}2$

FIGURE IV.2.4 – Histogramme de la loi géométrique de paramètre p.

Définition 4 (Loi géométrique). Soit $p \in]0,1]$. On appelle *loi géométrique de paramètre* p la probabilité sur l'espace $\big(\mathbb{N}^*, \mathcal{P}(\mathbb{N}^*)\big)$ déterminée par :

$$P(\{k\}) := p(1-p)^{k-1}, \qquad k \in \mathbb{N}^*.$$

Pour $p = 1$, on a $P(\{k\}) = 0$ pour tout $k \geqslant 2$ et $P(\{1\}) = 1$ car on utilise la convention $0^0 = 1$. En appliquant la formule de la somme d'une série géométrique, on retrouve bien que la somme des probabilités élémentaires vaut 1 :

$$\sum_{k=1}^{+\infty} P(\{k\}) = \sum_{k=1}^{+\infty} p(1-p)^{k-1} = \frac{p}{1-(1-p)} = 1.$$

Le nom de la loi vient de cette formule. On note cette loi $\mathcal{G}(p)$. Son histogramme a la forme dessinée sur la figure IV.2.4. Plus p est petit, et plus l'histogramme s'étale vers les grandes valeurs, car la probabilité de succès s'amenuisant, il faut beaucoup d'essais avant d'obtenir un succès.

V.a. géométrique. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité. Soit $X : \Omega \to \mathbb{N}^*$ une v.a. discrète. On dit que X est une *v.a. géométrique de paramètre* p, si elle a pour loi la loi géométrique de paramètre p.

2.4 Échantillonnage et loi hypergéométrique

Le problème de l'échantillonnage est central en statistique. La situation typique est la suivante. On souhaite connaître les intentions de vote dans une certaine population de m individus. Cette population se divise en m_a personnes qui souhaitent voter pour le candidat A, et m_b personnes qui souhaitent voter pour le candidat B. On suppose que $m_a + m_b = m$. On effectue un sondage sur n personnes tirées au hasard, par exemple en tirant au hasard des numéros dans l'annuaire téléphonique (en supposant que tout le monde a un numéro de téléphone et un seul). Quelle est la loi du nombre de personnes déclarant vouloir voter pour A ?

De manière plus générale, on peut supposer que l'on a affaire à une population de m individus divisée en deux classes complémentaires, que l'on tire n individus au hasard dans cette

population, et que l'on souhaite savoir combien d'individus tirés appartiennent à une classe donnée. La formulation la plus neutre et classique de ce problème est celle de l'urne : on dispose d'une urne contenant m_a boules noires et m_b boules blanches. On tire n boules de cette urne, et l'on compte le nombre de boules noires. Il y a en fait deux manières de tirer les n boules (ou individus) :

1. Après avoir tiré un certain individu, et déterminé sa classe, on peut décider de ne plus le tirer, et donc de choisir au hasard parmi les individus restants de la population. On parle alors de tirage sans remise. Dans l'exemple du sondage sur les intentions de vote, après avoir tiré un numéro de téléphone et sondé une personne, on décide de ne plus l'appeler pour ce sondage, et de tirer un nouveau numéro au hasard dans l'annuaire, hormis les numéros déjà tirés. Dans l'exemple de l'urne avec les boules, après avoir tiré une boule et noté sa couleur, on ne remet pas cette boule dans l'urne, et l'on continue le tirage.
2. On peut décider, après avoir tiré un certain individu, et déterminé sa classe, de continuer le sondage comme si de rien n'était, et de retirer un nouvel individu dans la population globale. On peut donc retirer le même individu. On parle alors de tirage avec remise. Dans l'exemple de l'urne avec les boules, après avoir tiré une boule et noté sa couleur, on remet cette boule dans l'urne, et l'on continue le tirage.

On peut modéliser le problème de l'urne comme suit.
L'espace fondamental est $\Omega = \{A, B\}^n$, où $\omega_i = A$ (respectivement B) signifie que la $i^{ème}$ boule tirée est noire (respectivement blanche). Les expériences du tirage avec ou sans remise correspondent à deux probabilités différentes sur Ω.
Pour un tirage avec remise, le problème se ramène en fait à une suite de n expériences de Bernoulli avec une probabilité de succès (i.e. tirer une boule noire) égale à $p = m_a/(m_a + m_b)$. Par conséquent, le nombre total S de boules noires tirées suit une loi binomiale de paramètres (n, p) :

$$\mathbb{P}(S = k) = \binom{n}{k} \left(\frac{m_a}{m_a + m_b}\right)^k \left(\frac{m_b}{m_a + m_b}\right)^{n-k} .$$

Pour un tirage sans remise, il faut faire un peu de combinatoire. Pour éviter de se retrouver avec des sous-cas qui compliqueraient inutilement la situation, on va supposer que n est plus petit que m_a et m_b, si bien qu'il existe effectivement des configurations où l'on tire k boules noires, avec k compris entre 0 et n (si $m_a < n$, alors on ne peut clairement pas tirer n boules noires de l'urne, et si $m_b < n$, alors on ne peut clairement pas tirer n boules blanches, soit 0 boule noire). D'une part, il y a en tout $\binom{m_a+m_b}{n}$ manières possibles de tirer n boules de l'urne. D'autre part, il y a $\binom{m_a}{k}\binom{m_b}{n-k}$ manières possibles de tirer k boules noires parmi les m_a boules noires de l'urne et $n - k$ boules blanches parmi les m_b boules blanches de l'urne. Comme tous ces tirages sont équiprobables, on obtient, pour tout $k \in [\![0, n]\!]$:

$$\mathbb{P}(S = k) = \frac{\binom{m_a}{k}\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} \cdot \tag{6}$$

En développant les facteurs combinatoires du numérateur, et en les regroupant de manière

différente :

$$\mathbb{P}(S=k)=\frac{\binom{m_b}{n}}{\binom{m_a+m_b}{n}}\times\frac{\binom{n}{k}\binom{m_a}{k}}{\binom{m_b-n+k}{k}}\,.$$

Le second facteur apparaît dans le développement en série entière d'une fonction tabulée célèbre, la fonction hypergéométrique :

$$F^n_{m_a,m_b}(x)=\sum_{k=0}^{+\infty}\frac{\binom{n}{k}\binom{m_a}{k}}{\binom{m_b-n+k}{k}}x^k\,, \tag{7}$$

paramétrée par (m_a,m_b,n). Cette fonction apparaît dans les solutions d'équations de la physique. Le nom de la loi définie par (6) vient de cette écriture.

Proposition 4. Lors du tirage sans remise de n boules dans une urne contenant m_a boules noires et m_b boules blanches, avec $m_a\geqslant n$ et $m_b\geqslant n$, le nombre total S de boules noires tirées suit la loi hypergéométrique de paramètres (m_a,m_b,n) :

$$\mathbb{P}(S=k)=\frac{\binom{m_a}{k}\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}}=\frac{\binom{m_b}{n}}{\binom{m_a+m_b}{n}}\times\frac{\binom{n}{k}\binom{m_a}{k}}{\binom{m_b-n+k}{k}}\,,\quad k\in[\![0,n]\!]\,.$$

3 Espérance de variables aléatoires discrètes réelles

Dans ce paragraphe on considère des variables aléatoires à valeurs dans un sous-ensemble fini ou dénombrable de $\mathbb{R}$ que l'on appelle des v.a. discrètes réelles. Bien qu'une v.a. discrète réelle fluctue d'une expérience à l'autre, on peut lui attacher plusieurs valeurs « typiques » : sa valeur moyenne, sa valeur la plus probable, ... Toutes ces quantités sont d'intérêt, mais la plus importante, et souvent la plus utile, est la moyenne. Le vocabulaire utilise la dénomination d'espérance pour cette quantité (apparue pour la première fois en 1657 dans un livre de Christiaan Huygens). Elle est bien connue de tous à cause de l'espérance de vie, par exemple si importante dans le domaine de l'assurance, qui a motivé depuis le XVIIIe siècle de nombreux développements de la théorie des probabilités.

3.1 Définition de l'espérance

Définition 5 (v.a. intégrable). Soit X une v.a. discrète définie sur un espace de probabilité $(\Omega,\mathcal{A},\mathbb{P})$ à valeurs dans $E:=\{x_i\mid i\geqslant 1\}\subset\mathbb{R}$ fini ou dénombrable. On dit que X est *intégrable* si :

$$\sum_{i\geqslant 1}|x_i|p_X(x_i)<+\infty\,,$$

où $p_X(x):=\mathbb{P}(X=x)$ pour tout $x\in E$.

Dans le cas d'une v.a. discrète réelle à valeurs dans un espace fini, la condition d'intégrabilité est automatiquement vérifiée.

Définition 6 (espérance).
L'*espérance* d'une v.a. discrète intégrable $X : \Omega \to E = \{x_i \mid i \geqslant 1\}$ est définie par :

$$\mathbb{E}[X] := \sum_{i \geqslant 1} x_i p_X(x_i)\,. \tag{8}$$

Exemples.
1) Si A est un événement, l'espérance de la v.a. de Bernoulli $\mathbf{1}_A$ est :

$$\mathbb{E}[\mathbf{1}_A] = 0 \times \mathbb{P}(\mathbf{1}_A = 0) + 1 \times \mathbb{P}(\mathbf{1}_A = 1) = 0 \times \mathbb{P}(A^c) + 1 \times \mathbb{P}(A)\,,$$

soit :

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)\,. \tag{9}$$

Cela montre que la notion d'espérance prolonge celle de probabilité, et elle en héritera des propriétés.
2) Si X est une v.a. constante, soit $X := a$ pour un $a \in \mathbb{R}$, alors elle est intégrable et d'espérance $\mathbb{E}[X] = a$.

3.2 Propriétés élémentaires de l'espérance

Soit $L^1_d(\Omega, \mathcal{A}, \mathbb{P})$ l'ensemble des v.a. discrètes intégrables définies sur un même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$.

Proposition 5 (Linéarité et positivité de l'espérance). Soient X et Y deux v.a. discrètes réelles appartenant à $L^1_d(\Omega, \mathcal{A}, \mathbb{P})$, et λ, μ deux réels.
1) La v.a. discrète réelle $Z := \lambda X + \mu Y$ appartient à L^1_d et $\mathbb{E}[Z] = \lambda\mathbb{E}[X] + \mu\mathbb{E}[Y]$.
2) Si $X \geqslant Y$, alors $\mathbb{E}[X] \geqslant \mathbb{E}[Y]$.

Démonstration. Soient $\{a_i \mid i \in I\}$ et $\{b_j \mid j \in J\}$ les ensembles de valeurs prises par X et Y respectivement, avec $I, J \subset \mathbb{N}$.
Notons $A_i := \{X = a_i\}$, $B_j := \{Y = b_j\}$, et $C_{ij} := A_i \cap B_j$. Comme la tribu $\mathcal{A}$ est stable par intersection, les C_{ij} sont des événements. Les valeurs possibles prises par Z forment l'ensemble :

$$E = \big\{\lambda a_i + \mu b_j \mid (i,j) \in I \times J\big\}\,.$$

Comme $I \times J$ est au plus dénombrable, l'ensemble E est au plus dénombrable et l'on peut numéroter ses éléments par un ensemble $K \subset \mathbb{N}$:

$$E = \big\{c_k \mid k \in K\big\}\,.$$

Comme $\{Z = c_k\} = \bigcup_{(i,j) \in I \times J \mid \lambda a_i + \mu b_j = c_k} C_{ij}$ est une union au plus dénombrable d'événements, c'est un événement, pour tout $k \in K$. Donc Z est une v.a. réelle discrète. Sa loi est déterminée par :

$$\mathbb{P}(Z = c_k) = r_k\,, \text{ avec } r_k = \sum_{(i,j) \in I \times J \mid \lambda a_i + \mu b_j = c_k} \mathbb{P}(C_{ij}).$$

Or la série :

$$\begin{aligned}
\sum_{k\in K}|c_k|r_k &= \sum_{k\in K}|c_k| \sum_{(i,j)\in I\times J|\lambda a_i+\mu b_j=c_k} \mathbb{P}(C_{ij}) \\
&= \sum_{k\in K}\sum_{(i,j)\in I\times J|\lambda a_i+\mu b_j=c_k} |\lambda a_i+\mu b_j|\mathbb{P}(C_{ij}) \\
&= \sum_{i\in I,\, j\in J} |\lambda a_i+\mu b_j|\mathbb{P}(C_{ij}) \\
&\leqslant |\lambda|\sum_{i\in I}|a_i|\sum_{j\in J}\mathbb{P}(C_{ij}) + |\mu|\sum_{j\in J}|b_j|\sum_{i\in I}\mathbb{P}(C_{ij}) \\
&= |\lambda|\sum_{i\in I}|a_i|\mathbb{P}(X=a_i) + |\mu|\sum_{j\in J}|b_j|\mathbb{P}(Y=b_j)\,,
\end{aligned}$$

est convergente (on a appliqué le théorème de Fubini pour les séries à termes positifs (*cf.* le livre [L2], module « Suites et séries de fonctions »), donc Z est intégrable. Enfin, en appliquant le théorème de Fubini pour les séries absolument convergentes (*cf.* le livre [L2], même module) :

$$\begin{aligned}
\mathbb{E}[Z] = \sum_{k\in K} c_k r_k &= \sum_{k\in K} c_k \sum_{(i,j)\in I\times J|\lambda a_i+\mu b_j=c_k} \mathbb{P}(C_{ij}) \\
&= \sum_{k\in K}\sum_{(i,j)\in I\times J|\lambda a_i+\mu b_j=c_k} (\lambda a_i+\mu b_j)\mathbb{P}(C_{ij}) \\
&= \sum_{i\in I,\, j\in J} (\lambda a_i+\mu b_j)\mathbb{P}(C_{ij}) \\
&= \lambda\sum_{i\in I} a_i \sum_{j\in J}\mathbb{P}(C_{ij}) + \mu\sum_{j\in J} b_j\sum_{i\in I}\mathbb{P}(C_{ij}) \\
&= \lambda\sum_{i\in I} a_i\mathbb{P}(X=a_i) + \mu\sum_{j\in J} b_j\mathbb{P}(Y=b_j) \\
&= \lambda\mathbb{E}[X] + \mu\mathbb{E}[Y]\,.
\end{aligned}$$

Quant à l'inégalité du 2), il suffit d'observer que $\mathbb{E}[X-Y] \geqslant 0$ par définition de l'espérance ($X-Y$ est une v.a. qui ne prend que des valeurs positives), et d'appliquer ensuite la linéarité qui vient d'être démontrée. ■

Le corollaire suivant ne fait que reformuler la proposition en des termes d'algèbre linéaire.

Corollaire 6. $L^1_d(\Omega,\mathcal{A},\mathbb{P})$ est un espace vectoriel, et $\mathbb{E} : \begin{array}{ccc} L^1_d & \longrightarrow & \mathbb{R} \\ X & \longmapsto & \mathbb{E}[X] \end{array}$ est une forme linéaire positive.

3.3 Propriétés de transport

Considérons une v.a. réelle discrète $X : \Omega \to E$ (c'est-à-dire que E est une partie au plus dénombrable de $\mathbb{R}$). Étant donnée une fonction f définie sur E (à valeurs dans l'ensemble discret $E' := f(E)$), la loi de la v.a. discrète $Y := f(X)$ se calcule aisément à l'aide de la loi de X. Notons $p_X(x) := \mathbb{P}(X = x)$, $x \in E$, et $p_Y(y) := \mathbb{P}(Y = y)$, $y \in E'$. Puisque $\{Y = y\} = \bigcup_{x \in E | f(x) = y} \{X = x\}$ on a :

$$p_Y(y) = \sum_{x \in E} \mathbf{1}_y\big(f(x)\big)\, p_X(x)\,, \quad y \in E'\,, \tag{10}$$

où $\mathbf{1}_y\big(f(x)\big) = 1$ si $f(x) = y$ et 0 sinon.

Il arrive souvent que l'on ait besoin de calculer, non l'espérance de la variable aléatoire X, mais l'espérance d'une fonction $Y := f(X)$ de celle-ci. Si l'on applique la définition de l'espérance, cela suppose que l'on calcule la loi de Y, ce qui peut être incommode. La proposition suivante montre que l'on n'a pas forcément besoin de calculer explicitement la loi d'une v.a. $Y = f(X)$ pour calculer son espérance.

Proposition 7. Soient $X : \Omega \to E$ une v.a. discrète et $f : E \to \mathbb{R}$ une fonction à valeurs réelles. Si $\sum_{x \in E} \big|f(x)\big| p_X(x) < +\infty$ alors $Y := f(X)$ est une v.a. discrète intégrable et l'on a :

$$\mathbb{E}[f(X)] = \sum_{x \in E} f(x) p_X(x)\,. \tag{11}$$

Démonstration. Montrons que $Y : \Omega \to E' = f(E) \subset \mathbb{R}$ est intégrable. En utilisant (10) :

$$\sum_{y \in E'} |y| p_Y(y) = \sum_{y \in E'} |y| \sum_{x \in E} \mathbf{1}_y\big(f(x)\big) p_X(x)\,.$$

On applique la version du théorème de Fubini valable pour les séries positives :

$$\sum_{y \in E'} |y| p_Y(y) = \sum_{x \in E} \sum_{y \in E'} |y| \mathbf{1}_y\big(f(x)\big)\, p_X(x) = \sum_{x \in E} \big|f(x)\big| p_X(x)\,.$$

Cela montre que $\sum_{y \in E'} |y| p_Y(y) < +\infty$ et donc Y est intégrable. Enfin, on peut reproduire la même démonstration en enlevant les $|\cdot|$ et en utilisant la version de Fubini valable pour les séries absolument convergentes, et l'on obtient ainsi (11). ∎

3.4 Quelques remarques

V.a. discrètes à valeurs dans $\mathbb{N}$

On aura souvent affaire à des v.a. discrètes à valeurs dans $\mathbb{N}$, ou un sous-ensemble de $\mathbb{N}$. Dans ce cas on dispose d'une autre relation entre l'espérance de la v.a. et sa loi de probabilité.

Proposition 8. Soit X une v.a. discrète à valeurs dans $\mathbb{N}$ telle que $\mathbb{E}[X] < +\infty$. Alors :

$$\mathbb{E}[X] = \sum_{n=0}^{+\infty} \mathbb{P}(X > n)\,. \tag{12}$$

Démonstration. On décompose la somme qui constitue la définition de l'espérance :

$$\mathbb{E}[X] = \sum_{n=0}^{+\infty} n\mathbb{P}(X = n) = \sum_{n=0}^{+\infty}\sum_{k=0}^{n-1} \mathbb{P}(X = n)\,,$$

puis l'on applique le théorème de Fubini (pour les séries à termes positifs) :

$$\mathbb{E}[X] = \sum_{k=0}^{+\infty}\sum_{n=k+1}^{+\infty} \mathbb{P}(X = n) = \sum_{k=0}^{+\infty} \mathbb{P}(X \geqslant k+1)\,,$$

ce qu'il fallait démontrer. ∎

Une application de la linéarité de l'espérance : formule d'inclusion-exclusion

Cette formule a déjà été démontrée dans le module précédent, page 677. On propose ici une démonstration plus simple. La règle d'inclusion-exclusion s'applique à une suite $(A_i)_{i\in[\![1,n]\!]}$ d'événements pour lesquels on connaît *a priori* les probabilités des conjonctions d'événements. On peut ainsi calculer la probabilité de $A = A_1 \cup A_2 \cup \cdots \cup A_n$, i.e. la probabilité que « au moins l'un des A_i est réalisé ». Plus précisément, soit $(A_i)_{i\in[\![1,n]\!]}$ une suite d'événements. Alors :

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\leqslant i_1<i_2<\cdots<i_k\leqslant n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right)\,. \tag{13}$$

Démonstration. Il est facile de vérifier d'une part que :

$$1 - \mathbf{1}_{\bigcup_{i=1}^{n} A_i} = \prod_{i=1}^{n}(1 - \mathbf{1}_{A_i})\,, \tag{14}$$

et d'autre part que, pour tout sous-ensemble d'indices $1 \leqslant i_1 < \cdots < i_k \leqslant n$:

$$\prod_{j=1}^{k} \mathbf{1}_{A_{i_j}} = \mathbf{1}_{\bigcap_{j=1}^{k} A_{i_j}}\,.$$

En prenant l'espérance de l'égalité (14) et en développant le membre de droite, on obtient directement la formule d'inclusion-exclusion (13) avec (9) (page 718). ∎

3.5 Variance

Pour $k \in \mathbb{N}^*$, le $k^{ème}$ *moment* de la v.a. réelle X est par définition l'espérance $\mathbb{E}[X^k]$ qui est bien définie si X^k est intégrable. La proposition 7 permet de calculer ces moments à l'aide de la loi de X. Les moments les plus utilisés sont le moment d'ordre 1 (l'espérance), et celui d'ordre 2 sous forme centrée (la variance). On dit que X est de carré intégrable si la v.a. X^2 est intégrable.

Définition 7. On appelle *variance* d'une variable aléatoire X de carré intégrable, le nombre positif :

$$\mathrm{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] .$$

Ce nombre positif est l'écart quadratique moyen de X à son espérance $\mathbb{E}[X]$, il mesure donc l'ampleur de la dispersion de la loi de la v.a. X. En fait, puisque la variance a la dimension d'un carré, c'est plutôt l'écart-type de X :

$$\sigma(X) := \sqrt{\mathrm{Var}(X)} ,$$

qui donne l'ordre de grandeur des fluctuations autour de la moyenne.

Proposition 9. Soit X une v. a. discrète de carré intégrable.
1) Une expression équivalente pour la variance est (formule de Huygens) :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 . \tag{15}$$

2) Si $a, b \in \mathbb{R}$, alors $aX + b$ est de carré intégrable, et l'on a :

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X) . \tag{16}$$

3) Si $\mathrm{Var}(X) = 0$, alors $\mathbb{P}(X = \mathbb{E}[X]) = 1$.

La formule (15) est utile en pratique pour calculer la variance.
Démonstration. Pour le 1), on part de la définition de $\mathrm{Var}(X)$, on développe le carré, et par linéarité, on obtient :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}\left[2X\mathbb{E}[X]\right] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 .$$

On vérifie aisément le 2) en utilisant la linéarité de l'espérance.
Pour démontrer le 3), on introduit la v.a. discrète $Y := X - \mathbb{E}[X]$. On a donc :

$$0 = \mathbb{E}[Y^2] = \sum_y y^2 \mathbb{P}(Y = y) .$$

La somme dans le membre de droite ne comporte que des termes positifs ou nuls. Comme la somme est nulle on a $\mathbb{P}(Y = y) = 0$ pour tout $y \neq 0$. Comme Y est à valeurs dans un espace fini ou dénombrable, on en déduit que $\mathbb{P}(Y = 0) = 1$, ce qui donne le résultat. ■

Exemple. Si X est une v.a. de Bernoulli de paramètre p, alors on calcule facilement $\mathbb{E}[X] = \mathbb{E}[X^2] = p$, et par (15) :

$$\mathrm{Var}(X) = p(1 - p) .$$

En étudiant la fonction $p \mapsto p(1 - p)$, on trouve que la variance est maximale lorsque $p = 1/2$ et elle est nulle pour $p = 0$ et $p = 1$.

Proposition 10 (Inégalité de Bienaymé-Tchebychev). Pour toute v.a. X de carré intégrable, on a pour tout $a > 0$:

$$\mathbb{P}\left(\left|X - \mathbb{E}[X]\right| \geqslant a\right) \leqslant \frac{1}{a^2}\mathrm{Var}(X)\,. \tag{17}$$

Démonstration. En examinant séparément les cas $|y| \geqslant a$ et $|y| < a$, on trouve que, pour tout $y \in \mathbb{R}$, on a :

$$\mathbf{1}_{[a,+\infty[}(|y|) \leqslant \frac{1}{a^2}y^2\,.$$

Par conséquent, si Y est une v.a., alors :

$$\mathbf{1}_{\{|Y|\geqslant a\}} \leqslant \frac{1}{a^2}Y^2\,.$$

En supposant que Y est de carré intégrable, on peut prendre l'espérance, et en utilisant (9) (page 718), on trouve :

$$\mathbb{P}(|Y| \geqslant a) \leqslant \frac{1}{a^2}\mathbb{E}[Y^2]\,.$$

L'inégalité cherchée s'obtient en prenant $Y = X - \mathbb{E}[X]$. ■

Cette proposition montre que la variance mesure effectivement la dispersion d'une v.a. autour de sa moyenne. En effet, en appliquant l'inégalité de Bienaymé-Tchebychev avec $a = 3{,}16\sigma$ par exemple, on trouve qu'il est improbable (probabilité inférieure à 10%) qu'une v.a. X dévie de $\mathbb{E}[X]$ de plus de $3{,}16$ fois son écart-type $\sigma = \sqrt{\mathrm{Var}(X)}$. L'étude des (grandes) déviations d'une v.a. est une branche importante de la théorie des probabilités, car elle concerne les probabilités d'événements rares tels que des pannes de systèmes ou des accidents. Il existe en fait des théorèmes donnant des inégalités plus fines que l'inégalité de Bienaymé-Tchebychev, mais qui dépassent le cadre de ce cours.

3.6 Espérances et variances pour des lois usuelles

Proposition 11. Si X suit la loi binomiale $\mathcal{B}(n,p)$, alors :

$$\mathbb{E}[X] = np\,, \qquad \mathrm{Var}(X) = np(1-p)\,.$$

Démonstration. On verra une démonstration très rapide de ce résultat dans la suite. On donne ici une preuve complète calculatoire. Par définition des moments de la loi binomiale :

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \times \binom{n}{k} p^k (1-p)^{n-k}, \qquad \mathbb{E}[X^2] = \sum_{k=0}^{n} k^2 \times \binom{n}{k} p^k (1-p)^{n-k}\,.$$

Si $p = 1$, alors on trouve immédiatement que

$$\mathbb{E}[X] = \sum_{k=0}^{n} k \times \binom{n}{k} 1^k 0^{n-k} = n\,, \qquad \mathbb{E}[X^2] = \sum_{k=0}^{n} k^2 \times \binom{n}{k} 1^k 0^{n-k} = n^2\,,$$

et donc $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = n^2 - n^2 = 0$.

Supposons dorénavant que $p \in [0,1[$. On peut écrire :

$$\mathbb{E}[X] = (1-p)^n \left[\sum_{k=0}^{n} k \binom{n}{k} x^k \right]_{x=p/(1-p)} . \tag{18}$$

Si l'on dérive par rapport à x l'identité du binôme :

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k ,$$

alors on obtient après avoir multiplié par x :

$$nx(1+x)^{n-1} = \sum_{k=0}^{n} k \binom{n}{k} x^k . \tag{19}$$

En substituant dans l'expression (18) de $\mathbb{E}[X]$ cette égalité évaluée en $x = p/(1-p)$, on trouve :

$$\mathbb{E}[X] = (1-p)^n \left[nx(1+x)^{n-1} \right]_{x=p/(1-p)} = np .$$

Un argument similaire permet de calculer $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = (1-p)^n \left[\sum_{k=0}^{n} k^2 \binom{n}{k} x^k \right]_{x=p/(1-p)} .$$

On dérive (19) et l'on multiplie par x à nouveau :

$$n(1+x)^{n-2}(x+nx^2) = \sum_{k=0}^{n} k^2 \binom{n}{k} x^k ,$$

ce qui donne :

$$\mathbb{E}[X^2] = (1-p)^n \left[n(1+x)^{n-2}(x+nx^2) \right]_{x=p/(1-p)} = np + n(n-1)p^2 ,$$

et finalement :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = np + n(n-1)p^2 - n^2p^2 = np(1-p) .$$

Proposition 12. Si X suit la loi géométrique $\mathcal{G}(p)$, alors :

$$\mathbb{E}[X] = \frac{1}{p} , \qquad \mathrm{Var}(X) = \frac{1-p}{p^2} .$$

Démonstration. L'espérance d'une v.a. géométrique de paramètre p est donnée par :

$$\mathbb{E}[X] = \sum_{k=1}^{+\infty} k \times p(1-p)^{k-1} = p \left[\sum_{k=1}^{+\infty} kx^{k-1} \right]_{x=1-p} . \tag{20}$$

On utilise la somme d'une série géométrique et le théorème de dérivation terme à terme d'une série entière (*cf.* le livre [L2], module « Suites et séries de fonctions ») pour calculer,

pour tout $x \in [0,1[$:

$$\frac{1}{(1-x)^2} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{1-x}\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{k=0}^{+\infty} x^k\right) = \sum_{k=0}^{+\infty}\left(\frac{\mathrm{d}}{\mathrm{d}x}x^k\right) = \sum_{k=1}^{+\infty} kx^{k-1}\,. \tag{21}$$

En substituant dans (20) :

$$\mathbb{E}[X] = p\left[\frac{1}{(1-x)^2}\right]_{x=1-p} = \frac{1}{p}\,.$$

Un argument similaire permet de calculer $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \sum_{k=1}^{+\infty} k^2 p(1-p)^{k-1} = p\left[\sum_{k=1}^{+\infty} k^2 x^{k-1}\right]_{x=1-p}\,.$$

En effet, en repartant de (21) :

$$\frac{1+x}{(1-x)^3} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x}{(1-x)^2}\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\sum_{k=0}^{+\infty} kx^k\right) = \sum_{k=0}^{+\infty}\left(\frac{\mathrm{d}}{\mathrm{d}x}kx^k\right) = \sum_{k=1}^{+\infty} k^2x^{k-1}\,,$$

ce qui donne :

$$\mathbb{E}[X^2] = p\left[\frac{1+x}{(1-x)^3}\right]_{x=1-p} = \frac{2-p}{p^2}\,,$$

et finalement :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}\,.$$

Proposition 13. Si X suit la loi hypergéométrique de paramètres (m_a, m_b, n), alors :

$$\mathbb{E}[X] = \frac{m_a n}{m_a + m_b}\,, \qquad \mathrm{Var}(X) = \frac{n(m_a+m_b-n)m_a m_b}{(m_a+m_b-1)(m_a+m_b)^2}\,.$$

Démonstration. Par définition de l'espérance et de la loi hypergéométrique :

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} k\frac{\binom{m_a}{k}\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} = \sum_{k=0}^{n} k\frac{m_a!}{k!(m_a-k)!}\frac{\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} \\
&= \sum_{k=1}^{n} \frac{m_a!}{(k-1)!(m_a-k)!}\frac{\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}} = \sum_{j=0}^{n-1} \frac{m_a!}{j!(m_a-1-j)!}\frac{\binom{m_b}{n-1-j}}{\binom{m_a+m_b}{n}} \\
&= m_a \sum_{j=0}^{n-1} \frac{\binom{m_a-1}{j}\binom{m_b}{n-1-j}}{\binom{m_a+m_b}{n}}.
\end{aligned}$$

Or la somme des probabilités élémentaires de la loi hypergéométrique de paramètres $(m_a-1, m_b, n-1)$ doit être égale à 1 :

$$\sum_{j=0}^{n-1} \frac{\binom{m_a-1}{j}\binom{m_b}{n-1-j}}{\binom{m_a+m_b-1}{n-1}} = 1\,,$$

ce qui implique que

$$\mathbb{E}[X] = m_a \frac{\binom{m_a+m_b-1}{n-1}}{\binom{m_a+m_b}{n}} = \frac{m_a n}{m_a + m_b}\,.$$

Le calcul de la variance s'effectue de manière similaire (laissé en exercice). ■

Le loto. On considère tout d'abord la forme classique du loto, qui a été utilisée de 1976 à 2008. Le joueur coche 6 numéros sur une grille en comportant 49, dans le cas d'un bulletin simple. Les 6 numéros gagnants sont déterminés par tirage au sort. L'espace fondamental est ici l'ensemble des parties à 6 éléments (« tirage ») de $[\![1, 49]\!]$ muni de la probabilité uniforme. Son cardinal est $\mathrm{card}\,(\Omega) = \binom{49}{6} = 13\ 983\ 816$. Notant N le nombre de numéros gagnants figurant parmi les numéros cochés par le joueur (on rappelle que la grille du joueur comporte 6 numéros cochés et 43 non cochés), l'événement $\{N = n\}$ $(n \in [\![0, 6]\!])$ est réalisé si le tirage produit n numéros cochés et $6 - n$ non cochés. La v.a. N a donc pour loi la loi hypergéométrique de paramètres $(6, 43, 6)$ (voir page 715) :

$$\mathbb{P}(N = n) = \frac{\binom{6}{n}\binom{43}{6-n}}{\binom{49}{6}}, \quad n \in [\![0, 6]\!]\,,$$

et certaines valeurs numériques sont données dans le tableau ci-dessous. Au premier tirage du 10 mai 2006, on recevait pour une mise de $0{,}3$ euros (soit une grille) le gain $G = g(N)$ suivant :

n numéros gagnants	gain $g(n)$	probabilité $\mathbb{P}(N = n)$
6	789 177,00 euros	$7{,}2\ 10^{-8}$
5	1 813,80 euros	$1{,}8\ 10^{-5}$
4	30,70 euros	$9{,}7\ 10^{-4}$
3	2,70 euros	$1{,}8\ 10^{-2}$

Le gain moyen est $\mathbb{E}[G] = 2{,}70 \times 1{,}8\ 10^{-2} + 30{,}70 \times 9{,}7\ 10^{-4} + \cdots \simeq 0{,}168$ euros, tandis que l'écart-type $\sigma(G) = (\mathbb{E}[G^2] - \mathbb{E}[G]^2)^{1/2} \simeq 212$ euros. On voit que le jeu est défavorable au joueur, dont le bénéfice moyen est $\mathbb{E}[G] - 0{,}3 = -0{,}132$ euros, et la grande valeur de σ vient de ce que parfois (mais très rarement), cela rapporte gros.
La forme moderne du loto, mise en place à partir de 2008, est légèrement différente : le joueur coche 5 numéros sur une grille en comportant 49 et un « numéro chance » sur une autre grille en comportant 10. Donc la probabilité qu'un joueur trouve la bonne combinaison est :

$$p = \frac{1}{\binom{49}{5}} \frac{1}{10} = \frac{1}{19\ 068\,840} \simeq 5{,}2\,10^{-8}\,,$$

car il y a $\binom{49}{5}$ choix possibles des cinq premiers numéros, et 10 choix possibles du « numéro chance ». Notez que l'on a moins de chance de gagner le gros lot avec la nouvelle version qu'avec la version classique.

Poursuivons la discussion engagée page 715 au sujet des deux méthodes possibles d'échantillonnage (avec ou sans remise). Si l'on tire n individus au hasard dans une population se divisant en une classe de m_a individus de type A et m_b individus de type B, alors le nombre d'individus tirés dans la classe A suit une loi binomiale de paramètres (n, p), avec $p = m_a/(m_a + m_b)$, dans le cas d'un tirage avec remise, et une loi hypergéométrique de paramètres (m_a, m_b, n) dans le cas sans remise. Dans les deux cas, l'espérance de la variable aléatoire est $nm_a/(m_a + m_b)$, comme le montrent les propositions précédentes. De plus, si m_a et m_b sont nettement plus grands que n, alors la variance de la variable aléatoire hypergéométrique vérifie :

$$\mathrm{Var}(X) = \frac{n(m_a + m_b - n)m_a m_b}{(m_a + m_b - 1)(m_a + m_b)^2} \simeq n\frac{m_a m_b}{(m_a + m_b)^2} = np(1-p)\,,$$

c'est-à-dire qu'elle est à peu près égale à la variance de la variable aléatoire binomiale du cas avec remise. Cette comparaison peut être menée encore plus loin, comme nous allons le voir dans la section suivante.

4 Problèmes d'approximation

On a examiné plusieurs lois usuelles, qui interviennent naturellement dans des problèmes classiques. Plusieurs d'entre elles comprennent des facteurs exponentiels ou factoriels, ce qui laisse présager la manipulation de très grands nombres lorsque les paramètres des lois deviennent grands. Il apparaît cependant que, lorsque les paramètres des lois deviennent grands, on peut souvent approcher les lois correspondantes par des lois limites plus simples. On verra une théorie générale sur ces théorèmes limites dans le module IV.4, mais on peut déjà rendre compte de résultats d'approximation élémentaires mais très importants en pratique.

4.1 Approximation de la loi hypergéométrique

La loi hypergéométrique intervient lorsque l'on cherche à décrire le résultat d'un tirage sans remise de n individus dans une population divisée en deux groupes de m_a individus du groupe A et m_b individus du groupe B. Le nombre total S d'individus du groupe A tirés suit la loi hypergéométrique de paramètres (m_a, m_b, n) :

$$\mathbb{P}(S = k) = \frac{\binom{m_a}{k}\binom{m_b}{n-k}}{\binom{m_a+m_b}{n}}\,, \qquad k \in [\![0, n]\!]\,.$$

La formule définissant cette loi n'est pas très pratique, surtout lorsque l'on cherche à l'appliquer dans des situations où m_a et m_b sont très grands. Si l'on pense à l'exemple du sondage d'opinion, m_a et m_b sont de l'ordre de plusieurs millions, et n est de l'ordre de mille. Dans une telle situation, où n est nettement plus petit que m_a et m_b, on introduit le cardinal total de la population $m_a + m_b$ et la fraction $p \in [0, 1]$ de la population totale qui appartient au groupe A :

$$p = \frac{m_a}{m}\,, \qquad 1 - p = \frac{m_b}{m}\,\cdot$$

En développant les facteurs combinatoires :

$$\mathbb{P}(S=k)=\frac{m_a!}{k!(m_a-k)!}\frac{m_b!}{(n-k)!(m_b-n+k)!}\frac{n!(m-n)!}{m!}\,, \tag{22}$$

on peut réécrire :

$$\begin{aligned}\frac{m_a!}{(m_a-k)!}&=m_a\times(m_a-1)\times\cdots\times(m_a-k+1)\\&=m_a^k\times 1\times\left(1-\frac{1}{m_a}\right)\times\cdots\times\left(1-\frac{k-1}{m_a}\right)\,,\end{aligned}$$

donc, avec $m_a=pm$, on trouve que :

$$\frac{m_a!}{(m_a-k)!}\overset{m\gg 1}{\simeq}p^km^k\,.$$

En appliquant le même raisonnement aux autres facteurs de (22), on obtient :

$$\mathbb{P}(S=k)\overset{m\gg 1}{\simeq}\frac{p^km^k}{k!}\frac{(1-p)^{n-k}m^{n-k}}{(n-k)!}\frac{n!}{m^n}\,,$$

et donc :

$$\mathbb{P}(S=k)\overset{m\to+\infty}{\longrightarrow}\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}=\binom{n}{k}p^k(1-p)^{n-k}\,. \tag{23}$$

On reconnaît la loi binomiale de paramètres (n,p).

Proposition 14. La loi hypergéométrique de paramètres $(mp,m(1-p),n)$ peut être approchée lorsque m tend vers $+\infty$ par la loi binomiale de paramètres (n,p).

Ce résultat n'est pas surprenant, car il exprime seulement le fait que la différence entre un tirage avec remise et un tirage sans remise s'efface lorsque la population sur laquelle on effectue le tirage devient très grande. Cela vient du fait que, lors du tirage avec remise, la probabilité de tirer un même individu plus d'une fois tend vers 0.

Exercice 1.

Montrer que, lors du tirage avec remise de n individus dans une population totale de m individus, la probabilité de l'événement $A=$« les n individus tirés sont tous distincts » est :

$$\mathbb{P}(A)=\prod_{k=0}^{n-1}\left(1-\frac{k}{m}\right)\overset{m\gg n}{\simeq}\exp\left(-\frac{n(n-1)}{2m}\right)\,.$$

Solution.

On introduit la suite décroissante d'événements $A_j=$« les j individus tirés lors des j premiers tirages sont tous distincts ». On a bien sûr $A_1=\Omega$, $A_n=A$, et pour tout $j\in[\![2,n]\!]$:

$$\mathbb{P}(A_j|A_{j-1})=\frac{m-j+1}{m},$$

et donc

$$\mathbb{P}(A) = \mathbb{P}(A_n|A_{n-1})\mathbb{P}(A_{n-1}) = \Big[\prod_{j=2}^{n} \mathbb{P}(A_j|A_{j-1})\Big]\mathbb{P}(A_1) = \prod_{j=2}^{n} \frac{m-j+1}{m},$$

ce qui donne le résultat espéré :

$$\mathbb{P}(A) = \prod_{k=0}^{n-1} \left(1 - \frac{k}{m}\right).$$

Lorsque m devient grand, on a $1 - k/m \simeq \exp(-k/m)$ et donc

$$\mathbb{P}(A) \simeq \exp\left(-\sum_{k=0}^{n-1} \frac{k}{m}\right) = \exp\left(-\frac{n(n-1)}{2m}\right).$$

4.2 Approximation de la loi binomiale, loi de Poisson

Définition 8. Soit θ un réel strictement positif. La *loi de Poisson* $\mathcal{P}(\theta)$ de paramètre θ est la probabilité sur $\big(\mathbb{N}, \mathcal{P}(\mathbb{N})\big)$ déterminée par :

$$P(\{k\}) := e^{-\theta}\frac{\theta^k}{k!}, \quad k \in \mathbb{N}. \tag{24}$$

Le développement en série entière de la fonction exponentielle nous assure qu'il s'agit bien d'une loi de probabilité :

$$\sum_{k\in\mathbb{N}} P(\{k\}) = e^{-\theta}\sum_{k\in\mathbb{N}} \frac{\theta^k}{k!} = e^{-\theta}e^{\theta} = 1.$$

D'autre part l'espérance de cette loi coïncide avec son paramètre puisque, si X suit la loi $\mathcal{P}(\theta)$:

$$\mathbb{E}[X] = \sum_{k\in\mathbb{N}} kP(\{k\}) = \sum_{k\geqslant 1} e^{-\theta}\frac{\theta^k}{(k-1)!} = e^{-\theta}\theta\sum_{k\geqslant 0}\frac{\theta^k}{k!} = \theta.$$

On calcule de la même manière $\mathbb{E}[X^2] = \theta^2 + \theta$ et donc $\mathrm{Var}(X) = \theta$. Les lois de Poisson s'introduisent le plus souvent dans les modèles aléatoires comme limites de lois binomiales, dans les conditions précisées par le lemme suivant.

Lemme 15. Soit $\theta > 0$. Si $n \in \mathbb{N}^*$ et $p_n = \theta/n$, alors la loi binomiale $\mathcal{B}(n, p_n)$ peut être approchée par la loi de Poisson $\mathcal{P}(\theta)$ quand n tend vers $+\infty$.

Cela veut dire que, si $X \sim \mathcal{B}(n, \theta/n)$, alors $\forall k \in \mathbb{N}$:

$$\mathbb{P}(X = k) \overset{n\to+\infty}{\longrightarrow} e^{-\theta}\frac{\theta^k}{k!}. \tag{25}$$

Démonstration. En effet, pour tout entier $k \geqslant 0$ fixé, dès que $n \geqslant k$, on a :

$$\begin{aligned}\mathbb{P}(X = k) &= \binom{n}{k}(p_n)^k(1-p_n)^{n-k} \\ &= (1-p_n)^n \times \frac{(np_n)^k}{k!} \times \frac{n(n-1)\cdots\times(n-k+1)}{n^k} \times (1-p_n)^{-k}\,.\end{aligned}$$

Lorsque n tend vers $+\infty$, les quatre facteurs du dernier membre tendent respectivement vers $e^{-\theta}$, $\theta^k/k!$, 1 et 1, ce qui donne le résultat. ■

Ainsi est-il fréquent d'approcher la loi binomiale $\mathcal{B}(n,p)$ par la loi de Poisson de paramètre np lorsque 1) le nombre n de tirages est grand, 2) la probabilité p de succès est faible, et 3) le produit np est de l'ordre de l'unité (en pratique, on peut considérer que l'approximation est bien justifiée lorsque $n \geqslant 50$ et $np \leqslant 5$).

5 Familles de variables aléatoires

Dans la plupart des expériences aléatoires, il apparaît naturellement non pas une seule variable aléatoire, mais plusieurs.

Définition 9. Soit $n \geqslant 2$. Si $X_1, \ldots, X_n$ sont des v.a. définies sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, alors $(X_1, \ldots, X_n)$ est appelé *vecteur aléatoire*, et si $n = 2$, *couple aléatoire*.

5.1 Loi d'un vecteur aléatoire

Proposition 16. Soient $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité et $n \geqslant 2$. Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite finie de v.a. discrètes définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans E_i ($i \in [\![1,n]\!]$).
Le vecteur aléatoire $X := (X_1, \ldots, X_n)$ est une v.a. discrète au sens de la définition 1, à valeurs dans l'espace discret $E := E_1 \times \cdots \times E_n$. Sa loi est donnée par les nombres :

$$p_X(x) := \mathbb{P}(X = x), \quad x := (x_1, \ldots, x_n) \in E\,,$$

où :

$$\{X = x\} = \{X_1 = x_1, \ldots, X_n = x_n\} = \bigcap_{i=1}^{n}\{X_i = x_i\}\,.$$

On appelle souvent la loi du vecteur X, *loi jointe*, pour la distinguer des lois p_{X_i} des composantes X_i ($i \in [\![1,n]\!]$) que l'on appelle *lois marginales*. On obtient ces lois marginales en sommant la loi jointe sur les v.a. que l'on veut éliminer.

Proposition 17 (Formule des marginales). Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire. La loi de X_1 se retrouve à partir de la loi de X par la formule :

$$p_{X_1}(x_1) = \sum_{x_2 \in E_2, \ldots, x_n \in E_n} p_X(x_1, x_2, \ldots, x_n)\,. \tag{26}$$

Démonstration. Il suffit de noter que :

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \mathbb{P}\left(\bigcup_{x_2, \ldots, x_n \in E_2 \times \cdots \times E_n} \{X = (x_1, \ldots, x_n)\} \right)\,,$$

et d'appliquer ensuite la σ-additivité de la probabilité $\mathbb{P}$. ∎

On obtient de la même manière la loi d'un sous-vecteur de X, par exemple :

$$p_{(X_1,X_2)}(x_1, x_2) = \sum_{x_3 \in E_3, \ldots, x_n \in E_n} p_X(x_1, \ldots, x_n)\,.$$

Pour toute fonction $f : E \to \mathbb{R}$, l'espérance de la v.a. discrète réelle $Y := f(X)$ est encore définie par (8) sous réserve que Y soit intégrable ou de manière équivalente $\sum_{x \in E} \big|f(x)\big| p_X(x) < +\infty$ ou $\sum_y |y| p_Y(y) < +\infty$, on a $\mathbb{E}[f(X)] = \sum_y y p_Y(y)$, l'équation (10) s'applique encore, et l'on calculera $\mathbb{E}[f(X)]$ à l'aide de (11), c'est-à-dire ici :

$$\mathbb{E}[f(X_1, \ldots, X_n)] = \sum_{x_1 \in E_1, \ldots, x_n \in E_n} f(x_1, \ldots, x_n) p_X(x_1, \ldots, x_n)\,.$$

La proposition 5 de la page 718 donnant les propriétés de $\mathbb{E}$ s'applique toujours.

Loi multinomiale. Dans une population de taille m divisée en 3 groupes de tailles respectives m_1, m_2, m_3 ($m_1 + m_2 + m_3 = m$), on choisit successivement n individus au hasard (tirage avec remise : on peut choisir plusieurs fois le même individu). On s'intéresse au triplet (N_1, N_2, N_3) avec $N_i :=$ nombre d'individus du $i^{\text{ème}}$ groupe tirés, $i = 1, 2, 3$. Les variables N_i sont liées par la relation $N_1 + N_2 + N_3 = n$. Chaque tirage avec remise est une application :

$$\omega : [\![1, n]\!] \to \{1, 2, 3\}\,,$$

avec $\omega(j)$ le numéro du groupe où est tiré le $j^{\text{ème}}$ individu. D'après l'équiprobabilité :

$$\mathbb{P}(\{\omega\}) = \frac{m_1^{n_1} m_2^{n_2} m_3^{n_3}}{m^n} = p_1^{n_1} p_2^{n_2} p_3^{n_3}\,,$$

avec $n_i = N_i(\omega)$ et $p_i = m_i/m$, $i = 1, 2, 3$. Il faut ensuite faire un peu de combinatoire pour calculer, pour tout triplet (n_1, n_2, n_3) d'entiers de somme égale à n, le nombre $C(\omega)$ d'applications ω telles que $N_i(\omega) = n_i$ $\forall i \in [\![1, 3]\!]$. On choisit d'abord le sous-ensemble $\omega^{-1}(\{1\})$ de $[\![1, n]\!]$ des indices des tirages dans le groupe 1 (soit $\binom{n}{n_1}$ possibilités), puis le sous-ensemble $\omega^{-1}(\{2\})$ dans $[\![1, n]\!] \backslash \omega^{-1}(\{1\})$ des indices des tirages dans le groupe 2 (soit $\binom{n-n_1}{n_2}$ possibilités). Les autres tirages sont alors forcément dans le

groupe 3. Donc le nombre $C(\omega)$ est donné par le coefficient multinomial $n!/(n_1!n_2!n_3!)$ et :

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \frac{n!}{n_1!n_2!n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}\,, \tag{27}$$

pour n_1, n_2, n_3 entiers tels que $n_1 + n_2 + n_3 = n$. Cette loi s'appelle la loi multinomiale.

Exercice 2.

On reprend l'exemple précédent (loi multinomiale).

1. Généraliser la formule (27) de la loi multinomiale, au cas où la population est divisée en d classes.
2. Vérifier que les lois marginales des N_i sont binomiales.
3. Vérifier que, pour $d = 2$, on a $(N_1, N_2) = (S, n - S)$ où S suit une loi binomiale de paramètres (n, p) avec $p = p_1$ (*cf.* page 711).

Solution. 1. Dans une population de taille m divisée en d groupes de tailles respectives $m_1, \ldots, m_d$ (avec $m_1 + \cdots + m_d = m$), on choisit successivement n individus au hasard avec remise. On s'intéresse au d-uplet $(N_1, \ldots, N_d)$ avec $N_i :=$ nombre d'individus du $i^{ème}$ groupe tirés, $i \in [\![1, d]\!]$. En notant $p_j = m_j/m$, $j \in [\![1, d]\!]$, on trouve :

$$\mathbb{P}(N_1 = n_1, \ldots, N_d = n_d) = \frac{n!}{n_1! \cdots n_d!} p_1^{n_1} \cdots p_d^{n_d}\,,$$

pour tout d-uplet d'entiers $(n_1, \ldots, n_d)$ tel que $n_1 + \cdots + n_d = n$, et $\mathbb{P}(N_1 = n_1, \ldots, N_d = n_d) = 0$ si $n_1 + \cdots + n_d \neq n$.
2. La loi marginale de N_1 s'obtient pas la formule des marginales :

$$\begin{aligned}
\mathbb{P}(N_1 = n_1) &= \sum_{n_2, \ldots, n_d \geqslant 0} \mathbb{P}(N_1 = n_1, \ldots, N_d = n_d) \\
&= \frac{n!}{n_1!} p_1^{n_1} \sum_{n_2, \ldots, n_d \geqslant 0, \sum_{j=2}^{d} n_j = n - n_1} \frac{1}{n_2! \cdots n_d!} p_2^{n_2} \cdots p_d^{n_d} \\
&= \frac{n!}{(n - n_1)! n_1!} p_1^{n_1} (1 - p_1)^{n - n_1} \sum_{n_2, \ldots, n_d \geqslant 0, \sum_{j=2}^{d} n_j = n - n_1} \frac{(n - n_1)!}{n_2! \cdots n_d!} \prod_{j=2}^{d} \Big(\frac{p_j}{1 - p_1}\Big)^{n_j}\,,
\end{aligned}$$

si $n_1 \in [\![0, n]\!]$, et $\mathbb{P}(N_1 = n_1) = 0$ sinon. On reconnaît dans la somme finale les probabilités d'une loi multinomiale, dont la somme est égale à 1, et il vient :

$$\mathbb{P}(N_1 = n_1) = \frac{n!}{(n - n_1)! n_1!} p_1^{n_1} (1 - p_1)^{n - n_1}\,,$$

qui est bien une loi binomiale de paramètres (n, p_1).
3. En posant $S = N_1$, on a $N_2 = n - S$. Comme S suit la loi binomiale de paramètres (n, p_1), on a le résultat désiré.

5.2 Covariance et corrélation

Dans le cas d'un couple de v.a., les notions d'espérance et de variance jouent un rôle fondamental pour caractériser les coordonnées séparément. Afin de les étudier simultanément, on introduit la notion de covariance et de corrélation de deux v.a.

Définition 10. Soient X et Y deux v.a. de carré intégrable définies sur le même espace de probabilité. On appelle *covariance* de X et de Y, le nombre :

$$\mathrm{Cov}(X,Y) := \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] . \tag{28}$$

Si $\mathrm{Var}(X) > 0$ et $\mathrm{Var}(Y) > 0$, le *coefficient de corrélation* $\mathrm{Cor}(X,Y)$ est le nombre défini par :

$$\mathrm{Cor}(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \cdot \tag{29}$$

Le fait que XY soit intégrable si X et Y sont de carré intégrable est établi dans la preuve de la proposition 19. Notez qu'en développant le membre de droite de (28) on obtient que la covariance peut aussi s'écrire :

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[XY\right] - \mathbb{E}[X]\mathbb{E}[Y] .$$

Ces quantités décrivent les variations simultanées des v.a. X et Y. Si X et Y « varient dans le même sens », leur covariance et donc leur corrélation est positive. Par exemple, $\mathrm{Cov}(X,X) = \mathrm{Var}(X) \geqslant 0$, tandis que $\mathrm{Cov}(X,-X) = -\mathrm{Var}(X) \leqslant 0$.

Lemme 18. L'ensemble des v.a. de carré intégrable sur un même espace de probabilité est un espace vectoriel.

Démonstration. Soient X et Y deux v.a. de carré intégrable définies sur le même espace de probabilité, et $\mu, \lambda \in \mathbb{R}$. Comme le montre la démonstration de la proposition 5 de la page 718, $Z := \lambda X + \mu Y$ est une v.a. discrète réelle. Remarquons que, si $x, y \in \mathbb{R}$, alors $2xy \leqslant x^2 + y^2$ car $x^2 + y^2 - 2xy = (x-y)^2 \geqslant 0$. On a donc :

$$Z^2 = \lambda^2 X^2 + \mu^2 Y^2 + 2\lambda\mu XY \leqslant 2\lambda^2 X^2 + 2\mu^2 Y^2 \leqslant \kappa(X^2 + Y^2) ,$$

avec $\kappa = 2\max(\lambda^2, \mu^2)$. Comme $X^2 + Y^2$ est intégrable, Z^2 l'est aussi. ■

Proposition 19 (Inégalité de Cauchy-Schwarz). Soient X et Y deux v.a. de carré intégrable définies sur le même espace de probabilité. Alors XY est intégrable et l'on a :

$$\mathbb{E}[XY]^2 \leqslant \mathbb{E}[X^2]\mathbb{E}[Y^2] .$$

Démonstration. Remarquons que si $x, y \in \mathbb{R}$, alors $|xy| \leqslant (x^2 + y^2)/2$.

Donc on a $|XY| \leqslant (X^2 + Y^2)/2$. Comme X et Y sont de carré intégrable, par linéarité, $(X^2 + Y^2)/2$ est de carré intégrable. On en déduit que $|XY|$, et donc XY, sont intégrables.

Si $\mathbb{E}[Y^2] = 0$, cela implique que $Y = 0$ avec probabilité 1, et donc $XY = 0$ avec probabilité 1 et l'inégalité de la proposition est triviale. Supposons que $\mathbb{E}[Y^2] > 0$. Par positivité et linéarité de l'espérance, on a pour tout $\lambda \in \mathbb{R}$:

$$0 \leqslant \mathbb{E}[(X - \lambda Y)^2] = \mathbb{E}[X^2] - 2\lambda\mathbb{E}[XY] + \lambda^2\mathbb{E}[Y^2]\,.$$

En prenant $\lambda = \mathbb{E}[XY]/\mathbb{E}[Y^2]$, on obtient :

$$0 \leqslant \mathbb{E}[X^2] - \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}\,,$$

ce qui donne l'inégalité recherchée. ∎

L'inégalité de Cauchy-Schwarz montre que le coefficient de corrélation est un nombre réel compris entre -1 et 1.

Corollaire 20. Soient X et Y deux v.a. de carré intégrable définies sur le même espace de probabilité telles que $\mathrm{Var}(X) > 0$ et $\mathrm{Var}(Y) > 0$. Alors $\mathrm{Cor}(X, Y) \in [-1, 1]$.

Démonstration. Il suffit d'appliquer l'inégalité de Cauchy-Schwarz aux variables centrées $X' = X - \mathbb{E}[X]$ et $Y' = Y - \mathbb{E}[Y]$. ∎

La variance, à la différence de l'espérance, n'est pas additive. Cependant l'importance de la notion de covariance vient de la relation :

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)\,,$$

dont la forme générale est donnée dans la proposition suivante :

Proposition 21. La variance de la somme $S_n = \sum_{i=1}^{n} X_i$ de n v.a. réelles de carré intégrable définies sur le même espace de probabilité est donnée par la formule :

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \leqslant l < m \leqslant n} \mathrm{Cov}(X_l, X_m)\,. \tag{30}$$

En fait la variance définit une forme quadratique sur l'espace vectoriel des v.a. discrètes de carré intégrable définies sur un même espace de probabilité. La covariance est la forme bilinéaire associée à cette forme quadratique.

Démonstration. Puisque $\mathbb{E}[S_n] = \sum_{i=1}^{n} \mathbb{E}[X_i]$, on a :

$$S_n - \mathbb{E}[S_n] = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\,,$$

et donc

$$(S_n - \mathbb{E}[S_n])^2 = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])^2 + 2 \sum_{1 \leqslant l < m \leqslant n} (X_l - \mathbb{E}[X_l])(X_m - \mathbb{E}[X_m])\,.$$

La formule de la proposition s'obtient alors en prenant l'espérance des deux membres. ∎

5.3 Indépendance

Définition 11. Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite finie de v.a. discrètes définies sur le même espace de probabilité à valeurs dans des espaces $(E_i)_{i\in[\![1,n]\!]}$. On dit que les v.a. sont *indépendantes* si, pour tout $x_1 \in E_1, \ldots, x_n \in E_n$, on a :

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right) . \tag{31}$$

Exemple. Dans le cas du jeu de Pile ou Face de durée n (décrit page 711), nous avons défini $\Omega := \{F, P\}^n$ et $\mathbb{P}(\{\omega\}) := p^k(1-p)^{n-k}$ pour tout résultat comportant exactement k Piles. Pour $i \in [\![1, n]\!]$, on note $X_i : \Omega \to \{0, 1\}$ la v.a. de Bernoulli :

$$X_i := \mathbf{1}_{A_i} ,$$

où $A_i := \{\omega \in \Omega \mid \omega(i) = P\}$ est l'événement « le $i^{ème}$ lancer est un Pile ».
Soient $x_i \in \{0, 1\}$, $i \in [\![1, n]\!]$. On se rend compte que la conjonction d'événements :

$$\{X_1 = x_1, \ldots, X_n = x_n\} = \bigcap_{i=1}^{n} \{X_i = x_i\}$$

n'est réalisée que pour un unique $\widetilde{\omega} \in \Omega$ qui est tel que la $i^{ème}$ coordonnée $\widetilde{\omega}(i)$ est P si $x_i = 1$ et F sinon. Donc :

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(\{\widetilde{\omega}\}) = p^{\bar{x}}(1-p)^{n-\bar{x}} ,$$

où $\bar{x} := \sum_{i=1}^{n} x_i$. D'autre part, pour tout $i \in [\![1, n]\!]$:

$$\mathbb{P}(X_i = x_i) = p^{x_i}(1-p)^{1-x_i} .$$

On en déduit que les X_i sont indépendantes. Il s'agit donc d'une suite de v.a. indépendantes et identiquement distribuées selon la loi de Bernoulli de paramètre p.

> ***Remarque***. Si $(X_i)_{i\in[\![1,n]\!]}$ est une suite de v.a. indépendantes, alors pour toute sous-famille d'indices $I \subset [\![1, n]\!]$, $(X_i)_{i\in I}$ est une suite de v.a. indépendantes. Il suffit d'appliquer la formule des marginales.

Proposition 22. Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite finie de v.a. discrètes définies sur le même espace de probabilité. Les v.a. sont indépendantes si, et seulement si, pour toutes fonctions à valeurs réelles $f_1, \ldots, f_n$ telles que $f_i(X_i)$ est intégrable pour tout $i \in [\![1, n]\!]$, on a :

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}\left[f_i(X_i)\right] . \tag{32}$$

Démonstration. On note E_i les espaces des valeurs prises par X_i.
Montrons le sens direct. Soient f_i, $i \in [\![1, n]\!]$, des fonctions à valeurs réelles telles

que $f_i(X_i)$ est intégrable. On considère :

$$I_f := \sum_{x_1 \in E_1, \ldots, x_n \in E_n} \left(\prod_{i=1}^{n} \left| f_i(x_i) \right| \right) \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \,.$$

Comme les v.a. sont indépendantes :

$$I_f = \sum_{x_1 \in E_1, \ldots, x_n \in E_n} \left(\prod_{i=1}^{n} \left| f_i(x_i) \right| \right) \left(\prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \right) \,,$$

et en factorisant (avec l'aide du théorème de Fubini valable pour des séries à termes positifs) :

$$I_f = \prod_{i=1}^{n} \sum_{x_i \in E_i} \left| f_i(x_i) \right| \mathbb{P}(X_i = x_i) = \prod_{i=1}^{n} \mathbb{E}\left[\left| f_i(X_i) \right| \right] \,,$$

qui est fini car les $f_i(X_i)$ sont intégrables. Donc la v.a. $\prod_{i=1}^{n} f_i(X_i)$ est intégrable. On peut donc reprendre la démonstration sans les $|\cdot|$, et avec l'aide du théorème de Fubini valable pour les séries absolument convergentes, on obtient (32).

Montrons la réciproque. Soient $x_i \in E_i$, $i \in [\![1, n]\!]$. On applique (32) avec les fonctions indicatrices $f_i := \mathbf{1}_{\{x_i\}}$:

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n\right) = \mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i) \right] = \prod_{i=1}^{n} \mathbb{E}\left[f_i(X_i) \right] = \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right) \,,$$

ce qui donne le résultat. ∎

Définition 12. Soit $(X_i)_{i \in \mathbb{N}}$ une suite de v.a. discrètes définies sur le même espace de probabilité à valeurs dans des espaces $(E_i)_{i \in \mathbb{N}}$. On dit que les v.a. sont *indépendantes* si, pour toute sous-famille d'indices finie $I \subset \mathbb{N}$, la famille $(X_i)_{i \in I}$ est indépendante au sens de la définition 11.

D'après la remarque de la page 735, il suffit de montrer que, pour tout $n \in \mathbb{N}$, la suite $(X_i)_{i \in [\![0,n]\!]}$ est indépendante au sens de la définition 11 de la page précédente pour assurer que la suite infinie $(X_i)_{i \in \mathbb{N}}$ est indépendante.

5.4 Indépendance et covariance

S'il est facile de montrer que deux v.a. indépendantes ont une covariance nulle, il faut cependant résister à la tentation et ne pas croire en la réciproque : deux v.a. ayant une covariance nulle ne sont pas forcément indépendantes !

Proposition 23. *1.* Si deux v.a. de carré intégrable définies sur le même espace de probabilité sont indépendantes, alors leur covariance est nulle.
2. La réciproque est fausse en générale.
3. Deux v.a. de Bernoulli sont indépendantes si, et seulement si, leur covariance est nulle.

Démonstration. *1.* Soient X et Y deux v.a. indépendantes. Alors :

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0\,.$$

2. Il existe des v.a. X et Y qui ne sont pas indépendantes, et dont la covariance est nulle. En effet, considérons par exemple l'espace $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ avec $\Omega := \{-1,0,1\}$ et $\mathbb{P}(\{-1\}) := 1/4$, $\mathbb{P}(\{0\}) := 1/2$ et $\mathbb{P}(\{1\}) := 1/4$. La v.a. $X(\omega) := \omega$ est une v.a. réelle discrète à valeurs dans $\{-1,0,1\}$ telle que :

$$\mathbb{P}(X=-1) = \frac{1}{4}\,, \quad \mathbb{P}(X=0) = \frac{1}{2}\,, \quad \mathbb{P}(X=1) = \frac{1}{4}\,\cdot$$

Introduisons la v.a réelle discrète $Y := |X|$. On trouve successivement :

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{4}\times(-1) + \frac{1}{2}\times 0 + \frac{1}{4}\times 1 = 0\,,\\
\mathbb{E}[Y] &= \frac{1}{4}\times 1 + \frac{1}{2}\times 0 + \frac{1}{4}\times 1 = \frac{1}{2}\,,\\
\mathbb{E}[XY] &= \frac{1}{4}\times(-1)\times 1 + \frac{1}{2}\times 0\times 0 + \frac{1}{4}\times 1\times 1 = 0\,,
\end{aligned}$$

et donc $\mathrm{Cov}(X,Y) = 0$. Et pourtant les v.a. X et Y ne sont pas indépendantes :

$$\mathbb{P}(X=1, Y=0) = 0 \neq \frac{1}{8} = \frac{1}{4}\times\frac{1}{2} = \mathbb{P}(X=1)\mathbb{P}(Y=0)\,.$$

3. Le sens direct est une conséquence du premier point. Montrons la réciproque. Soit un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, sur lequel sont définies deux v.a. de Bernoulli X et Y telles que $\mathrm{Cov}(X,Y) = 0$. Notons A l'événement $\{X=1\}$ et B l'événement $\{Y=1\}$. On a $X = \mathbf{1}_A$ et $Y = \mathbf{1}_B$, donc :

$$0 = \mathrm{Cov}(X,Y) = \mathbb{E}[\mathbf{1}_A\mathbf{1}_B] - \mathbb{E}[\mathbf{1}_A]\mathbb{E}[\mathbf{1}_B] = \mathbb{E}[\mathbf{1}_{A\cap B}] - \mathbb{E}[\mathbf{1}_A]\mathbb{E}[\mathbf{1}_B]\,,$$

la dernière égalité venant du fait que $\mathbf{1}_A(\omega)\mathbf{1}_B(\omega) = 1$ si, et seulement si, $\omega \in A$ et $\omega \in B$, soit $\omega \in A\cap B$. Comme $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$, on obtient donc que :

$$0 = \mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\,,$$

ce qui montre que les événements A et B sont indépendants, et l'on peut donc écrire :

$$\mathbb{P}(\{X=1\}\cap\{Y=1\}) = \mathbb{P}(X=1)\mathbb{P}(Y=1)\,.$$

Par la proposition 11 de la page 689, les couples d'événements (A, B^c), (A^c, B) et (A^c, B^c) sont indépendants, donc, en tenant compte du fait que $A^c = \{X=0\}$ et $B^c = \{Y=0\}$, on obtient que pour tous $i, j \in \{0,1\}$:

$$\mathbb{P}(\{X=i\}\cap\{Y=j\}) = \mathbb{P}(X=i)\mathbb{P}(Y=j)\,,$$

ce qui démontre l'indépendance de X et Y. ∎

Le troisième point de la proposition montre qu'une covariance nulle pour deux v.a. de Bernoulli assure l'indépendance, mais une telle affirmation est spécifique aux v.a. de Bernoulli et ne peut pas être généralisée, comme le dit le deuxième point de la proposition et le contre-exemple associé (dans la démonstration).

Corollaire 24. Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite finie de v.a. discrètes de carré intégrable définies sur le même espace de probabilité. Si les X_i sont indépendantes deux à deux (c'est-à-dire que X_i et X_j sont indépendantes pour tous $i \neq j$), alors on a $\mathrm{Cov}(X_i, X_j) = 0$ pour tous $i \neq j$ et :

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)\,.$$

La réciproque est fausse !

Démonstration. Il suffit d'appliquer la formule (30) en tenant compte du *1.* de la proposition 23 pour obtenir le résultat. Pour se convaincre que la réciproque est fausse, on peut reprendre l'exemple fourni dans la preuve du *2.* de la proposition 23 : sur cet exemple, on a $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, et pourtant X et Y ne sont pas indépendantes. ■

Corollaire 25. Soit un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, sur lequel est définie une suite de n v.a. de Bernoulli indépendantes de même paramètre $p \in [0,1]$, notées X_i, $i \in [\![1,n]\!]$. On note S la somme des X_i :

$$S := \sum_{i=1}^{n} X_i\,.$$

Alors S suit la loi binomiale de paramètres (n,p).

Démonstration. En introduisant les événements indépendants $A_i := \{X_i = 1\}$, on se retrouve exactement dans la situation de la proposition 3 de la page 712, ce qui donne immédiatement le résultat. ■

Application : on peut retrouver facilement les expressions de l'espérance et de la variance d'une v.a. binomiale $\mathcal{B}(n,p)$. En effet, en reprenant les notations du corollaire 25 et en utilisant la linéarité de l'espérance, on obtient :

$$\mathbb{E}[S] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\mathbb{E}[X_1] = np\,.$$

De plus, comme les v.a. X_i sont indépendantes, on a aussi :

$$\mathrm{Var}(S) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\mathrm{Var}(X_1) = np(1-p)\,.$$

5.5 Schéma succès-échec infini

Comme on l'a vu dans l'exemple de la page 735, il est facile de construire une suite finie de v.a. indépendantes, de loi de Bernoulli ou autre. La question est plus délicate lorsque l'on cherche à construire une suite infinie de v.a. indépendantes, car alors l'espace de probabilité sous-jacent ne peut pas être défini en posant la probabilité de chacun de ses événements élémentaires. Si l'on essaye de le faire, on se heurte à la difficulté fondamentale que tous les événements élémentaires (réalisation d'une suite infinie particulière) ont des probabilités nulles !

5.5.1 Schéma de Bernoulli

La proposition suivante présente une construction d'une suite infinie de v.a. de Bernoulli indépendantes. Vous pouvez la passer en première lecture en admettant l'existence de suites infinies de v.a. indépendantes. Ce résultat découle en fait de résultats plus généraux d'existence de v.a. et de suites de v.a. (théorème d'extension de Kolmogorov, voir le livre [Tao]).

Proposition 26 (Schéma de Bernoulli). Il existe un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ et une famille de v.a. $(X_n)_{n\in\mathbb{N}^*}$ définie sur cet espace qui est une suite indépendante et identiquement distribuée selon la loi de Bernoulli de paramètre p.

Démonstration. Nous choisissons l'intervalle $[0,1]$ pour l'espace fondamental Ω, la tribu borélienne sur $[0,1]$ pour $\mathcal{A}$, et la mesure de Lebesgue λ sur $[0,1]$ pour $\mathbb{P}$. Nous proposons la démonstration dans le cas $p = 1/2$.

À chaque réel $\omega \in \Omega$ nous associons son développement dyadique propre :

$$\omega = \sum_{n=1}^{+\infty} 2^{-n} X_n(\omega), \quad X_n(\omega) \in \{0,1\}.$$

Pour rappel sur les développements dyadiques, on peut définir les X_n de la manière suivante :

$$X_n := \mathbf{1}_{A_n} \text{ où } A_n := \bigcup_{k=0}^{2^{n-1}-1} \left[\frac{2k+1}{2^n}, \frac{2k+2}{2^n}\right[.$$

Les ensembles A_n sont des boréliens, car ce sont des unions finies d'intervalles, les fonctions X_n sont donc des v.a. de Bernoulli. Le paramètre de ces lois est $1/2$ car :

$$\begin{aligned}\mathbb{P}(X_n = 1) = \lambda(A_n) &= \lambda\left(\bigcup_{k=0}^{2^{n-1}-1} \left[\frac{2k+1}{2^n}, \frac{2k+2}{2^n}\right[\right) \\ &= \sum_{k=0}^{2^{n-1}-1} \lambda\left(\left[\frac{2k+1}{2^n}, \frac{2k+2}{2^n}\right[\right) = \sum_{k=0}^{2^{n-1}-1} \frac{1}{2^n} = \frac{1}{2}.\end{aligned}$$

Montrons que la suite $(X_n)_{n\in\mathbb{N}^*}$ est une suite de v.a. indépendantes. Il faut donc montrer que, pour tout $n \in \mathbb{N}^*$, la famille $(X_i)_{i\in[\![1,n]\!]}$ est constituée de v.a. indépendantes.

Soient $(x_i)_{i\in[\![1,n]\!]} \in \{0,1\}^n$.

Par construction du développement dyadique l'événement $\{X_1 = x_1, \ldots, X_n = x_n\}$ est l'inter-

valle $I = \left[\sum_{i=1}^{n} 2^{-i}x_i, \sum_{i=1}^{n} 2^{-i}x_i + 2^{-n}\right[$, si bien que :

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \lambda(I) = 2^{-n}\,,$$

qui est égal au produit $\mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n)$. ■

Exercice 3.
Généraliser la preuve de la proposition 26 au cas $p \in [0, 1]$.
On pourra se placer sur le même espace de probabilité et poser pour tout entier n :

$$X_n := \mathbf{1}_{A_n} \text{ où } A_n := \bigcup_{k=0}^{2^{n-1}-1} [a_{2k+1}^{(n)}, a_{2k+2}^{(n)}[\,,$$

et la suite $\{a_k^{(n)} \mid k \in [\![0, 2^n]\!], n \in \mathbb{N}\}$ est définie par la récurrence suivante : $a_0^{(0)} = 0$, $a_1^{(0)} = 1$, et pour tout entier $n \geqslant 1$, $a_{2^n}^{(n)} = 1$ et pour tout entier $k \in [\![0, 2^{n-1} - 1]\!]$:

$$a_{2k}^{(n)} = a_k^{(n-1)}, \quad a_{2k+1}^{(n)} = a_{2k}^{(n)} + p(a_{k+1}^{(n-1)} - a_k^{(n-1)})\,.$$

5.5.2 Les lois géométrique et binomiale revisitées

Sur l'espace de probabilité du schéma de Bernoulli, en interprétant les X_i comme une suite de lancers de Pile ou Face ($\{X_i = 1\}$:=« le $i^{ème}$ lancer est Pile ») on peut définir d'une part la v.a. T comme le premier instant où un Pile sort :

$$T := \inf\{j \geqslant 1 \mid X_j = 1\}\,,$$

et d'autre part on peut introduire pour tout $n \in \mathbb{N}$ les v.a. S_n définies comme le nombre de Pile obtenus lors des n premiers lancers :

$$S_n := \sum_{i=1}^{n} X_i\,.$$

Pour être rigoureux, il faudrait aussi définir la valeur de T (par exemple, 0 ou $+\infty$), lorsque l'on a une réalisation constituée d'une suite infinie de Faces. Cependant, cet événement a une probabilité nulle, et l'on peut donc le négliger.
On vérifie facilement que T suit la loi géométrique de paramètre p tandis que S_n suit la loi binomiale de paramètres (n, p). L'avantage de cette présentation par rapport à celle des exemples des pages 711 et 714 est que l'on peut ici considérer simultanément les v.a. S_n et T car elles sont définies sur le même espace de probabilité. Ainsi peut-on écrire que $\{T > n\} = \{S_n = 0\}$ (c'est-à-dire que le premier pile arrive après le $n^{ème}$ lancer si, et seulement s'il n'y a eu aucun Pile sur les n premiers lancers). On en déduit immédiatement que $\mathbb{P}(T > n) = \mathbb{P}(S_n = 0)$ ce que l'on peut aussi vérifier par le calcul de chacun des membres.

5.5.3 Loi binomiale négative

La loi binomiale permet de décrire le nombre de succès que l'on peut obtenir lors d'une série de n expériences. On peut considérer le problème inverse, c'est-à-dire essayer de décrire le nombre d'expériences qu'il faut faire pour arriver à obtenir k succès, k donné.

Clairement, si $k = 0$, ce nombre est 0. Si $k = 1$, alors on cherche en fait l'instant T_1 du premier succès, et l'on a vu que ce temps suivait une loi géométrique. Le cas général $k \geqslant 1$ est le sujet de ce paragraphe.
Considérons une suite infinie de v.a. de Bernoulli indépendantes et identiquement distribuées X_i de même paramètre p. On peut encore une fois interpréter cette suite comme une suite de lancers de Pile ou Face ($\{X_i = 1\}$:=« le $i^{ème}$ lancer est Pile »). On s'intéresse au temps d'attente T_k du $k^{ème}$ succès (Pile) :

$$T_1 := \inf\{j > 0 \mid X_j = 1\}\,, \qquad T_k := \inf\{j > T_{k-1} \mid X_j = 1\}\,.$$

On note P_k la loi de T_k. On peut la déterminer par des calculs combinatoires. D'une part, il faut au moins k lancers pour obtenir k Piles, donc :

$$P_k(\{n\}) = 0 \ \text{ si } n < k\,. \tag{33}$$

D'autre part, si $n \geqslant k$, pour que $T_k = n$, il faut que le $n^{ème}$ lancer donne Pile, et qu'il y ait $k-1$ lancers entre les instants 1 et $n-1$ qui donnent Pile (et donc, $(n-1)-(k-1)$ lancers qui donnent Face). Donc :

$$P_k(\{n\}) = p \times p^{k-1}(1-p)^{(n-1)-(k-1)}\binom{n-1}{k-1} = p^k(1-p)^{n-k}\binom{n-1}{k-1} \ \text{ si } \ n \geqslant k\,. \tag{34}$$

Définition 13. Soient $k \in \mathbb{N}^*$ et $p \in \,]0,1[$. La *loi binomiale négative* de paramètres (k,p) est la probabilité P sur $\{k, k+1, \ldots\}$ déterminée par :

$$P(\{n\}) := \binom{n-1}{k-1} p^k (1-p)^{n-k}\,, \quad n \geqslant k\,.$$

On la note $\mathcal{NB}(k,p)$.

6 Fonctions génératrices

6.1 Définition

Pour les v.a. à valeurs entières positives, les fonctions génératrices que nous allons introduire permettent fréquemment de simplifier les calculs que l'on peut avoir à faire sur les lois et les espérances et variances de ces v.a.

Définition 14. La *fonction génératrice* G_X d'une v.a. entière positive $X : \Omega \to \mathbb{N}$ est la somme de la série entière :

$$\sum_{x \in \mathbb{N}} z^x p_X(x)\,,$$

dont on note R le rayon de convergence.

Nous utiliserons les propriétés classiques des séries entières (voir module « Suites et séries de fonctions » du livre [L2]).

Proposition 27.
1) G_X est une série entière de rayon de convergence supérieur ou égal à 1, et pour $z \in [-1,1]$:

$$G_X(z) = \mathbb{E}[z^X] = \sum_{x\in\mathbb{N}} z^x p_X(x)\,. \tag{35}$$

2) G_X est une fonction continue sur $[-1,1]$ et indéfiniment différentiable sur $]-1,1[$.
3) G_X détermine entièrement la loi p_X. En particulier, pour tout $n \in \mathbb{N}$:

$$p_X(n) = \frac{1}{n!}\frac{\mathrm{d}^n G_X}{\mathrm{d}z^n}(0)\,. \tag{36}$$

Le fait que la fonction génératrice détermine la loi sera fréquemment utilisé comme suit : on fait un « catalogue » des fonctions génératrices des lois classiques (voir le tableau 6.4.4 à la page 747). Considérant un problème particulier, si l'étude de la fonction génératrice d'une v.a. l'identifie à une fonction du catalogue, alors on en déduit que la loi de cette v.a. est la loi correspondante.

Démonstration. Il est clair que la série entière converge absolument pour $z \in [-1,1]$ puisque $|p_n z^n| \leqslant p_n$ et $\sum\limits_{n\in\mathbb{N}} p_n = 1$. Donc $R \geqslant 1$.

Le caractère $\mathcal{C}^\infty$ de G_X sur $]-1,1[$ découle de résultats classiques sur les séries entières. De plus comme $\sum\limits_n p_n = 1$ la série entière converge normalement pour $z \in [-1,1]$ et sa limite pour z tendant vers 1 en croissant est $G_X(1) = 1$. Cela établit la continuité à gauche de G_X en $+1$ et l'on obtient de même la continuité à droite en -1.
Le 3) découle du lien entre la formule de Taylor et la somme d'une série entière. ∎

La fonction génératrice caractérise entièrement la loi d'une v.a. à valeurs dans $\mathbb{N}$, donc en particulier ses moments. La proposition qui suit donne la relation entre le comportement d'une fonction génératrice en $z = 1$ et les moments d'une v.a.

Proposition 28. Pour tout $k \in \mathbb{N}$, le moment $\mathbb{E}[X^k]$ existe si, et seulement si, la dérivée à gauche d'ordre k au point 1 de la fonction G_X existe et est finie. Dans ce cas :

$$\frac{\mathrm{d}^k G_X}{\mathrm{d}z^k}(1) = \sum_{n\in\mathbb{N}} n(n-1)\cdots(n-k+1)p_X(n) = \mathbb{E}[X(X-1)\cdots(X-k+1)]\,.$$

En particulier, si G_X est au moins deux fois dérivable à gauche en 1, alors X est de carré intégrable et :

$$\mathbb{E}[X] = \frac{\mathrm{d}G_X}{\mathrm{d}z}(1)\,, \qquad \mathbb{E}[X^2] = \frac{\mathrm{d}^2 G_X}{\mathrm{d}z^2}(1) + \frac{\mathrm{d}G_X}{\mathrm{d}z}(1)\,. \tag{37}$$

Démonstration. Nous donnons la preuve pour $k = 1$. Le principe pour k quelconque est le même. Supposons d'abord que $\mathbb{E}[X]$ existe, c'est-à-dire que $\sum\limits_n np_n$ converge. Montrons qu'alors la dérivée à gauche en 1 de G_X existe et est finie. Celle-ci est définie comme la

limite quand z croît vers 1 de :

$$\frac{G_X(z)-G_X(1)}{z-1}=\frac{1-G_X(z)}{1-z}=\sum_{n=0}^{+\infty}p_n\frac{1-z^n}{1-z}=\sum_{n=0}^{+\infty}p_n\left(1+z+\cdots+z^{n-1}\right).$$

Or, si $z\in[0,1]$, alors on a $|1+z+\cdots+z^{n-1}|\leqslant n$. Comme $\sum_n np_n$ converge la série précédente converge normalement et sa limite pour $z\nearrow 1$ est $\mathbb{E}[X]$.

Inversement, supposons que la dérivée à gauche en 1, notée $G'_X(1)$, existe. Appliquons le théorème des accroissements finis à l'intervalle $[z,1]$ et à la fonction G_X. Il existe donc $c(z)\in\,]z,1[$ tel que :

$$\frac{1-G_X(z)}{1-z}=G'_X\big(c(z)\big)=\sum_{n=1}^{+\infty}np_nc(z)^{n-1}.$$

Cela tend vers $G'_X(1)$ si z croît vers 1 par hypothèse. Soit n_0 un entier. On a alors :

$$G'_X(1)=\lim_{z\to 1}\sum_{n=1}^{+\infty}np_nc(z)^{n-1}\geqslant\lim_{z\to 1}\sum_{n=1}^{n_0}np_nc(z)^{n-1}=\sum_{n=1}^{n_0}np_n\,,$$

car $c(z)$ tend vers 1 quand z tend vers 1. Comme cela est vrai pour tout n_0, toutes les sommes partielles de la série $\sum_n np_n$ sont inférieures ou égales à $G'_X(1)$ ce qui prouve que cette série converge. Donc X est intégrable, et de plus $G'_X(1)\geqslant\mathbb{E}[X]$. Enfin, trivialement :

$$\sum_{n=1}^{+\infty}np_nc(z)^{n-1}\leqslant\sum_{n=1}^{+\infty}np_n\,,$$

ce qui montre finalement que $G'_X(1)\leqslant\mathbb{E}[X]$ et donc $G'_X(1)=\mathbb{E}[X]$. ■

Remarque. Si l'on suppose que le rayon de convergence R est strictement supérieur à 1, la démonstration de la proposition devient simple et immédiate par le théorème de dérivation d'une série entière à l'intérieur de l'intervalle de convergence.

6.2 Cas des vecteurs aléatoires

La notion de fonction génératrice s'étend aux vecteurs aléatoires à coordonnées entières positives. Pour un tel vecteur aléatoire $X=(X_1,\ldots,X_n)$ on pose pour tout $z\in[-1,1]^n$:

$$G_X(z):=\mathbb{E}\left[z_1^{X_1}z_2^{X_2}\cdots z_n^{X_n}\right]=\sum_{x\in\mathbb{N}^n}z_1^{x_1}\cdots z_n^{x_n}p_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n)\,. \tag{38}$$

De manière analogue à (36), la relation :

$$p_X(k_1,\ldots,k_n)=\frac{1}{k_1!\cdots k_n!}\frac{\partial^{k_1+\cdots+k_n}G_X}{\partial z_1^{k_1}\cdots\partial z_n^{k_n}}(0,\ldots,0)\,, \tag{39}$$

montre que G_X détermine la loi du vecteur X. Sur la formule :

$$G_{X_1,\ldots,X_n}(z_1,\ldots,z_n)=\mathbb{E}[z_1^{X_1}\cdots z_n^{X_n}]\,,$$

on voit immédiatement que la fonction génératrice du vecteur X détermine celle de toute coordonnée de X ou plus généralement de tout sous-vecteur de X, puisque l'on peut éliminer du second membre les X_j indésirables simplement en prenant les z_j correspondants égaux à 1. Ainsi par exemple si $k \in [\![1, n]\!]$:

$$G_{X_1,\dots,X_k}(z_1,\dots,z_k) = G_{X_1,\dots,X_n}(z_1,\dots,z_k,1,\dots,1)\,.$$

6.3 Indépendance et convolution

Proposition 29. Les v.a. entières positives $X_1,\dots,X_n$ sont indépendantes si, et seulement si, la fonction génératrice du vecteur aléatoire $(X_1,\dots,X_n)$ se met sous la forme produit :

$$G_{X_1,\dots,X_n}(z_1,\dots,z_n) = G_{X_1}(z_1)\cdots G_{X_n}(z_n)\,, \tag{40}$$

pour tous $z_1,\dots,z_n \in [-1,1]$, où G_{X_i} est la fonction génératrice de la v.a. X_i.
La fonction génératrice de la somme $S_n = X_1 + \cdots + X_n$ de n v.a. entières positives indépendantes $X_1,\dots,X_n$ est égale au produit (ordinaire) des fonctions génératrices des v.a. X_j, soit :

$$G_{S_n}(z) = \prod_{j=1}^{n} G_{X_j}(z)\,, \tag{41}$$

pour tout $z \in [-1,1]$. En particulier $G_{S_n} = (G_X)^n$ si les v.a. X_j suivent la même loi et ont donc la même fonction génératrice G_X.

Démonstration. D'après la définition (38) et la proposition 22 de la page 735 appliquée aux fonctions $f_j(x) = z_j^x$ ($x \in \mathbb{N}$), nous avons sous l'hypothèse d'indépendance des X_j :

$$G_{X_1,\dots,X_n}(z_1,\dots,z_n) = \mathbb{E}[z_1^{X_1}\cdots z_n^{X_n}] = \mathbb{E}[z_1^{X_1}]\cdots\mathbb{E}[z_n^{X_n}] = G_{X_1}(z_1)\cdots G_{X_n}(z_n)\,,$$

ce qui démontre (40).
Réciproquement, si l'on a (40) pour tout n-uplet z, alors en dérivant les deux membres de cette égalité on obtient par (39) et (36) que :

$$p_{(X_1,\dots,X_n)}(k_1,\dots,k_n) = p_{X_1}(k_1)\cdots p_{X_n}(k_n)\,,$$

pour tout n-uplet d'entiers $(k_1,\dots,k_n)$, ce qui est précisément la définition de l'indépendance de v.a.
La formule (41) se démontre de même en écrivant que :

$$G_{S_n}(z) = \mathbb{E}[z^{S_n}] = \mathbb{E}\left[\prod_{j=1}^{n} z^{X_j}\right] = \prod_{j=1}^{n}\mathbb{E}[z^{X_j}]\,,$$

ou alternativement en utilisant (40) :

$$G_{S_n}(z) = G_{X_1,\dots,X_n}(z,\dots,z) = \prod_{j=1}^{n} G_{X_j}(z)\,. \qquad \blacksquare$$

La formule (41) détermine de manière synthétique la loi de la somme $Z = X+Y$ de $n = 2$ variables entières positives X et Y indépendantes. Une méthode alternative consiste à écrire :

$$\mathbb{P}(X+Y=z) = \sum_x \mathbb{P}(X=x, Y=z-x) = \sum_x \mathbb{P}(X=x)\mathbb{P}(Y=z-x)\,,$$

par indépendance. La loi de la somme $Z = X+Y$ est donnée par :

$$p_{X+Y}(z) = \sum_x p_X(x)p_Y(z-x) = \sum_y p_X(z-y)p_Y(y)\,.$$

La valeur commune des deux sommes est appelée la convolée de p_X et p_Y, on la note $p_X * p_Y(z)$ et donc :

$$p_{X+Y} = p_X * p_Y\,,$$

lorsque X et Y sont indépendantes. La relation (41) montre que la fonction génératrice d'une convolée est le produit des fonctions génératrices.

6.4 Fonctions génératrices de lois usuelles

6.4.1 La loi de Bernoulli

Si $0 < p < 1$, alors X suit la loi de Bernoulli de paramètre p si $\mathbb{P}(X=0) = 1-p$ et $\mathbb{P}(X=1) = p$. Sa fonction génératrice est :

$$G_X(z) = (1-p) + pz\,,$$

son espérance est p et sa variance est $(1-p)p$.

6.4.2 La loi binomiale $\mathcal{B}(n,p)$

C'est la loi du nombre S de Pile dans une suite de n tirages Pile ou Face :

$$\mathbb{P}(S=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in [\![0,n]\!]\,.$$

Sa fonction génératrice est, d'après la formule du binôme :

$$G_S(z) = \sum_{k=0}^{n} \binom{n}{k} (zp)^k (1-p)^{n-k} = \big((1-p)+pz\big)^n\,.$$

Donc en prenant sa dérivée en 1, son espérance est np. Quant à sa variance, c'est $n(1-p)p$. On remarque que si X et Y sont indépendantes et de lois respectives $\mathcal{B}(n,p)$ et $\mathcal{B}(m,p)$, alors la fonction génératrice de $X+Y$ est par la proposition 29 $\big((1-p)+pz\big)^{n+m}$ qui est la fonction génératrice de la loi $\mathcal{B}(n+m,p)$. Comme la fonction génératrice caractérise la loi d'une v.a., la loi de $X+Y$ est $\mathcal{B}(n+m,p)$:

PROPRIÉTÉ. *La somme de deux v.a. indépendantes de lois respectives $\mathcal{B}(n,p)$ et $\mathcal{B}(m,p)$ est une v.a. de loi $\mathcal{B}(n+m,p)$.*

Un bon moyen de retenir ces résultats sur la loi binomiale est d'observer que si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes de même loi de Bernoulli de paramètre p, alors $S = X_1 + \cdots + X_n$ est de loi binomiale $\mathcal{B}(n,p)$ comme on le voit par la fonction génératrice G_S.

6.4.3 La loi de Poisson $\mathcal{P}(\lambda)$

Si $\lambda > 0$, on dit que la v.a. Y à valeurs dans $\mathbb{N}$ suit la loi de Poisson $\mathcal{P}(\lambda)$ si, pour tout $k \in \mathbb{N}$:

$$\mathbb{P}(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!}\cdot$$

Sa fonction génératrice est :

$$G_Y(z) = \sum_{k=0}^{+\infty} z^k e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{+\infty}\frac{(\lambda z)^k}{k!} = \exp\bigl(\lambda(z-1)\bigr)\,,$$

son espérance et sa variance sont toutes deux égales à λ. On remarque que si X et Y sont indépendantes et de lois respectives $\mathcal{P}(\lambda)$ et $\mathcal{P}(\theta)$, alors la loi de $X+Y$ est $\mathcal{P}(\lambda+\theta)$, comme on le voit par application de la proposition 29.

PROPRIÉTÉ. *La somme de deux v.a. indépendantes de lois respectives $\mathcal{P}(\lambda)$ et $\mathcal{P}(\theta)$ est une v.a. de loi $\mathcal{P}(\lambda+\theta)$.*

La manière la plus courante de rencontrer une loi de Poisson dans la nature est en tant qu'approximation de la loi binomiale (voir le lemme 15 de la page 729).

6.4.4 Loi géométrique et loi négative binomiale

Considérons une suite infinie de tirages Pile ou Face, c'est-à-dire une suite $(X_i)_{i\in\mathbb{N}^*}$ de v.a. de Bernoulli indépendantes de même paramètre p (Pile$:=\{X=1\}$, Face$:=\{X=0\}$). La loi du temps d'attente T_1 du premier Pile, soit $T_1 := \inf\{n \geqslant 1 \mid X_n = 1\}$, se calcule facilement en remarquant que $T_1 > n$ si, et seulement si, les n premiers essais ont été des Faces, un événement de probabilité $(1-p)^n$. Par conséquent :

$$\mathbb{P}(T_1 = n) = \mathbb{P}(T_1 > n-1) - \mathbb{P}(T_1 > n) = (1-p)^{n-1} - (1-p)^n = (1-p)^{n-1}p\,,$$

ce qui montre que la loi de T_1 est la loi géométrique de paramètre p donnée par :

$$p_{T_1}(n) = p(1-p)^{n-1}, \quad n \in \mathbb{N}^*\,.$$

Sa fonction génératrice est la fonction homographique :

$$G_{T_1}(z) = \sum_{n=1}^{+\infty} z^n p(1-p)^{n-1} = pz\sum_{n=1}^{+\infty}[z(1-p)]^{n-1} = \frac{pz}{1-(1-p)z}\,,$$

son espérance est $\mathbb{E}[T_1] = 1/p$ et sa variance est $\mathrm{Var}(T_1) = (1-p)/p^2$.
Intéressons-nous maintenant au temps d'attente T_k du $k^{\text{ème}}$ Pile. D'après ce que l'on a vu page 740, cette v.a. suit la loi binomiale négative de paramètres (k,p).
Montrons que c'est la somme de k variables aléatoires indépendantes $(\tau_i)_{i\in[\![1,k]\!]}$, de même loi que T_1 : la v.a. τ_k représente l'intervalle de temps entre les $(k-1)^{\text{ème}}$ et $k^{\text{ème}}$ Piles.

Démonstration. Considérons une suite de k variables aléatoires indépendantes $(\tau_i)_{i\in[\![1,k]\!]}$ de loi géométrique de paramètre p, et notons $\widetilde{T}_k$ leur somme. Sa fonction génératrice est :

$$G_{\widetilde{T}_k}(z) = [G_{\tau_1}(z)]^k = \frac{(pz)^k}{(1-(1-p)z)^k}\cdot$$

Toutefois, la loi de $\widetilde{T}_k$ est concentrée sur les entiers supérieurs ou égaux à k, et il y a avantage à considérer plutôt la loi de $\widetilde{T}_k - k$, concentrée sur $\mathbb{N}$, de fonction génératrice :

$$\begin{aligned} G_{\widetilde{T}_k - k}(z) = G_{\widetilde{T}_k}(z) z^{-k} &= \left(\frac{p}{1-(1-p)z} \right)^k \\ &= \sum_{n=0}^{+\infty} k(k+1)\cdots(k+n-1) p^k (1-p)^n z^n \,, \end{aligned}$$

obtenue en développant la série entière $1/(1-x)^k$ en $x = (1-p)z$. Cela entraîne donc que si $n \geqslant k$:

$$\begin{aligned} \mathbb{P}(\widetilde{T}_k = n) = \mathbb{P}(\widetilde{T}_k - k = n - k) &= \frac{1}{(n-k)!} k(k+1)\cdots(n-1) p^k (1-p)^{n-k} \\ &= \binom{n-1}{k-1} p^k (1-p)^{n-k} \,. \end{aligned}$$

En comparant avec la définition 13, on obtient que $\widetilde{T}_k$ suit bien une loi négative binomiale de paramètres (k, p). ■

On a donc montré :

PROPRIÉTÉ. *La somme de k v.a. indépendantes et identiquement distribuées de loi $\mathcal{G}(p)$ est une v.a. de loi $\mathcal{NB}(k, p)$.*

Un corollaire immédiat de ce résultat est qu'une v.a. suivant la loi $\mathcal{NB}(k, p)$ a pour espérance k/p et pour variance $k(1-p)/p^2$.

Loi usuelle $X \sim$	loi de probabilité $P_X(k) =$	espérance $\mathbb{E}[X] =$	variance $\mathrm{Var}(X) =$	fonction génératrice $G_X(z) =$
Bernoulli $\mathcal{B}(p)$	$\begin{cases} 1-p & \text{si } k = 0 \\ p & \text{si } k = 1 \end{cases}$	p	$p(1-p)$	$pz + (1-p)$
Binomiale $\mathcal{B}(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ si $k \in [\![0, n]\!]$	np	$np(1-p)$	$[pz + (1-p)]^n$
Géométrique $\mathcal{G}(p)$	$p(1-p)^{k-1}$ si $k \geqslant 1$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$	$\dfrac{pz}{1-(1-p)z}$
Poisson $\mathcal{P}(\lambda)$	$e^{-\lambda} \dfrac{\lambda^k}{k!}$ si $k \geqslant 0$	λ	λ	$e^{\lambda(z-1)}$

TABLE IV.2.1 – Tableaux des espérances, variances et fonctions génératrices des principales lois discrètes usuelles à valeurs dans $\mathbb{N}$.

6.5 Sommes aléatoires de v.a. indépendantes

On considère dans ce paragraphe une application moins triviale mais révélatrice de la puissance de cet outil qu'est la fonction génératrice. Il arrive fréquemment en Probabilités que l'on ait à considérer des sommes de variables aléatoires dont le nombre de termes est lui-même aléatoire.

Exemples. 1) Service journalier $S := \sum_{i=1}^{\nu} X_i$ effectué par un organe de service lorsque les temps de service X_i d'une part et le nombre ν de clients d'autre part sont aléatoires.
2) Nombre de garçons d'une famille $N := \sum_{i=1}^{\nu} \mathbf{1}_{G_i}$ puisque le sexe de chaque enfant ainsi que le nombre total d'enfants ν sont aléatoires (G_i désigne l'événement « le $i^{ème}$ enfant est un garçon »).

La fonction génératrice d'une telle somme s'obtient très simplement lorsque les v.a. mises en jeu sont toutes indépendantes.

Proposition 30. Si $(X_n)_{n\geqslant 1}$ est une suite infinie de v.a. entières positives, indépendantes et de même loi donnée par la fonction génératrice :

$$\varphi(z) := \mathbb{E}[z^{X_1}]\,,$$

et si ν est une v.a. entière positive, indépendante de la suite $(X_n)_{n\geqslant 1}$ (c'est-à-dire indépendante du vecteur $(X_1, \ldots, X_n)$ quel que soit n), de fonction génératrice :

$$G(z) := \mathbb{E}[z^{\nu}]\,,$$

alors la fonction génératrice de la somme :

$$S := \sum_{i=1}^{\nu} X_i \quad (S := 0 \text{ si } \nu = 0)$$

des ν premières v.a. X_n est la fonction composée de φ et G, soit :

$$\mathbb{E}[z^S] = G \circ \varphi(z)\,.$$

Démonstration. Soient $S_n = \sum_{i=1}^{n} X_i$, $n \geqslant 0$, les sommes partielles des X_i ($S_0 = 0$). Alors la v.a. S s'écrit explicitement sur Ω :

$$S(\omega) = S_{\nu(\omega)}(\omega)\,,$$

et dépend donc de deux manières de ω ; dans les calculs faisant intervenir S_ν, il convient de conditionner par la valeur de ν, puisque $S = S_n$ lorsque $\nu = n$. Ainsi, la fonction génératrice de S :

$$\mathbb{E}[z^S] = \sum_{s=0}^{+\infty} z^s \mathbb{P}(S = s)\,,$$

se calcule par la formule de Bayes :

$$\mathbb{P}(S=s)=\sum_{n=0}^{+\infty}\mathbb{P}(S=s\,|\,\nu=n)\mathbb{P}(\nu=n)=\sum_{n=0}^{+\infty}\mathbb{P}(S_n=s\,|\,\nu=n)\mathbb{P}(\nu=n)\,.$$

Or ν et S_n sont indépendants, donc $\mathbb{P}(S_n=s\,|\,\nu=n)=\mathbb{P}(S_n=s)$ ce qui implique que :

$$\mathbb{E}[z^S]=\sum_{n=0}^{+\infty}\mathbb{P}(\nu=n)\sum_{s=0}^{+\infty}z^s\mathbb{P}(S_n=s)=\sum_{n=0}^{+\infty}\mathbb{P}(\nu=n)\mathbb{E}[z^{S_n}]\,.$$

En appliquant la proposition 29 de la page 744 on obtient que $\mathbb{E}[z^{S_n}]=\varphi(z)^n$, donc :

$$\mathbb{E}[z^S]=\sum_{n=0}^{+\infty}\mathbb{P}(\nu=n)\varphi(z)^n\,,$$

qui n'est autre que $G\big(\varphi(z)\big)$. ■

Exercice 4.

Si ν est une v.a. de Poisson de paramètre θ, si les X_i sont des v.a. de Bernoulli de paramètre p, et si toutes ces v.a. sont indépendantes, déterminer la loi de $S:=\sum_{n=1}^{\nu}X_n$.

Solution. On calcule les fonctions génératrices de ν et des X_i :

$$E[z^\nu]=\exp[\theta(z-1)]\ \text{ et }\ \mathbb{E}[z^X]=zp+(1-p)\,.$$

On évalue la composée de ces deux fonctions, ce qui donne la fonction génératrice de S :

$$\mathbb{E}[z^S]=\exp(\theta[zp+(1-p)-1])=\exp[\theta p(z-1)]\,.$$

On reconnaît la fonction génératrice d'une loi de Poisson de paramètre θp, ce qui montre que S suit une loi de Poisson de paramètre θp.

Exercice type corrigé

IV.2.0 Un élément chimique émet des électrons pendant une période T. Le nombre d'électrons émis est une v.a. Y qui suit une loi de Poisson de paramètre λ. Chaque électron a une probabilité p d'avoir un effet biologique (on dira qu'il est efficace). Soit Z la v.a. égale au nombre d'électrons efficaces émis pendant une période T.

1. Déterminer la loi du couple (Y, Z).
2. Déterminer la loi de Z. Calculer son espérance.
3. Les v.a. Y et Z sont-elles indépendantes ?
4. On considère $X :=$ le nombre d'électrons émis non efficaces. Déterminer la loi du couple (X, Z). Les v.a. X et Z sont-elles indépendantes ?

On rappelle que, pour montrer que les deux v.a. entières Y et Z ne sont pas indépendantes, il suffit de trouver deux entiers y et z tels que $\mathbb{P}(Y = y, Z = z) \neq \mathbb{P}(Y = y)\mathbb{P}(Z = z)$.
Pour montrer que les deux v.a. entières X et Z sont indépendantes, il faut montrer que pour tous entiers x et z, on a $\mathbb{P}(X = x, Z = z) = \mathbb{P}(X = x)\mathbb{P}(Z = z)$.

Solution.

1. Les v.a. Y et Z sont à valeurs dans $\mathbb{N}$. Soient $y, z \in \mathbb{N}$. Si $y < z$, alors on a clairement :
$$\mathbb{P}(Y = y, Z = z) = 0\,.$$
Si $y \geqslant z$, alors :
$$\mathbb{P}(Y = y, Z = z) = \mathbb{P}(Z = z | Y = y)\mathbb{P}(Y = y)\,.$$
Or, si on sait que $Y = y$, alors Z suit une loi binomiale de paramètres (y, p), donc :
$$\mathbb{P}(Y = y, Z = z) = \binom{y}{z} p^z (1-p)^{y-z} \times e^{-\lambda}\frac{\lambda^y}{y!} = e^{-\lambda}\frac{p^z(1-p)^{y-z}\lambda^y}{z!(y-z)!}\,.$$
2. On utilise la formule des marginales. Si $z \in \mathbb{N}$, alors :
$$\begin{aligned}
\mathbb{P}(Z = z) &= \sum_{y=0}^{+\infty} \mathbb{P}(Y = y, Z = z) = \sum_{y=z}^{+\infty} e^{-\lambda}\frac{p^z(1-p)^{y-z}\lambda^y}{z!(y-z)!} \\
&= e^{-\lambda}\frac{p^z}{z!}\sum_{y=z}^{+\infty}\frac{(1-p)^{y-z}\lambda^y}{(y-z)!} = e^{-\lambda}\frac{p^z\lambda^z}{z!}\sum_{n=0}^{+\infty}\frac{(1-p)^n\lambda^n}{n!} \\
&= e^{-\lambda}\frac{p^z\lambda^z}{z!}e^{(1-p)\lambda} \\
&= e^{-p\lambda}\frac{(p\lambda)^z}{z!}\,.
\end{aligned}$$
On reconnaît la loi de Poisson de paramètre $p\lambda$. L'espérance de Z est donc $p\lambda$.
3. Considérons d'une part :
$$\mathbb{P}(Y = 0, Z = 1) = 0\,,$$
et d'autre part :
$$\mathbb{P}(Y = 0)\mathbb{P}(Z = 1) = e^{-p} \times e^{-p\lambda}\,.$$
Les deux expressions ne sont pas égales, donc les v.a. Y et Z ne sont pas indépendantes.

4. Les v.a. X et Z sont à valeurs dans $\mathbb{N}$. Soient $x, z \in \mathbb{N}$. Comme :

$$\{X = x, Z = z\} = \{Y = z + x, Z = z\},$$

on peut écrire :

$$\mathbb{P}(X = x, Z = z) = \mathbb{P}(Y = z + x, Z = z) = e^{-\lambda} \frac{p^z (1-p)^x \lambda^{x+z}}{z!x!}.$$

Par le même raisonnement qu'au 2., on trouve que X suit la loi de Poisson de paramètre $(1-p)\lambda$.
Pour tous $x, z \in \mathbb{N}$, on a donc :

$$\begin{aligned} \mathbb{P}(X = x, Z = z) &= e^{-\lambda} \frac{(p\lambda)^z [(1-p)\lambda]^x}{z!x!} = e^{-p\lambda} \frac{(p\lambda)^z}{z!} \times e^{-(1-p)\lambda} \frac{[(1-p)\lambda]^x}{x!} \\ &= \mathbb{P}(Z = z)\mathbb{P}(X = x), \end{aligned}$$

ce qui démontre l'indépendance de X et Z.

Pour aller plus loin

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

IV.2.1 Un concierge a n clefs dont une seule ouvre sa porte, mais il ne sait plus laquelle.

1. Il les essaie l'une après l'autre en éliminant après chaque essai la clef qui n'a pas convenu. Trouver le nombre moyen d'essais nécessaires pour trouver la bonne clef.
2. Le concierge rentre chez lui après une sortie bien arrosée... Après chaque essai, il remet la clef essayée dans le trousseau. Trouver le nombre moyen d'essais nécessaires pour trouver la bonne clef.

IV.2.2 Soit X une v.a. de carré intégrable. Montrer que :

$$\mathrm{Var}(X) = \inf_{c \in \mathbb{R}} \mathbb{E}\left[(X - c)^2\right],$$

et que cette borne inférieure est atteinte précisément pour $c = \mathbb{E}[X]$.

IV.2.3 Soit X une v.a. suivant une loi de Poisson de paramètre $\lambda > 0$. On considère les deux événements $E :=$« X est paire » et $F :=$« X est impaire ». Montrer que :

$$\mathbb{P}(E) > \mathbb{P}(F).$$

IV.2.4 Le nombre de blessés à l'œil par un bouchon de champagne arrivant aux urgences ophtalmologiques la nuit de la Saint-Sylvestre suit une loi de Poisson d'espérance 6. Sachant que les victimes potentielles sont réparties équitablement entre les deux sexes, on note X le nombre de femmes qui figurent parmi ces éborgnés du Jour de l'An et Y le nombre d'hommes.

1. Établir les lois de probabilité de X et Y. Préciser leurs espérances et leurs variances.
2. Montrer que X et Y sont indépendantes.
3. Calculer la probabilité pour qu'il y ait autant d'hommes que de femmes admis aux urgences.

IV.2.5 Calculer les sommes suivantes en précisant pour quelle valeurs de x elles sont valables :

$$f(x) = \sum_{k=0}^{+\infty} x^k\,, \quad g(x) = \sum_{k=0}^{+\infty} \frac{x^k}{k!}\,.$$

En déduire, pour tout $p \in]0,1[$ et tout $\lambda > 0$, les valeurs des séries suivantes :

$$\sum_{k=1}^{+\infty} kp^k\,, \quad \sum_{k=1}^{+\infty} k^2p^k\,, \quad \sum_{k=0}^{+\infty} s^kp^k\,, \quad \sum_{k=1}^{+\infty} k\frac{\lambda^k}{k!}\,, \quad \sum_{k=1}^{+\infty} k^2\frac{\lambda^k}{k!}\,, \quad \sum_{k=0}^{+\infty} s^k\frac{\lambda^k}{k!}\,.$$

IV.2.6 Un commerçant estime que la demande d'un certain produit saisonnier est une v.a. discrète X de loi :

$$\mathbb{P}(X = k) := \frac{p^k}{(1+p)^{k+1}} \text{ pour tout } k \in \mathbb{N}\,,$$

où p est le prix d'une campagne publicitaire de l'année précédente.

1. Vérifier que X suit bien une loi de probabilité.
2. Calculer l'espérance et la variance de X.
3. Connaissant son stock s, calculer la probabilité de rupture de stock.

IV.2.7 Cet exercice est classique, il possède de nombreuses variantes et on l'appelle le problème des rencontres.
On considère une urne contenant n boules numérotées de 1 à n. On les extrait successivement sans remise et l'on observe le numéro de la boule tirée. On dit qu'il y a rencontre au $k^{ème}$ tirage si la boule k est tirée au $k^{ème}$ tirage. Calculer la probabilité pour qu'il y ait au moins une rencontre.

IV.2.8 Cet exercice s'appelle le problème des boîtes d'allumettes de Banach.
Un fumeur a dans chacune de ses poches droite et gauche une boîte d'allumettes (de contenance N chacune). Lorsqu'il désire une allumette il doit choisir au hasard une de ses poches (chacune avec probabilité $1/2$). On considère le moment où, pour la première fois, le fumeur s'aperçoit que l'une des boîtes d'allumettes est vide. À ce moment, l'autre boîte peut contenir un nombre quelconque d'allumettes. On désigne par u_r la probabilité qu'elle en contienne r.

1. Calculer u_r, $0 \leqslant r \leqslant N$. Que vaut $\sum_{r=0}^{N} u_r$?
2. Calculer la probabilité u que les deux boîtes soient vides lorsque le fumeur s'aperçoit que l'une des boîtes d'allumettes est vide. Déterminer une expression approchée de u lorsque N est grand en utilisant la formule de Stirling $n! \sim (n/e)^n\sqrt{2\pi n}$.

IV.2.9

1. Soient X et Y deux v.a. à valeurs dans $\mathbb{N}^*$, indépendantes et de même loi géométrique de paramètre p donnée par :

$$\forall n \in \mathbb{N}^*\,, \quad \mathbb{P}(X = n) = p(1-p)^{n-1} \quad \text{où } p \in]0,1[\,.$$

On note V la v.a. égale au minimum de X et de Y et W celle égale à la différence $X - Y$.

(a) Déterminer la loi du couple (V, W).

(b) En déduire les lois de V et W.

(c) Montrer que V et W sont indépendantes.

2. Réciproquement, soient X et Y deux v.a. à valeurs dans $\mathbb{N}^*$, indépendantes et de même loi satisfaisant à :
$$\forall n \in \mathbb{N}^*, \quad \mathbb{P}(X = n) \neq 0.$$
Les v.a. V et W sont définies comme précédemment. Montrer que si V et W sont indépendantes, alors X et Y suivent une loi géométrique sur $\mathbb{N}^*$ dont on précisera le paramètre.

IV.2.10 Soient X_1 et X_2 deux v.a. géométriques de paramètres p_1 et p_2 :
$$\mathbb{P}(X_j = n) = p_j(1-p_j)^{n-1}, \quad n \in \mathbb{N}^*.$$

1. Déterminer la loi de $X_m := \min(X_1, X_2)$.
2. Calculer l'espérance de X_m.

IV.2.11 Soit $a > 0$. On considère les coefficients réels $p_{i,j} := \lambda \dfrac{a^{i+j}}{i!j!}$, avec $i \in \mathbb{N}, j \in \mathbb{N}$.

1. Donner la valeur de λ pour laquelle ces coefficients forment la loi d'un couple de v.a.
2. Soit (X, Y) un couple de v.a. à valeurs dans $\mathbb{N}$, admettant $(p_{i,j})_{i\in\mathbb{N}, j\in\mathbb{N}}$ pour loi jointe.
   (a) Déterminer les lois marginales de X et Y.
   (b) Les v.a. X et Y sont elles indépendantes ?
3. Déterminer la loi de la somme $X + Y$.

IV.2.12 Si ν, X_1, X_2, ... sont des v.a. strictement positives indépendantes, si ν suit une loi géométrique d'espérance a et si les X_i suivent la loi géométrique d'espérance b, montrer que $S := \sum_{i=1}^{\nu} X_i$ suit une loi géométrique d'espérance ab.

IV.2.13 Déduire de la proposition 30 de la page 748 que :
$$\mathbb{E}[S] = \mathbb{E}[\nu]\mathbb{E}[X], \quad \mathrm{Var}(S) = \mathrm{Var}(\nu)\,\mathbb{E}[X]^2 + \mathbb{E}[\nu]\mathrm{Var}(X).$$
Comment s'écrivent ces formules lorsque ν est une constante déterministe ?

Variables aléatoires à densité

IV.3

1 Variables aléatoires réelles

1.1 Définition

Une v.a. est une fonction $X : \omega \mapsto X(\omega)$ définie sur un espace fondamental Ω. Dans le cas de v.a. discrètes, la définition réclame aussi que l'événement « X prend une valeur donnée » est un événement de la tribu $\mathcal{A}$ associée à l'espace fondamental Ω. Dans le cas de v.a. pouvant prendre un nombre infini non dénombrable de valeurs, cette hypothèse doit être modifiée. La plupart des v.a. que nous considérons dans ce livre sont à valeurs réelles. Il est alors naturel de réclamer que la définition d'une v.a. X à valeurs réelles permette de considérer des événements du type $\{X \in I\}$ pour tout intervalle I de $\mathbb{R}$.

Définition 1. Une *variable aléatoire (v.a.) réelle* sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ est une application $X : \Omega \to \mathbb{R}$ telle que :

$$\{X \in I\} \in \mathcal{A} \text{ pour tout intervalle } I \text{ de } \mathbb{R}\,. \tag{1}$$

Proposition 1. *(i)* La condition (1) est vérifiée dès que :

$$\{X \leqslant x\} \in \mathcal{A} \text{ pour tout } x \in \mathbb{R}\,. \tag{2}$$

(ii) La condition (1) implique que :

$$\{X \in B\} \in \mathcal{A} \text{ pour tout } B \in \mathcal{B}(\mathbb{R})\,. \tag{3}$$

Rappelons que $\mathcal{B}(\mathbb{R})$ est la tribu borélienne de $\mathbb{R}$, c'est-à-dire la plus petite tribu contenant les intervalles de $\mathbb{R}$. En pratique, pour vérifier que X est une v.a. réelle, on se contente de vérifier la condition (2). Et une fois qu'on l'a montrée, on peut considérer et calculer les probabilités de tous les événements du type $\{X \in B\}$ pour tout $B \in \mathcal{B}(\mathbb{R})$.

Démonstration. Il faut d'abord remarquer que $\bigcup_n \{X \in A_n\} = \{X \in \bigcup_n A_n\}$ pour toute suite de boréliens (A_n) et $\{X \in A\}^c = \{X \in A^c\}$ pour tout $A \in \mathcal{B}(\mathbb{R})$.

Supposons que l'on ait (2) pour tout $x \in \mathbb{R}$. Soient deux réels $a < b$.

- Comme $]a,b] =]-\infty,b] \cap (]-\infty,a])^c$, on a $\{X \in]a,b]\} = \{X \leqslant b\} \cap \{X \leqslant a\}^c$. Or $\mathcal{A}$ est stable par intersection et par passage au complémentaire, donc $\{X \in]a,b]\} \in \mathcal{A}$.
- Comme $]a,b[= \bigcup_{n\geqslant 1}]a, b-1/n]$, on a $\{X \in]a,b[\} = \bigcup_{n\geqslant 1} \{X \in]a, b-1/n]\}$. Or $\mathcal{A}$ est stable par union dénombrable, donc $\{X \in]a,b[\} \in \mathcal{A}$.
- Comme $[a,b] = \bigcap_{n\geqslant 1}]a-1/n, b]$, on a $\{X \in [a,b]\} = \bigcap_{n\geqslant 1} \{X \in]a-1/n, b]\}$. Or $\mathcal{A}$ est stable par intersection dénombrable, donc $\{X \in [a,b]\} \in \mathcal{A}$.

Comme $[a,b[= \bigcap_{n\geqslant 1} \bigcup_{m\geqslant 1}]a-1/n, b-1/m]$, on a $\{X \in [a,b[\} \in \mathcal{A}$.

Ainsi, pour tout type d'intervalle borné I, on a $\{X \in I\} \in \mathcal{A}$. Dans le cas d'intervalles semi-infinis, par exemple $[a,+\infty[$, on part de $[a,+\infty[= \bigcup_{n\geqslant 1} [a,n[$ pour obtenir $\{X \in [a,b[\} \in \mathcal{A}$. Tout cela démontre (1).

La démonstration du *(ii)* est similaire. Si (1) est vraie, alors la stabilité de la tribu $\mathcal{A}$ par complémentation et union dénombrable montre que les parties $\{X \in B\}$ pour tout $B \in \mathcal{B}(\mathbb{R})$ sont aussi dans $\mathcal{A}$. ∎

Exemples. 1) Une v.a. discrète X au sens de la définition de la page 707 à valeurs dans un sous-ensemble fini ou dénombrable E de $\mathbb{R}$ est une v.a. réelle au sens de la définition 1 ci-dessus. En effet, on a dans ce cas $\{X \leqslant x\} = \bigcup_{y\in E | y\leqslant x} \{X = y\}$. Chacun des ensembles $\{X = y\}$ est un événement de la tribu $\mathcal{A}$, donc, par stabilité de $\mathcal{A}$ par union dénombrable, $\{X \leqslant x\} \in \mathcal{A}$.

2) Si $\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}([0,1])$ et $\mathbb{P} = \lambda$ est la mesure de Lebesgue sur $[0,1]$, alors $X(\omega) = \omega$ est une v.a. réelle. En effet, $\{X \leqslant x\}$ est simplement l'intervalle $[0,x]$.

1.2 Loi et fonction de répartition

Si X est une v.a. réelle, alors la formule $Q(B) = \mathbb{P}(X \in B)$ définit une fonction d'ensembles de $\mathcal{B}(\mathbb{R})$ dans $[0,1]$. Cette fonction est une probabilité.

Lemme 2. La fonction $Q : \mathcal{B}(\mathbb{R}) \to [0,1]$ donnée par $Q(B) := \mathbb{P}(X \in B)$ est une probabilité sur $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$.

Démonstration. On vérifie que les conditions requises dans la définition d'une probabilité sont vérifiées par Q. En effet, $Q(\mathbb{R}) = \mathbb{P}(X \in \mathbb{R}) = 1$ et, si $(B_i)_{i\geqslant 1}$ est une suite de boréliens deux à deux disjoints, alors :

$$Q\Big(\bigcup_{i\geqslant 1} B_i\Big) = \mathbb{P}\Big(X \in \bigcup_{i\geqslant 1} B_i\Big) = \mathbb{P}\Big(\bigcup_{i\geqslant 1} \{X \in B_i\}\Big) = \sum_{i\geqslant 1} \mathbb{P}\left(X \in B_i\right) = \sum_{i\geqslant 1} Q(B_i)\,.$$

∎

La probabilité Q est *la loi de probabilité* de la v.a. X.

Définition 2. Soit $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$ une v.a. réelle. La *loi* de cette v.a. est la probabilité Q sur $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ définie par :

$$Q(B) := \mathbb{P}(X \in B), \quad B \in \mathcal{B}(\mathbb{R}).$$

On a vu page 708 que, dans le cas d'une v.a. discrète $X : \Omega \to E$, la propriété de σ-additivité permet d'affirmer que la loi de X est caractérisée par la donnée des probabilités des singletons $\{x\}$, $x \in E$. Cette propriété est fausse dans le cas où l'espace E n'est pas dénombrable, car certains événements ne peuvent pas être décrits comme des unions dénombrables de singletons ou de complémentaires de singletons. On va voir dans la suite que, dans le cas d'une v.a. réelle générale $X : \Omega \to \mathbb{R}$, la loi de X est caractérisée par les valeurs $Q(I) = \mathbb{P}(X \in I)$ pour les intervalles I semi-infinis de la forme $]-\infty, x]$, si bien que l'on va aboutir à la conclusion que la fonction $F_X : \mathbb{R} \to [0,1]$ définie par :

$$F_X(x) = \mathbb{P}(X \leqslant x), \tag{4}$$

caractérise la loi de X. Cette fonction F_X est appelée fonction de répartition de X.

Définition 3. Soit F une fonction de $\mathbb{R}$ dans $\mathbb{R}$. On dit que F est une *fonction de répartition* si elle possède les propriétés suivantes :
(i) $F : \mathbb{R} \to [0,1]$ est croissante,
(ii) $\lim\limits_{x \to -\infty} F(x) = 0$, $\lim\limits_{x \to +\infty} F(x) = 1$,
(iii) F est continue à droite en tout point.

Proposition 3. *(i)* Soit Q une probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Soit F_Q la fonction de $\mathbb{R}$ dans $[0,1]$ définie par :

$$F_Q(x) := Q\left(]-\infty, x]\right).$$

Alors F_Q est une fonction de répartition.
(ii) Si F est une fonction de répartition, alors il existe une unique probabilité Q sur $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ telle que $F_Q = F$.

Démonstration. Montrons l'implication directe, à savoir que, si Q est une probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, alors $F_Q(x) = Q\left(]-\infty, x]\right)$ est une fonction de répartition. Pour cela, il faut que F_Q vérifie les trois propriétés *(i)*, *(ii)* et *(iii)* de la définition 3 ci-dessus.

Si $x \leqslant y$, alors $\{X \leqslant x\} \subset \{X \leqslant y\}$ et la propriété *(i)* résulte de la propriété de monotonie (4) de la proposition 1 de la page 669.

Les événements $A_n = \{X \leqslant -n\}$ constituent une suite décroissante avec $\bigcap\limits_{n \in \mathbb{N}} A_n = \varnothing$, si bien que $F_Q(n) = \mathbb{P}(A_n) \searrow \mathbb{P}(\varnothing) = 0$ lorsque n tend vers $+\infty$. Cela, avec la propriété *(i)*, montre la première partie de la propriété *(ii)*. Les événements $B_n = \{X \leqslant n\}$ constituent une suite croissante avec $\bigcup\limits_{n \in \mathbb{N}} B_n = \Omega$, si bien que $F_Q(n) = \mathbb{P}(B_n) \nearrow \mathbb{P}(\Omega) = 1$

lorsque n tend vers $+\infty$. Cela, avec la propriété *(i)*, montre la deuxième partie de la propriété *(ii)*.
Enfin, pour $x \in \mathbb{R}$ et $h > 0$, $F_Q(x+h) - F_Q(x) = \mathbb{P}(x < X \leqslant x+h)$. Si $h_n \searrow 0$, alors $C_n =]x, x+h_n]$ est une suite décroissante et $\bigcap_n C_n = \varnothing$, donc :

$$F_Q(x+h_n) - F_Q(x) = \mathbb{P}(C_n) \overset{n\to+\infty}{\searrow} \mathbb{P}(\varnothing) = 0,$$

ce qui montre que F_Q est continue à droite au point $x \in \mathbb{R}$ et vérifie donc la propriété *(iii)*.
Nous admettrons la réciproque, dont la démonstration nécessite des résultats de la théorie de l'intégration qui dépassent le cadre de ce cours. ■

Corollaire 4. Soit $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$ une v.a. réelle. La loi de cette v.a. est caractérisée par sa fonction de répartition $F_X : \mathbb{R} \to [0,1]$ définie par (4).

Nous étudions dans ce livre deux types de v.a. réelles :

- les v.a. discrètes, qui ont été étudiées dans le module précédent. Dans ce cas, la fonction de répartition est constante par morceaux, comme le montre la proposition suivante.
- les v.a. à densité, que l'on va étudier dans la prochaine section.

Proposition 5. Soient $E := \{x_1, x_2, \ldots\} \subset \mathbb{R}$ un ensemble discret et P une probabilité sur $(E, \mathcal{P}(E))$ définie par $P(\{x_i\}) = p_i$, $i \in \mathbb{N}^*$. La fonction de répartition de la probabilité P est :

$$F_P(x) = \sum_{i|x_i \leqslant x} p_i \,.$$

Démonstration. Soit $x \in \mathbb{R}$. En utilisant la σ-additivité de P, on obtient :

$$F_P(x) = P\left(]-\infty, x]\right) = P\left(\bigcup_{i|x_i \leqslant x} \{x_i\}\right) = \sum_{i|x_i \leqslant x} P(\{x_i\}) = \sum_{i|x_i \leqslant x} p_i \,.$$ ■

2 Variables à densité

2.1 Densité de probabilité

Définition 4.
(i) Une *densité de probabilité* est une fonction intégrable $p : \mathbb{R} \to \mathbb{R}_+$ continue par morceaux possédant au plus un nombre fini de points de discontinuité et d'intégrale $\displaystyle\int_{\mathbb{R}} p(x)\,\mathrm{d}x = 1$.
(ii) On dit qu'une v.a. réelle X est *à densité* p si sa fonction de répartition F vérifie pour tout $x \in \mathbb{R}$:

$$F(x) = \int_{-\infty}^{x} p(y)\,\mathrm{d}y \,. \tag{5}$$

La densité caractérise la loi d'une v.a. à densité, car elle permet de reconstruire la fonction de répartition qui caractérise la loi d'après la proposition 3 de la page 757.
Si p est une densité de probabilité, alors la fonction F définie par (5) est une fonction de répartition car elle vérifie toutes les conditions requises dans la définition 3 de la page 757, et cette fonction de répartition définit elle même une loi de probabilité sur $\mathbb{R}$.
L'équation (5) montre que la fonction de répartition d'une v.a. à densité est continue. La loi d'une v.a. réelle à densité X est dite diffuse, au sens où les probabilités des événements élémentaires sont nulles :

$$\forall x \in \mathbb{R},\ \ \mathbb{P}(X = x) = 0\,,$$

au contraire des lois discrètes. Cette propriété se montre en considérant la suite croissante $C_n =]-\infty, x - 1/n]$ dont la limite est $C =]-\infty, x[$:

$$\begin{aligned}\mathbb{P}(X = x) &= \mathbb{P}(X \leqslant x) - \mathbb{P}(X < x)\\ &= \mathbb{P}(X \leqslant x) - \lim_{n\to+\infty} \mathbb{P}\Big(X \leqslant x - \frac{1}{n}\Big)\\ &= \int_{-\infty}^{x} p(y)\,\mathrm{d}y - \lim_{n\to+\infty} \int_{-\infty}^{x-\frac{1}{n}} p(y)\,\mathrm{d}y\\ &= \int_{-\infty}^{x} p(y)\,\mathrm{d}y - \int_{-\infty}^{x} p(y)\,\mathrm{d}y = 0\,.\end{aligned}$$

On voit donc clairement ici que l'on ne peut pas caractériser la loi d'une v.a. réelle à densité par les probabilités des singletons $\{x\}$, $x \in \mathbb{R}$, comme on le faisait dans le cas discret.
Si X est une v.a. à densité p, alors on a :

$$\mathbb{P}(X \in I) = \int_I p(x)\,\mathrm{d}x\,, \tag{6}$$

pour tout intervalle I de $\mathbb{R}$. En effet, si I est un intervalle de la forme $]a, b]$, avec $a < b$, alors :

$$\mathbb{P}\left(X \in]a, b]\right) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a) = F(b) - F(a) = \int_a^b p(x)\,\mathrm{d}x\,.$$

Si I est un intervalle d'un autre type (fermé ou ouvert), alors la formule (6) reste valable car la loi de X est diffuse.

2.2 Lois usuelles

2.2.1 Loi uniforme

Pour tout couple de réels $a < b$ la fonction :

$$p(x) := \frac{1}{b-a}\mathbf{1}_{]a,b[}(x) \tag{7}$$

est la densité de la loi uniforme sur $[a, b]$ (voir Figure IV.3.1 de la page suivante). Une v.a. réelle de loi uniforme sur $[a, b]$ est appelée v.a. uniforme. La fonction de répartition de la loi uniforme sur $[a, b]$ est :

$$F(x) = \begin{cases} 0 & \text{si } x \leqslant a\\ \frac{x-a}{b-a} & \text{si } a < x < b\\ 1 & \text{si } x \geqslant b\end{cases} \tag{8}$$

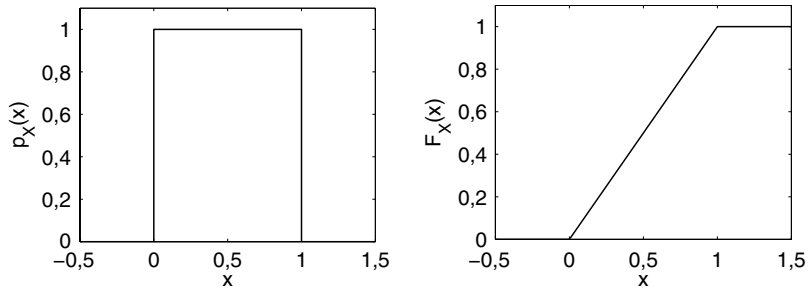


FIGURE IV.3.1 – Densité de probabilité et fonction de répartition d'une v.a. uniforme sur $[0,1]$.

On a déjà rencontré des lois uniformes dans le cas d'espaces de probabilité finis (voir le livre [L1]). On a montré alors qu'il existait une unique probabilité (dite probabilité uniforme) sur un espace fini qui satisfait la propriété d'équiprobabilité, qui demande que les probabilités des événements élémentaires soient toutes égales. En d'autres mots, si X est une v.a. discrète à valeurs dans un espace fini E, alors X suit la loi uniforme si, et seulement si, sa loi de probabilité satisfait la propriété d'équiprobabilité sur $(E, \mathcal{P}(E))$. Cette notion n'a de sens que dans le cas d'espaces discrets, car, comme on vient de le noter, dans le cas des v.a. réelles, toute loi à densité est diffuse, et les probabilités des événements élémentaires sont alors égales car toutes nulles. Ce n'est donc pas une caractérisation satisfaisante. La définition d'équiprobabilité est alors remplacée par une définition qui a du sens dans le cas d'un intervalle. On dit que la loi d'une v.a. réelle satisfait la propriété d'équiprobabilité sur un intervalle $I = [a,b]$ si $\mathbb{P}(X \in I) = 1$ et si les probabilités des sous-intervalles $J \subset I$ ne dépendent que des longueurs des sous-intervalles. On peut vérifier que la loi uniforme sur $I = [a,b]$ dont la densité est (7) satisfait la propriété d'équiprobabilité sur I. En effet, si $J \subset I$, alors J est de la forme $J = [c,d]$ (ou $[c,d[$, ou $]c,d]$, ou $]c,d[$), avec $a \leqslant c \leqslant d \leqslant b$, et donc :

$$\mathbb{P}(X \in J) = F(d) - F(c) = \frac{d-c}{b-a}\cdot$$

Réciproquement, on peut montrer que les lois uniformes définies par (7) sont les seules lois de probabilité qui satisfont la propriété d'équiprobabilité. En effet, si X satisfait la propriété d'équiprobabilité sur $[a,b]$, alors il existe une constante C telle que, pour tout $x \in [a,b[$, pour tout $h > 0$ tel que $x + h \leqslant b$, on a $\mathbb{P}(X \in]x, x+h]) = F(x+h) - F(x) = Ch$. De même, pour tout $x \in]a,b]$, pour tout $h > 0$ tel que $x - h \geqslant a$, on a $\mathbb{P}(X \in]x-h, x]) = F(x) - F(x-h) = Ch$. Cela montre que la fonction de répartition F de X est dérivable sur $[a,b]$ et sa dérivée est constante $F'(x) = C$. Donc F est affine sur $[a,b]$. De plus,

$$F(b) = \mathbb{P}(X \leqslant b) = \mathbb{P}(X < a) + \mathbb{P}(X \in [a,b]) = 0 + 1 = 1\,,$$

$$F(a) = \mathbb{P}(X \leqslant a) = \mathbb{P}(X < a) + \mathbb{P}(X \in \{a\}) = 0 + 0 = 0\,.$$

On peut alors identifier la fonction de répartition qui est nécessairement (8). Comme la

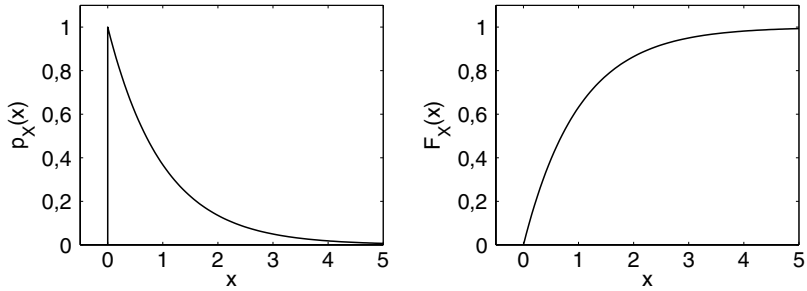


FIGURE IV.3.2 – Densité de probabilité et fonction de répartition d'une v.a. exponentielle de paramètre 1.

fonction de répartition caractérise la loi d'une v.a. réelle, on a ainsi montré que la propriété d'équiprobabilité était satisfaite uniquement par les lois uniformes.

2.2.2 Loi exponentielle

Pour tout réel $\lambda > 0$, la fonction :

$$p(x) := \lambda e^{-\lambda x} \mathbf{1}_{]0,+\infty[}(x) \tag{9}$$

est une densité de probabilité concentrée sur $]0, +\infty[$, la densité exponentielle de paramètre λ (voir Figure IV.3.2). Elle intervient en pratique pour décrire le délai séparant des événements imprévisibles (émission de particules radioactives, passages de métros à une station donnée, tremblements de terre, appels téléphoniques,...). La fonction de répartition d'une v.a. exponentielle de paramètre λ est :

$$F(x) = \begin{cases} 0 & \text{si } x \leqslant 0\,, \\ 1 - e^{-\lambda x} & \text{si } x > 0\,. \end{cases} \tag{10}$$

2.2.3 Loi gaussienne

La fonction :

$$g(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{11}$$

est une densité de probabilité sur $\mathbb{R}$, appelée densité gaussienne (voir Figure IV.3.3 de la page suivante). En effet, g est une fonction continue à valeurs positives, et $c := \displaystyle\int_{\mathbb{R}} g(x)\,\mathrm{d}x$ se calcule en écrivant le carré de cette intégrale comme une intégrale double :

$$\begin{aligned} c^2 &= \int_{\mathbb{R}^2} g(x)g(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbb{R}^2} \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right) \mathrm{d}x\,\mathrm{d}y \qquad \text{par Fubini} \\ &= \int_0^{+\infty} \int_0^{2\pi} \frac{1}{2\pi} \exp\left(-\frac{r^2}{2}\right) r\,\mathrm{d}r\,\mathrm{d}\theta \qquad \text{par passage en coordonnées polaires} \\ &= \frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\theta \int_0^{+\infty} r e^{-r^2/2}\,\mathrm{d}r = 1\,, \end{aligned}$$

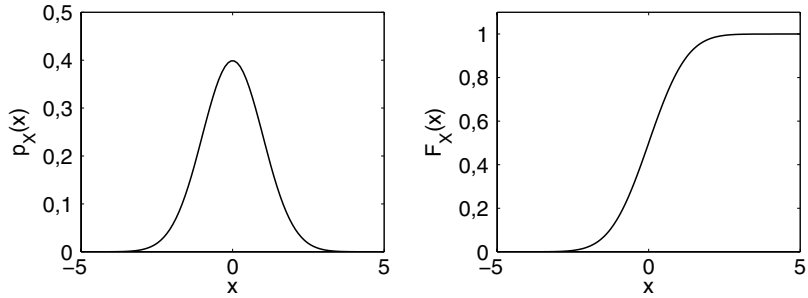


FIGURE IV.3.3 – Densité de probabilité et fonction de réparation d'une v.a. gaussienne.

et donc $c = 1$. Cette densité joue un rôle central en probabilité, car elle intervient pour décrire des lois limites de sommes de variables aléatoires, comme on va le voir dans le module IV.4. Sa fonction de répartition F ne peut pas s'exprimer en termes de fonctions transcendantes usuelles. Cependant, cette fonction de répartition est suffisamment importante pour qu'on lui ait donné un nom. Plus exactement :

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t = \begin{cases} \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{-x}{\sqrt{2}}\right) & \text{si } x \leqslant 0\,, \\ \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) & \text{si } x > 0\,, \end{cases}$$

où erf (error function) est la fonction tabulée de $[0, +\infty[$ dans $[0, 1[$ définie par :

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\, \mathrm{d}t\,.$$

Les programmes de calcul scientifique, matlab, maple, mathematica, python, ... connaissent la fonction erf.

2.3 Caractérisation des v.a. réelles à densité

La proposition qui suit permet en pratique d'identifier des v.a. à densité.

Proposition 6. Soit X une v.a. de fonction de répartition F. Si la dérivée de F existe partout sauf en un nombre fini de points, et si la fonction p définie par :

$$p(x) := \begin{cases} F'(x) & \text{si } F \text{ est dérivable en } x \\ 0 & \text{sinon} \end{cases}$$

est continue par morceaux et vérifie (5) pour tout $x \in \mathbb{R}$, alors p est une densité de probabilité et X est une v.a. à densité p.

La proposition est triviale à la vue de la définition 4 de la page 758, la seule chose à vérifier est que la fonction p en question satisfait les hypothèses de la définition 4 partie i. Or $\displaystyle\int_{\mathbb{R}} p(x)\, \mathrm{d}x = \lim_{x \to +\infty} F(x) = 1$. C'est donc bien une densité de probabilité.

Ainsi, pour déterminer si une v.a. possède une densité, on pourra chercher à dériver sa fonction de répartition. Les v.a. à densité généralisent naturellement les v.a. uniformes. Mais elles apparaissent aussi quand on effectue une transformation régulière d'une v.a. uniforme.

Exercice 1.

On considère une v.a. U de loi uniforme sur $[0,1]$. Montrer que la v.a. $X := -\ln(U)$ est à densité, et identifier la densité.

Solution. On commence par calculer la fonction de répartition F_X de X. On remarque que $X \leqslant x$ si, et seulement si, $U \geqslant \exp(-x)$ car la fonction $x \mapsto \exp(-x)$ est décroissante. Par conséquent, pour tout $x \in \mathbb{R}$:

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(U \geqslant \exp(-x)) \,.$$

Si $x < 0$, alors $\exp(-x) > 1$ et donc $F_X(x) = 0$. Si $x \geqslant 0$, alors $\exp(-x) \in [0,1]$ et donc :

$$F_X(x) = \mathbb{P}(U \geqslant \exp(-x)) = 1 - \mathbb{P}(U < \exp(-x)) = 1 - \exp(-x) \,.$$

Finalement :

$$F_X(x) = \begin{cases} 0 & \text{si } x \leqslant 0 \\ 1 - \exp(-x) & \text{sinon.} \end{cases}$$

La fonction F_X est continue sur $\mathbb{R}$ et dérivable partout sauf en 0. Sa dérivée est :

$$p_X(x) := \exp(-x)\mathbf{1}_{]0,+\infty[}(x) \,.$$

Par conséquent, la v.a. X suit la loi de densité p_X. On reconnaît la densité de la loi exponentielle de paramètre 1.

On aurait pu abréger la preuve sitôt la fonction de répartition connue. On aurait pu dire alors que l'on reconnaissait la fonction de répartition de la loi exponentielle de paramètre 1. Comme la fonction de répartition caractérise la loi, cela aurait été suffisant pour conclure que la v.a. X est une v.a. à densité exponentielle de paramètre 1.

Remarque. Il existe des v.a. réelles qui ne sont ni discrètes, ni à densité. Un exemple est $Y := \min(a, X)$ avec X à densité p et $a \in \mathbb{R}$ fixé. On peut décrire la loi de Y par sa fonction de répartition qui possède une partie à densité sur $]-\infty, a[$ et une masse au point a. Plus exactement, la fonction de répartition de Y est :

$$F_Y(y) := \begin{cases} \displaystyle\int_{-\infty}^{y} p(x)\,\mathrm{d}x & \text{si } y < a \,, \\ 1 & \text{si } y \geqslant a \,. \end{cases}$$

Cette fonction possède un saut en $y = a$, d'amplitude $\displaystyle\int_a^{+\infty} p(x)\,\mathrm{d}x$. Le lecteur trouvera une approche générale des v.a. réelles dans des cours de probabilité plus avancés.

3 Moments de variables aléatoires à densité

3.1 Espérance

On veut définir une notion d'espérance pour les v.a. réelles à densité qui corresponde à la notion intuitive de moyenne et qui soit compatible avec la définition adoptée pour les v.a. réelles discrètes. Donner une définition générale de la notion d'espérance de v.a. réelles, indépendantes de leur qualité discrète ou à densité, est en fait une des motivations à développer une véritable théorie des probabilités avec une intégrale générale. Mais ce propos dépasse le cadre de ce cours. Ici, on va se contenter d'une définition qui s'applique aux v.a. réelles à densité, et l'on va vérifier que cette définition est bien compatible avec la définition d'espérance déjà donnée pour les v.a. réelles discrètes.

Définition 5. Soit $X : \Omega \to \mathbb{R}$ une v.a. réelle à densité p. Soit $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction continue par morceaux. Si $\displaystyle\int_{\mathbb{R}} |\varphi(x)|p(x)\,\mathrm{d}x < +\infty$, on dit que $\varphi(X)$ est *intégrable* et l'*espérance* de $\varphi(X)$ est définie par :

$$\mathbb{E}[\varphi(X)] := \int_{\mathbb{R}} \varphi(x)p(x)\,\mathrm{d}x\,. \tag{12}$$

Dans le cas où $\varphi(x) = x$, si $\displaystyle\int_{\mathbb{R}} |x|p(x)\,\mathrm{d}x < +\infty$, on dit que la v.a. réelle X est intégrable et son espérance est alors définie par :

$$\mathbb{E}[X] = \int_{\mathbb{R}} xp(x)\,\mathrm{d}x\,.$$

Cette formule correspond à la notion intuitive de moyenne pondérée.
Vérifions que la définition 5 est compatible avec la définition de l'espérance pour les v.a. discrètes réelles. Considérons une v.a. réelle X à densité p et un intervalle $I \subset \mathbb{R}$. On pose $Y = \mathbf{1}_I(X)$. On a deux possibilités pour calculer l'espérance de la v.a. réelle Y. D'une part, on peut appliquer la définition 5 ci-dessus pour obtenir :

$$\mathbb{E}\left[\mathbf{1}_I(X)\right] = \int_{\mathbb{R}} \mathbf{1}_I(x)p(x)\,\mathrm{d}x = \int_I p(x)\,\mathrm{d}x\,.$$

D'autre part, on peut aussi remarquer que Y ne prend pour valeurs que 0 ou 1, c'est donc une v.a. de Bernoulli. Dans ce cas, avec la définition de l'espérance d'une v.a. discrète (voir page 718), on obtient :

$$\mathbb{E}\left[\mathbf{1}_I(X)\right] = \mathbb{P}(X \in I) = \int_I p(x)\,\mathrm{d}x\,,$$

et l'on retrouve (fort heureusement) la même expression.
On peut énoncer quelques propriétés simples vérifiées par l'espérance de v.a. à densité.

Proposition 7. Soit X une v.a. réelle à densité p.
(i) Si $a \in \mathbb{R}$ et $\varphi_a : \mathbb{R} \to \mathbb{R}$ est la fonction constante égale à a, alors $\mathbb{E}[\varphi_a(X)] = a$.
(ii) Si $\varphi : \mathbb{R} \to \mathbb{R}$ et $\psi : \mathbb{R} \to \mathbb{R}_+$ sont deux fonctions continues par morceaux telle que $|\varphi(x)| \leqslant \psi(x)$ pour tout $x \in \mathbb{R}$ et $\displaystyle\int_{\mathbb{R}} \psi(x)p(x)\,\mathrm{d}x < +\infty$, alors $\varphi(X)$ est intégrable et $\mathbb{E}[\varphi(X)] \leqslant \mathbb{E}[\psi(X)]$.
(iii) Si a, $b \in \mathbb{R}$, si φ_1 et φ_2 sont deux fonctions de $\mathbb{R}$ dans $\mathbb{R}$ continues par morceaux, et si $\displaystyle\int_{\mathbb{R}} |\varphi_j(x)|p(x)\,\mathrm{d}x < +\infty$ pour $j = 1$ et 2, alors $a\varphi_1(X)+b\varphi_2(X)$ est intégrable et son espérance vaut :

$$\mathbb{E}[a\varphi_1(X) + b\varphi_2(X)] = a\mathbb{E}[\varphi_1(X)] + b\mathbb{E}[\varphi_2(X)]\,.$$

Démonstration.

Le *(i)* est trivial d'après la définition de l'espérance et le fait que $\displaystyle\int_{\mathbb{R}} p(x)\,\mathrm{d}x = 1$.

Pour le *(ii)*, $|\varphi(x)| \leqslant \psi(x)$ pour tout $x \in \mathbb{R}$ implique que :

$$\int_{\mathbb{R}} |\varphi(x)|p(x)\,\mathrm{d}x \leqslant \int_{\mathbb{R}} \psi(x)p(x)\,\mathrm{d}x < +\infty\,,$$

donc l'espérance de $\varphi(X)$ existe. Ensuite, on utilise le fait que l'on a aussi $\varphi(x) \leqslant \psi(x)$ pour tout $x \in \mathbb{R}$ ce qui implique que :

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x)p(x)\,\mathrm{d}x \leqslant \int_{\mathbb{R}} \psi(x)p(x)\,\mathrm{d}x = \mathbb{E}[\psi(X)]\,.$$

Pour le *(iii)* on applique l'inégalité triangulaire $|a\varphi_1(x)+b\varphi_2(x)| \leqslant |a||\varphi_1(x)|+|b||\varphi_2(x)|$ pour obtenir le fait que $\displaystyle\int_{\mathbb{R}} |a\varphi_1(x) + b\varphi_2(x)|p(x)\,\mathrm{d}x < +\infty$. Puis on utilise la linéarité de l'intégrale pour obtenir :

$$\int_{\mathbb{R}} [a\varphi_1(x) + b\varphi_2(x)]p(x)\,\mathrm{d}x = a\int_{\mathbb{R}} \varphi_1(x)p(x)\,\mathrm{d}x + b\int_{\mathbb{R}} \varphi_2(x)p(x)\,\mathrm{d}x\,. \qquad \blacksquare$$

Proposition 8. Soit X une v.a. réelle à densité p. Soit $\varphi : \mathbb{R} \to \mathbb{R}_+$ continue par morceaux telle que $\displaystyle\int_{\mathbb{R}} \varphi(x)p(x)\,\mathrm{d}x < +\infty$.
(i) $\mathbb{E}[\varphi(X)] \geqslant 0$.
(ii) Si $\mathbb{E}[\varphi(X)] = 0$, alors $\mathbb{P}(\varphi(X) = 0) = 1$.

Démonstration. Preuve du *(i)*. $\mathbb{E}[\varphi(X)] = \displaystyle\int_{\mathbb{R}} \varphi(x)p(x)\,\mathrm{d}x$ est la valeur de l'intégrale d'une fonction positive, elle est donc positive.

Preuve du *(ii)*. Pour tout $n \in \mathbb{N}^*$ on introduit l'événement $A_n = \{\varphi(X) \geqslant 1/n\}$. On vérifie que la fonction $\psi : y \in \mathbb{R}_+ \mapsto ny - \mathbf{1}_{[1/n,+\infty[}(y)$ est toujours positive, ce qui

montre que $\mathbb{E}[\psi(\varphi(X))] \geqslant 0$. Par linéarité de l'espérance, et le fait que φ est à valeurs positives :

$$\begin{aligned} 0 \leqslant \mathbb{E}[\psi(\varphi(X))] &= n\mathbb{E}[\varphi(X)] - \mathbb{E}[\mathbf{1}_{[1/n,+\infty[}(\varphi(X))] \\ &= n\mathbb{E}[\varphi(X)] - \mathbb{P}(\varphi(X) \geqslant 1/n) \\ &= n\mathbb{E}[\varphi(X)] - \mathbb{P}(A_n)\,. \end{aligned}$$

Comme $\mathbb{E}[\varphi(X)] = 0$ et $\mathbb{P}(A_n) \geqslant 0$, cela implique qu'en fait $\mathbb{P}(A_n) = 0$ pour tout n. Or $(A_n)_{n\in\mathbb{N}^*}$ est une suite croissante d'événements et $\bigcup_n A_n = \{\varphi(X) > 0\}$, donc :

$$\mathbb{P}(\varphi(X) > 0) = \lim_n \mathbb{P}(A_n) = 0.$$

Comme $\mathbb{P}(\varphi(X) > 0) + \mathbb{P}(\varphi(X) = 0) = 1$ on obtient le résultat. ■

Remarque. On peut facilement pousser plus loin l'étude de la compatibilité entre la définition de l'espérance d'une v.a. à densité et la définition de l'espérance d'une v.a. discrète. En effet, considérons une v.a. réelle X à densité p, une fonction $\varphi : \mathbb{R} \to \mathbb{R}$ en escalier, c'est-à-dire une fonction de la forme :

$$\varphi(x) := \sum_{i=1}^{n} \alpha_i \mathbf{1}_{I_i}(x)\,,$$

où $n \in \mathbb{N}^*$, $(I_i)_{i\in[\![1,n]\!]}$ est une suite d'intervalles de $\mathbb{R}$ deux à deux disjoints et $(\alpha_i)_{i\in[\![1,n]\!]}$ une suite de réels. Alors, ou bien en appliquant la linéarité de l'espérance et la définition 5, ou bien en appliquant la définition de l'espérance d'une v.a. discrète, on obtient :

$$\mathbb{E}[\varphi(X)] = \sum_{i=1}^{n} \alpha_i \int_{I_i} p(x)\,\mathrm{d}x\,.$$

3.2 Moments

Définition 6. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité, et soit n un entier naturel. Soit X une v.a. réelle à densité p telle que $\int_{\mathbb{R}} |x|^n p(x)\,\mathrm{d}x < +\infty$. Le $n^{\text{ème}}$ *moment* de X est défini par :

$$\mathbb{E}[X^n] := \int_{\mathbb{R}} x^n p(x)\,\mathrm{d}x\,.$$

On dit que X *admet un moment d'ordre* n si X^n est intégrable.

Lemme 9. Soit X une v.a. réelle admettant un moment d'ordre n. Alors X admet des moments d'ordre j pour tout $j \in [\![0, n]\!]$.

Démonstration. Soit $j \in [\![0, n]\!]$. On a $|x|^j \leqslant 1 + |x|^n$ pour tout $x \in \mathbb{R}$, donc :

$$\int_{\mathbb{R}} |x|^j p(x)\,\mathrm{d}x \leqslant \int_{\mathbb{R}} (1 + |x|^n) p(x)\,\mathrm{d}x = 1 + \int_{\mathbb{R}} |x|^n p(x)\,\mathrm{d}x < +\infty\,.$$ ■

On utilise aussi le moment centré d'ordre n de la v.a. X, défini par $\mathbb{E}[(X - \mathbb{E}[X])^n]$. Remarquons au passage que si le moment non centré existe, alors le moment centré existe, puisque c'est l'espérance d'un polynôme en X de degré n et que l'on vient de voir que les moments de degré inférieur à n existaient. Le cas particulier réellement important est le cas où $n = 2$. Le moment centré d'ordre 2 de X s'appelle la variance de X.

Définition 7. Si X est une v.a. réelle de carré intégrable, alors la *variance* de X est définie par :

$$\mathrm{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] .$$

Notez que $\mathrm{Var}(X) \geqslant 0$. La racine carrée positive de la variance de X est l'écart-type de X, encore appelé déviation standard. Insistons encore sur le fait que l'écart-type a la dimension de la variable aléatoire : si celle-ci s'exprime en mètres, l'écart-type s'exprime en mètres et la variance en mètres carrés. Il faut connaître les deux formules suivantes :

Proposition 10. Si X a un moment d'ordre 2, alors pour λ, μ réels :

$$\mathrm{Var}(\lambda X + \mu) = \lambda^2 \mathrm{Var}(X) ,$$

et (formule de Huyghens) :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 .$$

En particulier, $\mathbb{E}[X]^2 \leqslant \mathbb{E}[X^2]$, avec égalité si, et seulement si, $\mathbb{P}(X = \mathbb{E}[X]) = 1$.

Démonstration. La première formule est immédiate. Pour la formule de Huygens :

$$\mathrm{Var}(X) = \mathbb{E}[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2] = \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 .$$

Ici, on a utilisé le fait que l'espérance d'une constante est la constante elle-même.

Quant à la dernière inégalité elle vient du fait qu'une variance est toujours positive ou nulle. Si la variance est nulle, alors appliquant le *(ii)* de la proposition 8 de la page 765 à $\varphi(x) = (x - \mathbb{E}[X])^2$, on obtient $\mathbb{P}(\varphi(X) = 0) = 1$. ■

Pour résumer, si X est une v.a. réelle à densité p définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, si $\displaystyle\int_{\mathbb{R}} x^2 p(x)\,\mathrm{d}x$ est fini, alors X est de carré intégrable, et la variance de X est donnée par :

$$\mathrm{Var}(X) = \int_{\mathbb{R}} x^2 p(x)\,\mathrm{d}x - \left(\int_{\mathbb{R}} x p(x)\,\mathrm{d}x\right)^2 .$$

3.3 Moments des lois usuelles

Variable uniforme

Une v.a. X est de loi uniforme sur l'intervalle $[a, b]$ $(-\infty < a < b < +\infty)$ si elle a pour densité $p(x) = (b - a)^{-1}\mathbf{1}_{]a,b[}(x)$. On calcule facilement :

$$\mathbb{E}[X] = \frac{1}{b-a}\int_a^b x\,\mathrm{d}x = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2} ,$$

c'est-à-dire que la moyenne est le milieu de l'intervalle $[a, b]$. La variance est :

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] &= \frac{1}{b-a}\int_a^b \left(x - \frac{a+b}{2}\right)^2 \mathrm{d}x \\ &= \frac{1}{b-a}\int_{-(b-a)/2}^{(b-a)/2} y^2\,\mathrm{d}y = \frac{(b-a)^2}{12}\,,\end{aligned}$$

ce qui montre que l'écart-type est proportionnel à la longueur de l'intervalle.

Variable exponentielle

Si X est de loi exponentielle de paramètre $\lambda > 0$ (dont la densité est (9), voir page 761), on a :

$$\mathbb{E}[X] = \int_0^{+\infty} x\lambda e^{-\lambda x}\,\mathrm{d}x = \left[-xe^{-\lambda x}\right]_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}\,,$$

en intégrant par parties. On calcule de même :

$$\mathbb{E}[X^2] = \int_0^{+\infty} x^2\lambda e^{-\lambda x}\,\mathrm{d}x = \left[-x^2e^{-\lambda x}\right]_0^{+\infty} + 2\int_0^{+\infty} xe^{-\lambda x}\,\mathrm{d}x = \frac{2}{\lambda^2}\,,$$

et donc :

$$\mathrm{Var}(X) = \frac{1}{\lambda^2}\cdot$$

Variables gaussiennes, lois normales

a) Une v.a. réelle X de densité gaussienne (11) (voir page 761) est appelée variable gaussienne (ou normale) centrée réduite, et sa loi est notée $\mathcal{N}(0,1)$. L'adjectif « centré » et le nombre « 0 » traduisent le fait que $\mathbb{E}[X] = 0$ (la densité g est paire, et donc $x \mapsto xg(x)$ est impaire), tandis que l'adjectif « réduit » et le nombre « 1 » traduisent que $\mathrm{Var}(X) = 1$, comme on peut le voir en calculant $\mathbb{E}[X^2]$ par intégration par parties :

$$\begin{aligned}\mathbb{E}[X^2] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^2 \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\left[-x\exp\left(-\frac{x^2}{2}\right)\right]_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x = 1\,.\end{aligned}$$

b) Soient $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}_+$. Si $X \sim \mathcal{N}(0,1)$, alors la v.a. $Y = \mu + \sigma X$ est appelée variable gaussienne de moyenne μ et variance σ^2. On vérifie en effet facilement que $\mathbb{E}[Y] = \mu + \mathbb{E}[X] = \mu$ et $\mathrm{Var}(Y) = \sigma^2\mathrm{Var}(X) = \sigma^2$. La loi de Y est notée $\mathcal{N}(\mu, \sigma^2)$, appelée loi gaussienne ou loi normale. De plus, si $\sigma > 0$, Y a pour densité :

$$g_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\,.$$

Si $\sigma = 0$, alors la Y est une v.a. constante égale à μ.

Démonstration. Le cas $\sigma = 0$ est trivial. Soit $\sigma > 0$. On exprime la fonction de répartition de Y en fonction de celle de X :

$$F_Y(y) = \mathbb{P}(\sigma X + \mu \leqslant y) = \mathbb{P}\left(X \leqslant \frac{y-\mu}{\sigma}\right) = F_X\left(\frac{y-\mu}{\sigma}\right)\,.$$

On dérive alors les deux membres extrêmes de la ligne ci-dessus : à gauche on obtient la densité cherchée de Y, à droite en utilisant le théorème de dérivation des fonctions composées et le fait que la densité de X est par hypothèse $g(x)$ on obtient la densité $g_{\mu,\sigma^2}(y)$ comme annoncé. ∎

c) Soient $\mu, \mu' \in \mathbb{R}$ et $\sigma, \sigma' \in \mathbb{R}_+$. Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $Y = \sigma' X + \mu'$ suit la loi $\mathcal{N}(\sigma'\mu + \mu', \sigma^2\sigma'^2)$. En particulier, si $\sigma > 0$, alors $Y = (X - \mu)/\sigma$ suit la loi $\mathcal{N}(0, 1)$. La preuve est la même que pour le cas précédent.
La loi normale se rencontre dans de nombreuses applications. On verra dans le module IV.4 qu'elle modélise les erreurs de mesure lors d'observations expérimentales.

3.4 Inégalités célèbres

Il existe (au moins) trois inégalités célèbres. On ne demande pas de les connaître par cœur, mais savoir qu'elles existent et savoir les retrouver pourra vous être utile.

Proposition 11 (Inégalité de Markov). Si Y est une v.a. réelle intégrable, à valeurs positives ou nulles, alors pour tout $y > 0$ on a :

$$\mathbb{P}\big(Y \geqslant y\big) \leqslant \frac{1}{y}\mathbb{E}[Y]\,.$$

Démonstration. Soit $y > 0$. En utilisant l'égalité triviale $1 = \mathbf{1}_{Y \geqslant y} + \mathbf{1}_{Y < y}$:

$$\mathbb{E}[Y] = \mathbb{E}[Y\mathbf{1}_{Y \geqslant y}] + \mathbb{E}[Y\mathbf{1}_{Y<y}] \geqslant \mathbb{E}[Y\mathbf{1}_{Y \geqslant y}] \geqslant \mathbb{E}[y\mathbf{1}_{Y \geqslant y}] = y\mathbb{E}[\mathbf{1}_{Y \geqslant y}] = y\mathbb{P}(Y \geqslant y)\,,$$

ce qui est équivalent à l'inégalité de Markov en divisant par y. ∎

Corollaire 12 (Inégalité de Bienaymé-Tchebychev). Si X est une v.a. de carré intégrable, alors pour tout $t > 0$ on a :

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \geqslant t\big) \leqslant \frac{1}{t^2}\mathrm{Var}(X)\,.$$

Démonstration. On note $m = \mathbb{E}[X]$. On applique l'inégalité de Markov à la v.a. $Y = (X - m)^2$ et au réel positif $y = t^2$. Comme :

$$\mathbb{P}(|X - m| \geqslant t) = \mathbb{P}\left((X - m)^2 \geqslant t^2\right) \leqslant \frac{1}{t^2}\mathbb{E}\left[(X - m)^2\right] = \frac{1}{t^2}\mathrm{Var}(X)\,,$$

l'inégalité de Bienaymé-Tchebychev est ainsi démontrée. ∎

Proposition 13 (Inégalité de Jensen). Soient X une v.a. réelle intégrable et $f : \mathbb{R} \to \mathbb{R}$ une fonction convexe, telle que $f(X)$ soit intégrable. Alors :

$$\mathbb{E}[f(X)] \geqslant f\big(\mathbb{E}[X]\big)\,.$$

Démonstration. Une caractérisation des fonctions convexes est la suivante (voir l'exercice 19 de la page 529 du livre [L2]) : f est convexe sur $\mathbb{R}$ si, et seulement si :

$$\forall x_0 \in \mathbb{R},\ \exists a \in \mathbb{R} \text{ tel que } f(x) \geqslant f(x_0) + a(x - x_0)\,\forall x \in \mathbb{R}\,.$$

On applique cette caractérisation avec $x_0 = \mathbb{E}[X]$, et l'on obtient qu'il existe a tel que $f(x) \geqslant f(\mathbb{E}[X]) + a(x - \mathbb{E}[X])$ pour tout $x \in \mathbb{R}$. Donc :

$$f(X) \geqslant f(\mathbb{E}[X]) + a(X - \mathbb{E}[X]).$$

En appliquant l'espérance aux deux membres de cette inégalité, on obtient le résultat. ■

On utilise souvent l'inégalité de Jensen avec $f(x) = |x|$:

$$\mathbb{E}\big[|X|\big] \geqslant \big|\mathbb{E}[X]\big|,$$

et avec $f(x) = x^2$:

$$\mathbb{E}[X^2] \geqslant \mathbb{E}[X]^2.$$

4 Vecteurs aléatoires

4.1 Définition et loi

Soit $(X_i)_{i\in[\![1,n]\!]}$ une famille de v.a. réelles définies sur le même espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. Si I_i est un intervalle de $\mathbb{R}$ pour tout $i \in [\![1, n]\!]$, alors :

$$\{(X_1, \ldots, X_n) \in I_1 \times \cdots \times I_n\} = \bigcap_{i=1}^{n} \{X_i \in I_i\}$$

appartient à $\mathcal{A}$ car les parties $\{X_i \in I_i\}$ appartiennent à $\mathcal{A}$ pour tout $i \in [\![1, n]\!]$. Rappelons que $\mathcal{B}(\mathbb{R}^n)$ est la plus petite tribu sur $\mathbb{R}^n$ contenant les pavés. Donc, pour tout $B \in \mathcal{B}(\mathbb{R}^n)$:

$$\{(X_1, \ldots, X_n) \in B\} \in \mathcal{A}.$$

Cette propriété nous permet de définir simplement un vecteur aléatoire comme un vecteur de v.a. réelles définies sur le même espace de probabilité et d'introduire la loi de ce vecteur.

Définition 8. Soient $n \geqslant 1$ et $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^n$.
(i) On dit que X est un *vecteur aléatoire* si ses coordonnées X_i sont des v.a. réelles définies sur $(\Omega, \mathcal{A}, \mathbb{P})$.
(ii) La *loi* du vecteur aléatoire est la probabilité Q sur $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$ définie par :

$$Q(B) := \mathbb{P}(X \in B), \quad B \in \mathcal{B}(\mathbb{R}^n).$$

Comme dans le cas $n = 1$, il suffit de connaître la loi sur les pavés semi-infinis pour la connaître entièrement.

Proposition 14. Soient $n \geqslant 1$ et $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^n$ un vecteur aléatoire. La loi du vecteur aléatoire X est caractérisée par sa fonction de répartition :

$$F_X(x_1, \ldots, x_n) := \mathbb{P}\left(X \in]-\infty, x_1] \times \cdots \times]-\infty, x_n]\right)$$

définie pour tout $(x_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^n$.

4.2 Vecteur aléatoire à densité

Définition 9. Le vecteur aléatoire X est de *densité* p_X si $p_X : \mathbb{R}^n \to \mathbb{R}_+$ est une fonction continue par morceaux et si pour tout $(x_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^n$:

$$\mathbb{P}\left(X \in]-\infty, x_1] \times \cdots \times]-\infty, x_n]\right) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p_X(y)\,\mathrm{d}y\,. \tag{13}$$

On a alors $\mathbb{P}(X \in B) = \displaystyle\int_B p_X(x)\,\mathrm{d}x$ pour tout B élément de la tribu borélienne de $\mathbb{R}^n$. Alors nécessairement, $p_X : \mathbb{R}^n \to \mathbb{R}_+$ est une fonction d'intégrale 1 :

$$1 = \mathbb{P}(X \in \mathbb{R}^n) = \int_{\mathbb{R}^n} p_X(x)\,\mathrm{d}x\,.$$

Proposition 15 (Formule des marginales). Si $X = (X_1, \ldots, X_n)$ est un vecteur aléatoire de densité p_X, alors X_1 est une v.a. à densité p_{X_1} donnée par :

$$p_{X_1}(x_1) = \int_{\mathbb{R}^{n-1}} p_X(x_1, x_2, \ldots, x_n)\,\mathrm{d}x_2 \ldots \mathrm{d}x_n\,. \tag{14}$$

Démonstration. La relation (13) et la formule de Fubini entraînent pour tout $x_1 \in \mathbb{R}$:

$$\mathbb{P}(X_1 \leqslant x_1) = \int_{-\infty}^{x_1} \left[\int_{\mathbb{R}^{n-1}} p_X(y_1, y_2, \ldots, y_n)\,\mathrm{d}y_2 \ldots \mathrm{d}y_n\right] \mathrm{d}y_1\,,$$

ce qui montre que X_1 est à densité p_{X_1} donnée par (14). ∎

La *densité marginale* p_{X_1} s'obtient donc en intégrant la *densité jointe* p_X par rapport aux variables que l'on veut éliminer. Le même raisonnement s'applique aux autres composantes $X_2, \ldots, X_n$ et aux sous-vecteurs, comme par exemple $(X_2, \ldots, X_n)$ qui a pour densité :

$$p_{(X_2,\ldots,X_n)}(x_2, \ldots, x_n) = \int_{\mathbb{R}} p_X(x_1, x_2, \ldots, x_n)\,\mathrm{d}x_1\,.$$

Si $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction continue par morceaux, l'espérance de $f(X)$ se calcule par la formule :

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}^n} f(x) p_X(x)\,\mathrm{d}x\,, \tag{15}$$

sous réserve que $\displaystyle\int_{\mathbb{R}^n} |f(x)| p_X(x)\,\mathrm{d}x < +\infty$. On dit alors que $f(X)$ est intégrable. L'espérance vérifie les propriétés énoncées dans les propositions 7 et 8 de la page 765.

Exemple. Achille tire à l'arc sur une cible circulaire de rayon unité. On suppose qu'Achille est suffisamment loin de la cible (ou maladroit) pour que le point M d'impact de la flèche soit uniformément distribué sur la cible. Les coordonnées cartésiennes de M dans le disque unité $D := \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$ constituent un couple aléatoire (X, Y) de densité :

$$p_{(X,Y)}(x, y) := \frac{1}{\pi} \mathbf{1}_D(x, y)\,,$$

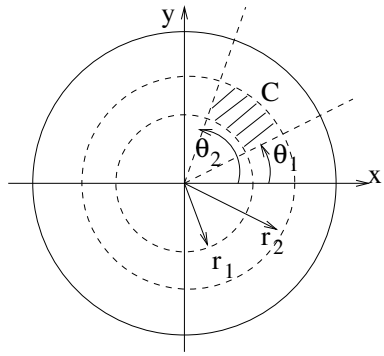


FIGURE IV.3.4 – Tir à l'arc

uniforme sur le disque D, par hypothèse. L'abscisse X est distribuée selon la densité marginale :

$$p_X(x) = \int_{\mathbb{R}} p_{(X,Y)}(x,y)\,\mathrm{d}y = \frac{2}{\pi}\sqrt{1-x^2}\mathbf{1}_{]-1,1[}(x)\,,$$

et Y a ici la même densité, $p_Y = p_X$. Il est naturel de paramétrer M par ses coordonnées polaires $R = \sqrt{X^2+Y^2} \in [0,+\infty[$, $\theta \in [0,2\pi[$ avec $\cos(\theta) = X/R$, $\sin(\theta) = Y/R$ si $(X,Y) \neq (0,0)$ et $\theta = 0$ si $(X,Y) = (0,0)$. La v.a. R est la distance au centre de la cible, θ est sa position angulaire. Pour $0 \leqslant r_1 \leqslant r_2 \leqslant 1$ et $0 \leqslant \theta_1 \leqslant \theta_2 < 2\pi$ on a $\{R \in [r_1,r_2],\ \theta \in [\theta_1,\theta_2]\} = \{M \in C\}$ où C est le secteur angulaire de couronne indiqué sur la figure IV.3.4, dont on calcule aisément la surface $|C|$. Donc :

$$\mathbb{P}(R \in [r_1,r_2],\ \theta \in [\theta_1,\theta_2]) = \frac{|C|}{|D|} = \frac{\theta_2-\theta_1}{2\pi}(r_2^2-r_1^2) = \int_{[r_1,r_2]\times[\theta_1,\theta_2]} q(r,\theta)\,\mathrm{d}r\,\mathrm{d}\theta\,,$$

avec $q(r,\theta) = (1/\pi)r\mathbf{1}_{]0,1[\times]0,2\pi[}(r,\theta)$. Puisque r_1, r_2, θ_1 et θ_2 sont arbitraires, on en déduit que le couple (R,θ) a pour densité $p_{(R,\theta)} = q$. Par la formule des marginales (14) on obtient $p_R(r) = 2r\mathbf{1}_{]0,1[}(r)$ et $p_\theta(\theta) = 1/(2\pi)\mathbf{1}_{]0,2\pi[}(\theta)$ (densité uniforme).

Exercice 2.

Déterminer la loi du résultat S d'un tir, ainsi que son espérance, dans les deux scénarios suivants :

a) $S := 1-$distance de la flèche au centre de la cible,

b) en divisant la cible en 3 couronnes d'épaisseur $1/3$, notées C_1 (= disque de rayon $1/3$), C_2 et C_3, on définit S par $S := \sum_{k=1}^{3}\left(\frac{4-k}{3}\right)\mathbf{1}_{M\in C_k}$.

Solution. Dans le scénario a, la loi de la v.a. $S = 1-R$ est telle que, pour $0 \leqslant s \leqslant 1$:

$$\mathbb{P}(S \leqslant s) = \mathbb{P}(R \geqslant 1-s) = \int_{1-s}^{1} p_R(r)\,\mathrm{d}r = \int_0^s p_R(1-r)\,\mathrm{d}r\,.$$

Elle est donc à densité :

$$p_S(s) = p_R(1-s) = 2(1-s)\mathbf{1}_{]0,1[}(s)\,,$$

et on a : $\mathbb{E}[S] = 1 - \mathbb{E}[R] = 1 - 2\displaystyle\int_0^1 r^2\,\mathrm{d}r = \frac{1}{3}$.

Dans le scénario b, S est une v.a. discrète à valeurs dans $\{1/3, 2/3, 1\}$. Sa loi est donnée par :

$$\mathbb{P}(S = 1) = \mathbb{P}\big(R \leqslant \frac{1}{3}\big) = 2\int_0^{1/3} r\,\mathrm{d}r = \frac{1}{9}\,,$$

$$\mathbb{P}\big(S = \frac{1}{3}\big) = \mathbb{P}\big(R > \frac{2}{3}\big) = 2\int_{2/3}^1 r\,\mathrm{d}r = 1 - \frac{4}{9} = \frac{5}{9}\,,$$

et on a bien sûr $\mathbb{P}(S = \frac{2}{3}) = 1 - \mathbb{P}(S = 1) - \mathbb{P}(S = \frac{1}{3}) = \frac{1}{3}$.

Donc $\mathbb{E}[S] = 1 \times \frac{1}{9} + \frac{2}{3} \times \frac{3}{9} + \frac{1}{3} \times \frac{5}{9} = \frac{14}{27}$.

---

5 Indépendance

5.1 Définition

Définition 10. Soient X et Y deux v.a. réelles définies sur le même espace de probabilité. Les v.a. sont dites *indépendantes* si pour tout choix d'intervalles I et J :

$$\mathbb{P}(X \in I,\ Y \in J) = \mathbb{P}(X \in I)\mathbb{P}(Y \in J)\,. \tag{16}$$

La notion d'indépendance pour des v.a. réelles prolonge la définition 4 de la page 686 d'indépendance pour des événements. Cette notion est assez intuitive, cependant, comme on l'a vu dans le cas de v.a. discrètes, l'intuition peut être mise en défaut et il faut prendre garde à ne pas annoncer un résultat d'indépendance basé sur un raisonnement fait avec les mains.

Proposition 16. Soient X et Y deux v.a. réelles définies sur le même espace de probabilité.
(i) Si X et Y sont indépendantes, alors pour tous boréliens A et B de $\mathbb{R}$ on a :

$$\mathbb{P}(X \in A,\ Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\,. \tag{17}$$

(ii) Si pour tous x, y réels on a :

$$\mathbb{P}(X \leqslant x,\ Y \leqslant y) = \mathbb{P}(X \leqslant x)\mathbb{P}(Y \leqslant y)\,, \tag{18}$$

alors X et Y sont indépendantes.

Démonstration. On a vu la démonstration dans le cas discret, mais ce n'est pas évident dans le cas général. Il faut soit faire des « découpages » comme dans la démonstration de la proposition 1 de la page 755, en examinant tous les cas possibles, soit utiliser des outils plus puissants qu'on ne considérera pas dans ce cours. On admettra le résultat ici. ■

Donc, pour montrer que deux v.a. sont indépendantes, (18) suffit, et une fois qu'on l'a montré, alors on peut appliquer (17). La définition s'étend au cas d'une suite finie ou infinie de v.a. réelles ou de vecteurs aléatoires définis sur le même espace $(\Omega, \mathcal{A}, \mathbb{P})$:

Définition 11. *(i)* Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite finie de v.a. réelles définies sur le même espace de probabilité. Les v.a. sont dites *indépendantes* si pour tout choix d'intervalles I_i, $i \in [\![1,n]\!]$:

$$\mathbb{P}(X_1 \in I_1,\ \ldots,\ X_n \in I_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in I_i)\,. \tag{19}$$

(ii) Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite finie de vecteurs aléatoires à valeurs dans $\mathbb{R}^{d_i}$ définis sur le même espace de probabilité. Les vecteurs sont dits *indépendants* si pour tout choix de pavés $J_i \subset \mathbb{R}^{d_i}$, $i \in [\![1,n]\!]$:

$$\mathbb{P}(X_1 \in J_1,\ \ldots,\ X_n \in J_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in J_i)\,. \tag{20}$$

(iii) Une suite infinie de v.a. (respectivement vecteurs aléatoires) $(X_i)_{i\in\mathbb{N}}$ est dite *indépendante*, si pour tout $n \geqslant 1$, la suite de v.a. (respectivement vecteurs) $(X_i)_{i\in[\![0,n]\!]}$ est indépendante.

5.2 Indépendance de variables aléatoires à densité

Lorsque X et Y sont discrètes à valeurs dans E, on a déjà constaté en prenant des singletons $I = \{x\}$ et $J = \{y\}$ que l'indépendance équivalait à la propriété :

$$p_{(X,Y)}(x,y) = p_X(x)p_Y(y)\,,\ \ \forall x,y \in E\,, \tag{21}$$

où $p_X(x) = \mathbb{P}(X = x)$. Une propriété similaire existe aussi dans le cas de v.a. à densité.

Proposition 17. Soit $X = (X_i)_{i\in[\![1,n]\!]}$ un vecteur aléatoire.
(i) Si X est à densité p_X et si la densité jointe p_X se factorise :

$$\forall (x_1,\ldots,x_n) \in \mathbb{R}^n, \quad p_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)\,, \tag{22}$$

alors les v.a. X_i sont indépendantes.
(ii) Si les v.a. X_i sont indépendantes et à densité p_{X_i} pour tout $i \in [\![1,n]\!]$, alors le vecteur aléatoire X est à densité p_X donnée par (22).

Démonstration. Preuve de *(i)*. Si la densité jointe se met sous la forme produit (22), alors pour toute suite d'intervalles I_i, on a :

$$\int_{I_1\times\cdots\times I_n} p_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n)\,\mathrm{d}x_1\ldots\mathrm{d}x_n = \int_{I_1\times\cdots\times I_n} \prod_{i=1}^{n} p_{X_i}(x_i)\,\mathrm{d}x_1\ldots\mathrm{d}x_n\,.$$

Par la formule de Fubini, l'intégrale se factorise, et l'on obtient précisément (19), ce qui montre que les v.a. X_i sont indépendantes.

Preuve de *(ii)*. Si les v.a. X_i sont indépendantes et à densité p_{X_i}, alors pour tout

$(x_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^n$ on a :

$$\mathbb{P}\left(X \in]-\infty, x_1] \times \cdots \times]-\infty, x_n]\right) = \mathbb{P}\left(X_1 \in]-\infty, x_1]\right) \cdots \mathbb{P}\left(X_n \in]-\infty, x_n]\right)$$
$$= \int_{-\infty}^{x_1} p_{X_1}(y_1)\,\mathrm{d}y_1 \cdots \int_{-\infty}^{x_n} p_{X_n}(y_n)\,\mathrm{d}y_n\,.$$

En utilisant la formule de Fubini, on obtient (13) avec p_X définie par (22). ■

Exercice 3.

Reprenons l'exemple du tir à l'arc de la page 771.
a) Montrer que les v.a. X et Y ne sont pas indépendantes.
b) Montrer que les v.a. R et θ sont indépendantes.

Solution. a) Pour montrer que les deux v.a. X et Y ne sont pas indépendantes, il suffit d'exhiber deux intervalles I et J tels que :

$$\mathbb{P}(X \in I, Y \in J) \neq \mathbb{P}(X \in I)\mathbb{P}(Y \in J)\,.$$

Prenons $I = J = [3/4, 1]$. Sans faire de calcul, on a clairement $\mathbb{P}(X \in I) > 0$ et $\mathbb{P}(Y \in J) > 0$, alors que :

$$\mathbb{P}(X \in I, Y \in J) = \mathbb{P}\Big(X \geqslant \frac{3}{4}, Y \geqslant \frac{3}{4}\Big) = 0\,,$$

car $(3/4)^2 + (3/4)^2 = 18/16 > 1$. Cela montre que :

$$\mathbb{P}(X \in I, Y \in J) = 0 \neq \mathbb{P}(X \in I)\mathbb{P}(Y \in J)\,,$$

et que les v.a. ne sont pas indépendantes.
b) Pour montrer que les deux v.a. R et θ sont indépendantes, un moyen est de montrer que pour *tous* intervalles I et J, on a $\mathbb{P}(R \in I, \theta \in J) = \mathbb{P}(R \in I)\mathbb{P}(\theta \in J)$. Mais il est plus efficace ici d'appliquer la proposition 17 et de montrer que le densité jointe du couple (R, θ) se factorise. La densité jointe est :

$$q(r, \theta) := \frac{1}{\pi} r \mathbf{1}_{]0,1[\times]0,2\pi[}(r, \theta) = p_R(r) p_\theta(\theta)\,,$$

avec $p_R(r) = 2r\mathbf{1}_{]0,1[}(r)$ et $p_\theta(\theta) = 1/(2\pi)\mathbf{1}_{]0,2\pi[}(\theta)$.
Les v.a. R et θ sont donc indépendantes.

Proposition 18. *(i)* Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite finie de v.a. réelles à densité indépendantes. Pour toutes fonctions $f_i : \mathbb{R} \to \mathbb{R}$ telles que $f_i(X_i)$ est intégrable, on a :

$$\mathbb{E}[f_1(X_1) \cdots f_n(X_n)] = \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)]\,. \tag{23}$$

(ii) Réciproquement, soit $X = (X_i)_{i\in[\![1,n]\!]}$ un vecteur aléatoire à densité tel que, pour toutes fonctions $f_i : \mathbb{R} \to \mathbb{R}$ telles que $f_i(X_i)$ soit intégrable, on a $\prod_{i=1}^{n} f_i(X_i)$ est intégrable et l'égalité (23) est vraie. Alors les v.a. $(X_i)_{i\in[\![1,n]\!]}$ sont indépendantes.

Cette proposition permet de simplifier les calculs d'espérance lorsque l'on a affaire à des v.a. indépendantes, et elle est aussi un outil pour montrer que des v.a. sont indépendantes. La proposition entraîne en particulier que toute sous-famille d'une famille de v.a. indépendante est encore indépendante : par exemple X_1 et X_2 sont indépendantes si $X_1, \ldots, X_n$ le sont. La proposition 18 est en réalité tout à fait générale, mais nous n'en donnons la preuve que dans le cas à densité, sachant que nous en avons déjà donné la preuve dans le cas discret.

Démonstration. Preuve du *(i)*. Par indépendance, le vecteur X est à densité p_X donnée par, pour tout $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$:

$$p_X(x) = \prod_{i=1}^{n} p_{X_i}(x_i)\,,$$

avec la notation $X = (X_1, \ldots, X_n)$.

Soient des fonctions f_i telles que $\displaystyle\int_{\mathbb{R}} |f_i(x)| p_{X_i}(x)\,\mathrm{d}x < +\infty$ pour tout $i \in [\![1, n]\!]$. On a :

$$\begin{aligned}\int_{\mathbb{R}^n} \left|\prod_{i=1}^{n} f_i(x_i)\right| p_X(x)\,\mathrm{d}x &= \int_{\mathbb{R}^n} \left(\prod_{i=1}^{n} |f_i(x_i)|\right)\left(\prod_{i=1}^{n} p_{X_i}(x_i)\right)\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} |f_i(x_i)| p_{X_i}(x_i)\,\mathrm{d}x_i < +\infty\,,\end{aligned}$$

ce qui montre que $\prod_{i=1}^{n} f_i(X_i)$ est intégrable. Il s'en suit :

$$\begin{aligned}\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] &= \int_{\mathbb{R}^n} \prod_{i=1}^{n} f_i(x_i) p_X(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^n} \left(\prod_{i=1}^{n} f_i(x_i)\right)\left(\prod_{i=1}^{n} p_{X_i}(x_i)\right)\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} f_i(x_i) p_{X_i}(x_i)\,\mathrm{d}x_i = \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)]\,.\end{aligned}$$

Preuve du *(ii)*. Soit I_i, $i \in [\![1, n]\!]$ des intervalles de $\mathbb{R}$.

En appliquant (23) avec les fonctions indicatrices $f_i = \mathbf{1}_{I_i}$, on obtient directement (19). ■

5.3 Covariance et variance

Toutes les propositions de cette section sont valables pour des v.a. discrètes ou à densité.

Définition 12. Soient X et Y deux v.a. réelles de carré intégrable définies sur le même espace de probabilité. La *covariance* de X et Y est définie par :

$$\mathrm{Cov}(X, Y) := \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]\,.$$

Cette définition se généralise à des vecteurs aléatoires de taille quelconque.

Définition 13. Soit $X = (X_i)_{i\in[\![1,d]\!]}$ un vecteur aléatoire, dont les coordonnées X_i sont de carré intégrable. La *matrice de covariance* de X est la matrice réelle C de taille $d \times d$ dont les coefficients sont :

$$C_{ij} := \mathrm{Cov}(X_i, X_j)\,, \qquad i, j \in [\![1, d]\!]\,.$$

Cette matrice est la généralisation de la variance pour les v.a. réelles.

Proposition 19. *(i)* La matrice de covariance d'un vecteur aléatoire $X = (X_i)_{i\in[\![1,d]\!]}$ est une matrice symétrique positive.
(ii) Soient d et n des entiers strictement positifs, et $X = (X_i)_{i\in[\![1,d]\!]}$ un vecteur aléatoire de moyenne $\mu = (\mu_i)_{i\in[\![1,d]\!]}$ et de matrice de covariance $C = (C_{ij})_{i,j\in[\![1,d]\!]}$. Soit A une matrice réelle de taille $n \times d$. Le vecteur aléatoire $Y := AX$ a pour moyenne $A\mu$ et pour matrice de covariance ACA^t où A^t est la transposée de A.

Démonstration. Preuve de *(i)*. La matrice de covariance C d'un vecteur aléatoire X est symétrique et réelle, donc diagonalisable. Soit λ une de ses valeurs propres, et soit $\nu = (\nu_1, \ldots, \nu_d)$ un vecteur propre (non nul) associé. On calcule la variance de la v.a. $\nu^t X = \sum_{i=1}^{d} \nu_i X_i$:

$$\begin{aligned}\mathrm{Var}(\nu^t X) &= \mathbb{E}[(\nu^t X)^2] - \mathbb{E}[\nu^t X]^2 = \sum_{i,j=1}^{d} \nu_i\nu_j \mathbb{E}[X_i X_j] - \sum_{i,j=1}^{d} \nu_i\nu_j \mathbb{E}[X_i]\mathbb{E}[X_j] \\ &= \sum_{i,j=1}^{d} \nu_i\nu_j \mathrm{Cov}(X_i, X_j) = \nu^t C \nu\,.\end{aligned}$$

Comme ν est vecteur propre de C, on a $C\nu = \lambda\nu$, et donc :

$$\mathrm{Var}(\nu^t X) = \lambda \nu^t \nu = \lambda \|\nu\|^2\,,$$

où $\|\cdot\|$ est la norme euclidienne sur $\mathbb{R}^d$. Comme une variance est un nombre positif, et que la norme euclidienne du vecteur ν est strictement positive, λ doit être positif.
Preuve de *(ii)*. La linéarité de l'espérance implique le fait que l'espérance de $Y = AX$ est $\mathbb{E}[Y] = A\mathbb{E}[X]$. Pour déterminer la matrice de covariance, on effectue le calcul suivant :

$$\begin{aligned}\mathrm{Cov}(Y_i, Y_j) &= \mathbb{E}[(AX)_i (AX)_j] - \mathbb{E}[(AX)_i]\mathbb{E}[(AX)_j] \\ &= \sum_{k,l=1}^{n} \mathbb{E}[A_{ik} X_k A_{jl} X_l] - \mathbb{E}[A_{ik} X_k]\mathbb{E}[A_{jl} X_l] \\ &= \sum_{k,l=1}^{n} A_{ik} A_{jl} \left(\mathbb{E}[X_k X_l] - \mathbb{E}[X_k]\mathbb{E}[X_l]\right) \\ &= \sum_{k,l=1}^{n} A_{ik} C_{kl} A_{jl} = (ACA^t)_{ij}\,.\end{aligned}$$

■

Considérons X et Y deux v.a. réelles indépendantes de carré intégrable. D'après la proposition 18 :

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right] = \mathbb{E}\left[X-\mathbb{E}[X]\right] \times \mathbb{E}\left[Y-\mathbb{E}[Y]\right] = 0\,.$$

Cela montre que la matrice de covariance d'un vecteur aléatoire composé de variables indépendantes est diagonale.

Proposition 20. *(i)* Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite finie de v.a. réelles de carré intégrable. Alors :

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i,X_j)\,.$$

(ii) Si les $(X_i)_{i\in[\![1,n]\!]}$ sont deux à deux indépendantes (c'est-à-dire, si X_i et X_j sont indépendantes pour tout $i \neq j$), alors :

$$\forall i \neq j,\ \ \mathrm{Cov}(X_i,X_j) = 0\,,$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)\,.$$

Démonstration. La preuve est identique à celle de la proposition 21 de la page 734. ■

On termine ce paragraphe par la célèbre inégalité de Cauchy-Schwarz.

Proposition 21 (Inégalité de Cauchy-Schwarz). Soient X et Y deux v.a. de carré intégrable. Alors XY est intégrable et l'on a :

$$\mathbb{E}[XY]^2 \leqslant \mathbb{E}[X^2]\mathbb{E}[Y^2]\,.$$

Démonstration. La preuve est rigoureusement la même que dans le cas discret (voir la proposition 19 de la page 733). ■

5.4 Somme de variables aléatoires à densité indépendantes

Proposition 22. Soient X_1 et X_2 deux v.a. réelles à densité p_1 et p_2 indépendantes. Alors $X = X_1 + X_2$ est une v.a. réelle à densité $p = p_1 * p_2$ donnée par :

$$p(x) = \int_{\mathbb{R}} p_1(y)p_2(x-y)\,\mathrm{d}y\,.$$

Le signe $*$ est conventionnel et désigne le produit de convolution.

Démonstration. Par indépendance, le vecteur aléatoire (X_1,X_2) est à densité $p_X(x_1,x_2)$ donnée par $p_1(x_1)p_2(x_2)$. On va calculer la fonction de répartition de X. Soit $x \in \mathbb{R}$:

$$\mathbb{P}(X \leqslant x) = \mathbb{P}(X_1+X_2 \leqslant x) = \int_{\mathbb{R}^2} \mathbf{1}_D(x_1,x_2)p_1(x_1)p_2(x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1\,,$$

où D est le domaine des points (x_1,x_2) du plan $\mathbb{R}^2$ tels que $x_1+x_2 \leqslant x$.

On peut donc écrire :

$$\mathbb{P}(X \leqslant x) = \int_{x_1=-\infty}^{+\infty} \int_{x_2=-\infty}^{x-x_1} p_1(x_1)p_2(x_2)\,\mathrm{d}x_2\,\mathrm{d}x_1\,.$$

En posant $x_2' = x_1 + x_2$:

$$\begin{aligned}
\mathbb{P}(X \leqslant x) &= \int_{x_1=-\infty}^{+\infty} \int_{x_2'=-\infty}^{x} p_1(x_1)p_2(x_2'-x_1)\,\mathrm{d}x_2'\,\mathrm{d}x_1 \\
&\overset{\text{Fubini}}{=} \int_{x_2'=-\infty}^{x} \left[\int_{x_1=-\infty}^{+\infty} p_1(x_1)p_2(x_2'-x_1)\,\mathrm{d}x_1\right]\mathrm{d}x_2' \\
&= \int_{x_2'=-\infty}^{x} p(x_2')\,\mathrm{d}x_2'\,,
\end{aligned}$$

avec $p(x) = p_1 * p_2(x)$. Comme la fonction de répartition caractérise la loi, on a le résultat cherché. ■

Proposition 23.
Soient $(X_i)_{i\in[\![1,n]\!]}$ des v.a. gaussiennes indépendantes de moyennes μ_i et de variances σ_i^2. La v.a. $X := \sum_{i=1}^{n} X_i$ suit une loi gaussienne de moyenne $\mu = \sum_{i=1}^{n} \mu_i$ et de variance $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$.

La moyenne de X est forcément μ, par linéarité de l'espérance. La variance de X est σ^2, d'après la proposition 20. Le résultat original est ici que la loi d'une somme de v.a. gaussiennes indépendantes est encore gaussienne.

Démonstration. On suppose que les σ_i sont tous non nuls. On commence par le cas où $n = 2$ et $\mu_1 = \mu_2 = 0$. La loi de $X = X_1 + X_2$ est une loi à densité p donnée par :

$$p_X(x) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2\sigma_1^2} - \frac{(x-y)^2}{2\sigma_2^2}}\,\mathrm{d}y\,.$$

Le terme dans l'exponentielle s'écrit aussi :

$$\frac{y^2}{\sigma_1^2} + \frac{(x-y)^2}{\sigma_2^2} = \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}y^2 - \frac{2x}{\sigma_2^2}y + \frac{x^2}{\sigma_2^2} = \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}\left(y - \frac{x}{1+\frac{\sigma_2^2}{\sigma_1^2}}\right)^2 + \frac{x^2}{\sigma_1^2+\sigma_2^2}\,.$$

L'intégrale se calcule alors par un changement de variable affine :

$$\begin{aligned}
p_X(x) &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left(-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(y - \frac{x}{1+\frac{\sigma_2^2}{\sigma_1^2}}\right)^2\right)\mathrm{d}y\, e^{-\frac{x^2}{2(\sigma_1^2+\sigma_2^2)}} \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\, e^{-\frac{x^2}{2(\sigma_1^2+\sigma_2^2)}} = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}\, e^{-\frac{x^2}{2(\sigma_1^2+\sigma_2^2)}}\,.
\end{aligned}$$

On reconnaît la densité de la loi gaussienne de moyenne nulle et de variance $\sigma_1^2 + \sigma_2^2$.

Dans le cas $n = 2$ et μ_1 et μ_2 sont quelconques, alors $\tilde{X}_1 := X_1 - \mu_1$ et $\tilde{X}_2 := X_1 - \mu_2$ sont des v.a. gaussiennes indépendantes de moyennes nulles et de variances σ_1^2 et σ_2^2. Donc $\tilde{X} = \tilde{X}_1 + \tilde{X}_2$ est une v.a. gaussienne à moyenne nulle et de variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$, et $X = \tilde{X} + \mu_1 + \mu_2$ est une v.a. gaussienne à moyenne $\mu_1 + \mu_2$ et de variance σ^2.
Le cas n quelconque se montre par récurrence sur n. Supposons que la propriété soit vraie jusqu'au rang $n - 1$, et considérons $(X_i)_{i \in [\![1,n]\!]}$ des v.a. gaussiennes indépendantes de moyennes μ_i et de variances σ_i^2. Alors, par l'hypothèse de récurrence, $\tilde{X} := \sum_{i=1}^{n-1} X_i$ est une v.a. gaussienne de moyenne $\tilde{\mu} = \sum_{i=1}^{n-1} \mu_i$ et de variance $\tilde{\sigma}^2 = \sum_{i=1}^{n-1} \sigma_i^2$. Cette v.a. est indépendante de X_n, car, pour tous intervalles $I, J \subset \mathbb{R}$:

$$\begin{aligned}
\mathbb{P}(\tilde{X} \in I, X_n \in J) &= \mathbb{E}[\mathbf{1}_I(\tilde{X})\mathbf{1}_J(X_n)] \\
&= \int_{\mathbb{R}^n} \mathbf{1}_I(x_1 + \cdots + x_{n-1})\mathbf{1}_J(x_n) \prod_{i=1}^{n} p_{X_i}(x_i)\,\mathrm{d}x_1 \ldots \mathrm{d}x_{n-1}\,\mathrm{d}x_n \\
&= \int_{\mathbb{R}^{n-1}} \mathbf{1}_I(x_1 + \cdots + x_{n-1}) \prod_{i=1}^{n-1} p_{X_i}(x_i)\,\mathrm{d}x_1 \ldots \mathrm{d}x_{n-1} \\
&\qquad \times \int_{\mathbb{R}} \mathbf{1}_J(x_n) p_{X_n}(x_n)\,\mathrm{d}x_n \\
&= \mathbb{E}[\mathbf{1}_I(\tilde{X})]\mathbb{E}[\mathbf{1}_J(X_n)] = \mathbb{P}(\tilde{X} \in I)\mathbb{P}(X_n \in J)\,.
\end{aligned}$$

On peut donc appliquer le résultat du cas $n = 2$ pour obtenir le résultat de la proposition.
Si certains des σ_i sont nuls, alors le résultat est encore vrai, car les v.a. X_i correspondantes sont des constantes $\mathbb{P}(X_i = \mu_i) = 1$, et la somme d'une v.a. gaussienne de moyenne μ et de variance σ^2 et d'une constante μ' est une v.a. gaussienne de moyenne $\mu + \mu'$ et de variance σ^2. ■

5.5 Vecteurs gaussiens

Définition 14. Soit d un entier strictement positif. On dit que le vecteur aléatoire $X = (X_i)_{i \in [\![1,d]\!]}$ est un *vecteur gaussien* si c'est la transformée affine d'un vecteur aléatoire constitué de variables aléatoires gaussiennes centrées réduites indépendantes.

En d'autres mots, il existe un entier n, un vecteur $\mu = (\mu_i)_{i \in [\![1,d]\!]}$ de $\mathbb{R}^d$, une matrice réelle A de taille $d \times n$ et n v.a. réelles Z_j indépendantes et identiquement distribuées selon la loi gaussienne centrée réduite tels que $X = \mu + AZ$. La proposition qui suit est une conséquence immédiate de la définition.

Proposition 24.
Toute transformation affine d'un vecteur gaussien est un vecteur gaussien.

Démonstration. Il suffit de noter que les transformations affines forment un groupe. ■

Proposition 25. Les lois marginales uni-dimensionnelles d'un vecteur gaussien sont des lois gaussiennes.

Démonstration. Soit $X := (X_i)_{i\in[\![1,d]\!]}$ un vecteur gaussien d-dimensionnel. Par définition, il existe un entier n, un vecteur $\mu := (\mu_i)_{i\in[\![1,d]\!]}$ de $\mathbb{R}^d$, une matrice réelle A de taille $d\times n$ et n v.a. réelles Z_j indépendantes et identiquement distribuées selon la loi gaussienne centrée réduite tels que $X = \mu + AZ$. La loi de X_i est celle de $\sum_{j=0}^{n} Y_j$ avec $Y_0 := \mu_i$ et $Y_j := A_{ij}Z_j$, $j \in [\![1,n]\!]$. Les Y_j sont des v.a. gaussiennes indépendantes, la proposition 23 de la page 779 donne alors le résultat. ∎

On a vu dans la proposition 19 de la page 777 que la matrice de covariance d'un vecteur aléatoire est une matrice symétrique positive. La proposition suivante montre que toute matrice symétrique positive est la matrice de covariance d'un vecteur aléatoire (gaussien).

Proposition 26. Si $\mu := (\mu_i)_{i\in[\![1,d]\!]}$ est un vecteur de $\mathbb{R}^d$ et $C := (C_{ij})_{i,j\in[\![1,d]\!]}$ est une matrice symétrique positive, alors il existe un vecteur gaussien dont la moyenne est μ et la matrice de covariance est C.

Démonstration. Si C est une matrice symétrique positive, il existe une matrice orthogonale U et des réels positifs $(\lambda_i)_{i\in[\![1,d]\!]}$ tels que $C = U^tDU$ où D est la matrice diagonale de coefficients diagonaux λ_i. Si l'on introduit la matrice diagonale D_2 de coefficients diagonaux $\lambda_i^{1/2}$ et la matrice symétrique $A = U^tD_2U$, alors on vérifie que :

$$AA^t = U^tD_2UU^tD_2U = U^tD_2^2U = U^tDU = C\,.$$

Si $Z := (Z_i)_{i\in[\![1,d]\!]}$ est une famille de v.a. gaussiennes centrées réduites indépendantes, alors le vecteur aléatoire $AZ+\mu$ est un vecteur gaussien, de moyenne μ et de covariance C par la proposition 19 de la page 777. ∎

Proposition 27.
Soit $X = AZ + \mu$ un vecteur gaussien, avec $\mu \in \mathbb{R}^d$, $A = (A_{ij})_{i,j\in[\![1,d]\!]}$ inversible, et $(Z_i)_{i\in[\![1,d]\!]}$ des v.a. gaussiennes centrées réduites indépendantes. La loi de X a pour densité $p_{\mu,C} : \mathbb{R}^d \to]0,+\infty[$ donnée par :

$$p_{\mu,C}(x) = \frac{1}{\sqrt{(2\pi)^d \det C}} \exp\left(-\frac{(x-\mu)^tC^{-1}(x-\mu)}{2}\right),$$

où $C := AA^t$ est la matrice de covariance de X. Cette loi est appelée loi gaussienne d-dimensionnelle de moyenne μ et de covariance C.

Démonstration. La matrice de covariance de X est $C = AA^t$. Comme A est inversible, l'application $z \in \mathbb{R}^d \mapsto Az + \mu$ est une bijection linéaire. Soient I_i des intervalles de $\mathbb{R}$, et $\chi : \mathbb{R}^d \to \{0,1\}$ la fonction indicatrice $\chi(x) := \mathbf{1}_{I_1}(x_1)\cdots\mathbf{1}_{I_d}(x_d)$.

On évalue la probabilité :

$$\mathbb{P}(X_1 \in I_1, \dots, X_d \in I_d) = \mathbb{E}[\chi(X)] = \mathbb{E}[\chi(AZ+\mu)] = \int_{\mathbb{R}^d} \chi(Az+\mu) p_Z(z)\,\mathrm{d}z\,,$$

où :

$$p_Z(z) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|z\|^2}{2}\right) = g(z_1)\dots g(z_d)\,,$$

et g est la densité de la loi gaussienne centrée réduite (11). On effectue le changement de variable $x := Az + \mu$:

$$\begin{aligned}\mathbb{P}(X_1 \in I_1, \dots, X_d \in I_d) &= \det(A)^{-1} \int_{\mathbb{R}^d} \chi(x) p_Z(A^{-1}(x-\mu))\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \chi(x) p_{\mu,C}(x)\,\mathrm{d}x = \int_{I_1\times\dots\times I_d} p_{\mu,C}(x)\,\mathrm{d}x\,,\end{aligned}$$

ce qui permet d'identifier la loi de $X = (X_1, \dots, X_d)$. ■

Comme on le verra dans le module IV.4, la somme des carrés des coordonnées d'un vecteur gaussien constitué de n v.a. gaussiennes centrées réduites indépendantes intervient beaucoup en statistique. On donne un nom à la loi de cette somme : la loi du χ^2_n (prononcer : la loi du khi-deux à n degrés de libertés).

Proposition 28. Soient n un entier strictement positif et $(X_i)_{i\in[\![1,n]\!]}$ une famille de v.a. gaussiennes centrées réduites indépendantes. Alors la v.a. réelle :

$$X := \sum_{i=1}^n X_i^2$$

admet une densité donnée par :

$$p_X(x) = c_n x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right) \mathbf{1}_{]0,+\infty[}(x)\,, \qquad \text{avec } c_n = \frac{1}{2^{n/2}\Gamma(n/2)}\,.$$

De plus, la v.a. réelle X et le vecteur $Y := (Y_i)_{i\in[\![1,n]\!]}$, où $Y_i := X_i/\sqrt{X}$ si $X > 0$ et $Y_i := 0$ si $X = 0$, sont indépendants.

La fonction Γ (fonction Gamma d'Euler) est une fonction tabulée célèbre définie par :

$$\Gamma(x) := \int_0^{+\infty} y^{x-1} e^{-y}\,\mathrm{d}y\,, \quad x \in]0,+\infty[\,.$$

Démonstration. On calcule la fonction de répartition de X. Soit $x < 0$. Comme X est une somme de carrés, on a $\mathbb{P}(X \leqslant x) = 0$. Soit $x \geqslant 0$. On a :

$$\mathbb{P}(X \leqslant x) = \mathbb{P}(X_1^2 + \dots + X_n^2 \leqslant x) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathbf{1}_{]-\infty,x]}(\|y\|^2) \exp\left(-\frac{\|y\|^2}{2}\right)\mathrm{d}y\,.$$

On passe en coordonnées sphériques sur $\mathbb{R}^n$:

$$y_1 = R\cos\theta_1\,, \quad y_2 = R\sin\theta_1\cos\theta_2\,, \quad y_3 = R\sin\theta_1\sin\theta_2\cos\theta_3\,, \ \dots$$

$$\dots\,, \quad y_{n-1} = R\sin\theta_1\cdots\sin\theta_{n-2}\cos\theta_{n-1}\,, \quad y_n = R\sin\theta_1\cdots\sin\theta_{n-1}\,,$$

avec $R \in [0, +\infty[$, $\theta_1, \ldots, \theta_{n-2} \in [0, \pi]$, et $\theta_{n-1} \in [0, 2\pi[$. Le jacobien de cette transformation est :

$$J(R, \theta_1, \ldots, \theta_{n-1}) = R^{n-1} \prod_{i=1}^{n-2} \big(\sin \theta_{n-i-1}\big)^i .$$

Cette expression du jacobien s'obtient par récurrence sur n (laissé en exercice).
On obtient alors :

$$\begin{aligned}
\mathbb{P}(X \leqslant x) &= 2c_n \int_0^{+\infty} \mathbf{1}_{]-\infty,x]}(R^2) \exp\left(-\frac{R^2}{2}\right) R^{n-1} \,\mathrm{d}R \\
&\overset{r=R^2}{=} c_n \int_0^{+\infty} \mathbf{1}_{]-\infty,x]}(r) \exp\left(-\frac{r}{2}\right) r^{\frac{n}{2}-1} \,\mathrm{d}r ,
\end{aligned}$$

$$\text{avec} \qquad c_n = \frac{1}{2(2\pi)^{n/2}} \int_0^{2\pi} \mathrm{d}\theta_{n-1} \int_0^{\pi} \mathrm{d}\theta_{n-2} \cdots \int_0^{\pi} \mathrm{d}\theta_1 \prod_{i=1}^{n-2} \big(\sin \theta_{n-i-1}\big)^i .$$

Ce calcul permet d'identifier la densité p_X de X. On peut déterminer la constante c_n par cette formulation intégrale (ce sont des intégrales dites de Wallis). On peut plus simplement se servir du fait que l'on sait que la constante de normalisation c_n est telle que l'intégrale sur $\mathbb{R}$ de p_X donne 1 :

$$c_n \int_0^{+\infty} y^{\frac{n}{2}-1} \exp\left(-\frac{y}{2}\right) \mathrm{d}y = 1 .$$

On obtient l'expression de c_n en termes de la fonction Γ en effectuant le changement de variables $x := y/2$. Pour prouver l'indépendance de X et de Y, prenons un intervalle $I \subset \mathbb{R}$ et un pavé $J \subset \mathbb{R}^n$ et considérons :

$$\begin{aligned}
\mathbb{P}(X \in I, Y \in J) = \mathbb{E}[\mathbf{1}_I(X)\mathbf{1}_J(Y)] &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{-\frac{x_1^2 + \cdots + x_n^2}{2}} \\
&\quad \times \mathbf{1}_I(x_1^2 + \cdots + x_n^2) \mathbf{1}_J\left(\frac{x_1}{\sqrt{x_1^2 + \cdots + x_n^2}}, \ldots, \frac{x_n}{\sqrt{x_1^2 + \cdots + x_n^2}}\right) \mathrm{d}x .
\end{aligned}$$

En passant en coordonnées sphériques :

$$\begin{aligned}
\mathbb{P}(X \in I, Y \in J) &= \frac{1}{(2\pi)^{n/2}} \int_0^{+\infty} \mathrm{d}R R^{n-1} \int_0^{2\pi} \mathrm{d}\theta_{n-1} \int_0^{\pi} \mathrm{d}\theta_{n-2} \cdots \int_0^{\pi} \mathrm{d}\theta_1 \\
&\quad \times \prod_{i=1}^{n-2} (\sin \theta_{n-i-1})^i e^{-\frac{R^2}{2}} \mathbf{1}_I(R^2) \mathbf{1}_J(\sin \theta_1, \ldots, \cos \theta_1 \cdots \cos \theta_{n-1}) \\
&= \left[c_n \int_0^{+\infty} r^{\frac{n}{2}-1} e^{-\frac{r}{2}} \mathbf{1}_I(r) \,\mathrm{d}r\right] \left[\frac{1}{2(2\pi)^{n/2} c_n} \int_0^{2\pi} \mathrm{d}\theta_{n-1}\right. \\
&\quad \left. \times \int_0^{\pi} \mathrm{d}\theta_{n-2} \cdots \int_0^{\pi} \mathrm{d}\theta_1 \prod_{i=1}^{n-2} (\sin \theta_{n-i-1})^i \mathbf{1}_J(\sin \theta_1, \ldots, \cos \theta_1 \cdots \cos \theta_{n-1})\right] \\
&= \mathbb{E}[\mathbf{1}_I(X)]\mathbb{E}[\mathbf{1}_J(Y)] = \mathbb{P}(X \in I)\mathbb{P}(Y \in J) ,
\end{aligned}$$

ce qui permet aussi d'identifier la loi de Y. ∎

Exercice 4.

Montrer qu'une v.a. suivant la loi χ^2_n a pour espérance n et pour variance $2n$.
Montrer que, si $X \sim \chi^2_n$ et $Y \sim \chi^2_p$ sont indépendantes, alors $X + Y \sim \chi^2_{n+p}$.

Solution. Soit $X \sim \chi^2_n$. Pour calculer les moments de X il y a beaucoup plus simple que de se lancer dans le calcul d'intégrales avec la densité de la loi. Il faut revenir à la définition. X a même loi que $\sum_{j=1}^{n} X_j^2$, où les X_j sont indépendantes et identiquement distribuées selon la loi gaussienne centrée réduite. Donc l'espérance de X est :

$$\mathbb{E}[X] = \sum_{j=1}^{n} \mathbb{E}[X_j^2] = n\,,$$

et le second moment de X est :

$$\mathbb{E}[X^2] = \sum_{i,j=1}^{n} \mathbb{E}[X_i^2 X_j^2] = \sum_{i=1}^{n} \mathbb{E}[X_i^4] + \sum_{i \neq j} \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] = n\mathbb{E}[X_1^4] + n(n-1)\mathbb{E}[X_1^2]^2\,.$$

On sait que $\mathbb{E}[X_1^2] = 1$, et l'on calcule par intégration par parties :

$$\mathbb{E}[X_1^4] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^4 e^{-\frac{x^2}{2}}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \left[-x^3 e^{-\frac{x^2}{2}}\right]_{-\infty}^{+\infty} + \frac{3}{\sqrt{2\pi}} \int_{\mathbb{R}} x^2 e^{-\frac{x^2}{2}}\,\mathrm{d}x = 3\,.$$

On obtient donc $\mathbb{E}[X^2] = 3n + n(n-1) = n^2 + 2n$, et :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = n^2 + 2n - n^2 = 2n\,.$$

La seconde question est immédiate avec la représentation en sommes de carrés de v.a. gaussiennes indépendantes.

Exercice 5.

Montrer que la constante de normalisation de la loi du χ^2_n vaut explicitement :

$$\begin{cases} c_{2k} = 1/[2^k(k-1)!] & \text{si } n = 2k \text{ est pair} \\ c_{2k+1} = 2^k k!/[(2\pi)^{1/2}(2k)!] & \text{si } n = 2k+1 \text{ est impair.} \end{cases}$$

Solution. La fonction Γ s'évalue simplement sur les entiers et les demi-entiers. D'une part, on montre par intégration par parties que :

$$\Gamma(x) = (x-1)\Gamma(x-1)\,, \text{ pour tout } x \in \,]0, +\infty[\,.$$

Ensuite, on calcule que $\Gamma(1) = 1$, et donc $\Gamma(x) = (x-1)!$ pour tout $x \in \mathbb{N}^*$. Cela donne l'expression de c_n lorsque n est pair. Enfin, on calcule par le changement de variables $y = t^2$ que $\Gamma(1/2) = \sqrt{\pi}$, et donc :

$$\Gamma\Big(x + \frac{1}{2}\Big) = \sqrt{\pi} \prod_{i=1}^{x} \Big(i - \frac{1}{2}\Big) = \frac{\sqrt{\pi}(2x)!}{2^{2x} x!}\,,$$

pour tout $x \in \mathbb{N}$. Cela donne l'expression de c_n lorsque n est impair.

Proposition 29. Si $X \sim \mathcal{N}(0,1)$ et $Y \sim \chi_n^2$ sont indépendantes, alors $Z = \sqrt{n}\frac{X}{\sqrt{Y}}$ suit une loi dite de Student à n degrés de libertés, notée T_n. Sa densité est :

$$p_{T_n}(z) = d_n \left(1 + \frac{z^2}{n}\right)^{-\frac{n+1}{2}}, \text{ avec } d_n = \frac{\Gamma(\frac{n+1}{2})}{(n\pi)^{\frac{1}{2}}\Gamma(\frac{n}{2})}.$$

Il est facile de voir que la densité de la loi T_n tend quand n tend vers $+\infty$ vers la densité gaussienne, car :

$$\left(1 + \frac{z^2}{n}\right)^{-\frac{n+1}{2}} \xrightarrow{n\to+\infty} \exp\left(-\frac{z^2}{2}\right).$$

On peut aussi montrer, avec quelques calculs supplémentaires, que la variance de T_n diminue avec n, à partir de $n = 4$ (une v.a. de loi T_n n'est pas de carré intégrable si $n \leqslant 3$). Les lois χ_n^2 et T_n sont tabulées dans tous les ouvrages classiques, leurs densités et fonctions de répartition sont aussi préprogrammées dans les outils de calcul scientifique et statistique usuel (matlab, ...). La loi de Student est importante en statistique, comme on le verra dans le module IV.4.

Démonstration. On peut voir cette démonstration comme un exercice sur les v.a. à densité. Soit I un intervalle.

$$\begin{aligned}
\mathbb{P}(Z \in I) = \mathbb{E}\Big[\mathbf{1}_I\Big(\sqrt{n}\frac{X}{\sqrt{Y}}\Big)\Big] &= \frac{c_n}{\sqrt{2\pi}}\int_0^{+\infty}\mathrm{d}y\int_{-\infty}^{+\infty}\mathrm{d}x\, y^{\frac{n}{2}-1}e^{-\frac{y}{2}}e^{-\frac{x^2}{2}}\mathbf{1}_I\Big(\sqrt{n}\frac{x}{\sqrt{y}}\Big) \\
&= \frac{2c_n}{\sqrt{2\pi}}\int_0^{+\infty}\mathrm{d}t\int_{-\infty}^{+\infty}\mathrm{d}x\, t^{n-1}e^{-\frac{t^2+x^2}{2}}\mathbf{1}_I\Big(\sqrt{n}\frac{x}{t}\Big) \quad \text{avec } y \mapsto t := \sqrt{y} \\
&= \frac{2c_n}{\sqrt{2\pi n}}\int_0^{+\infty}\mathrm{d}t\int_{-\infty}^{+\infty}\mathrm{d}z\, t^n e^{-\frac{t^2+t^2z^2/n}{2}}\mathbf{1}_I(z) \quad \text{avec } x \mapsto z := \sqrt{n}\frac{x}{t} \\
&= \frac{2c_n}{\sqrt{2\pi n}}\int_{-\infty}^{+\infty}\mathrm{d}z\mathbf{1}_I(z)\int_0^{+\infty}\mathrm{d}t t^n e^{-\frac{t^2}{2}\left(1+\frac{z^2}{n}\right)} \quad \text{par Fubini} \\
&= \frac{c_n}{\sqrt{2\pi n}}\int_I \mathrm{d}z\Big(1+\frac{z^2}{n}\Big)^{-\frac{n+1}{2}}\int_0^{+\infty}\mathrm{d}u u^{\frac{n-1}{2}}e^{-\frac{u}{2}} \quad \text{avec } t \mapsto u := t^2\Big(1+\frac{z^2}{n}\Big) \\
&= \frac{c_n}{\sqrt{2\pi n}c_{n+1}}\int_I \mathrm{d}z\Big(1+\frac{z^2}{n}\Big)^{-\frac{n+1}{2}},
\end{aligned}$$

où c_{n+1} est la constante de normalisation de la loi χ_{n+1}^2.
Ce calcul permet d'identifier la loi de Z et sa densité. ■

Exercice 6.

Montrer que la constante de normalisation de la loi T_n vaut explicitement

$$\begin{cases} d_{2k} = (2k)!k^{1/2}/[2^{2k+1/2}(k!)^2] & n = 2k \text{ est pair} \\ d_{2k+1} = 2^{2k}k!^2/[\pi(2k+1)^{1/2}(2k)!] & n = 2k+1 \text{ est impair.} \end{cases}$$

Solution. On a $d_n = \frac{c_n}{\sqrt{2\pi n}c_{n+1}}$ où c_n est la constante de normalisation de la loi χ_n^2. On prend les formules obtenues dans l'exercice 5 de la page 784 pour conclure.

6 Simulation de variables aléatoires

Dans certaines expériences « réelles », où le hasard intervient, il peut être intéressant de réaliser des expériences numériques où on simule les v.a. qui interviennent, plutôt que de réaliser physiquement les expériences en question. Ainsi, en physique nucléaire, les désintégrations des atomes fissibles sont décrits par des variables aléatoires (de loi exponentielle, de loi de Poisson...), et l'on peut chercher à simuler ces processus avant de construire des réacteurs.

Cette section se décompose en deux parties. On va d'abord présenter les principes d'un générateur de nombres aléatoires suivant une loi uniforme sur $[0,1]$. On va ensuite montrer qu'à partir d'un tel générateur, on peut simuler des variables aléatoires de loi quelconque.

6.1 Simulation d'une v.a. uniforme

On veut implémenter dans un ordinateur un processus (un code) qui génère une suite de nombres aléatoires suivant une loi uniforme sur $[0,1]$. Le problème est paradoxal : un ordinateur est une machine parfaitement déterministe, donc on ne voit pas comment on pourrait générer des nombres de manière aléatoire. En effet, on ne le peut pas, mais on peut générer une suite de nombres pseudo-aléatoires, qui vont mimer le hasard. Pour voir comment ces suites sont construites, on peut par exemple aller voir dans le logiciel de calcul scientifique SCILAB (qui est dans le domaine public) et ouvrir la boîte noire qu'est la fonction `rand`. La fonction `rand` fournit une suite de nombres censée être une réalisation d'une suite de v.a. uniformes sur $[0,1]$ et indépendantes. En fait, cette suite est construite sur une suite d'entiers Y_n définie par la relation de récurrence :

$$Y_{n+1} := aY_n + c \mod m$$

où $m = 2^{31}$, $a = 843314861$ et $c = 453816693$. Le terme initial de la suite est par défaut $Y_0 = 0$, mais on peut le changer (pour obtenir une autre « réalisation » de la suite). On se ramène ensuite à des nombres entre 0 et 1 en divisant Y_n par m, c'est-à-dire que l'on pose : $X_n := Y_n 2^{-31}$.

Notez que la suite (X_n) ainsi construite n'a rien d'aléatoire, elle est même parfaitement déterministe ! De plus, les valeurs possibles prises par la suite ne sont pas l'ensemble des réels entre 0 et 1, mais l'ensemble fini $\{k/m \mid k = 0, \ldots, m-1\}$. Enfin, la suite est périodique, de période m.

Cependant, l'ensemble fini $\{k/m \mid k = 0, \ldots, m-1\}$ est un quasi-continuum, la période m est très grande, et finalement ces suites parfaitement déterministes apparaissent comme aléatoires pour de bons choix de triplets (m, a, c), dans le sens où des tests statistiques d'uniformité (notion que nous ne verrons pas ici) ne font pas la différence. Cette méthode de simulation n'a rien de particulier à SCILAB, la grande majorité des générateurs de nombres aléatoires sur ordinateur sont basés sur ce modèle (dit de congruence linéaire), avec des choix particuliers de nombres entiers m, a et c. Par exemple, la méthode `Math.random` de JAVA utilise les valeurs $m = 2^{48}$, $a = 25214903917$ et $b = 11$. La fonction `rand` de MATLAB ou `random.rand` de la bibliothèque Numpy en PYTHON est basée sur le générateur de Mersenne Twister de période colossale $2^{19937} - 1$.

La mise au point de générateurs de nombres pseudo-aléatoires fait toujours l'objet de re-

cherches intensives. Dans ce cours, on va admettre que l'on sait simuler des suites de v.a. indépendantes et uniformes sur $[0,1]$, et l'on va montrer qu'alors, on peut simuler des v.a. suivant n'importe quelle loi donnée.

6.2 Méthode de la fonction inverse

La méthode de la fonction inverse est basée sur le résultat suivant.

Proposition 30. Soient X une v.a. réelle de fonction de répartition F et U une v.a. uniforme sur $[0,1]$. Si F est bijective, alors X et $F^{-1}(U)$ ont même loi.

Rappelons qu'une fonction de répartition est continue à droite et croissante. Pour que la fonction F soit bijective, il faut et il suffit qu'elle soit continue et strictement croissante. Cela est assuré en particulier si X est une v.a. à densité strictement positive sur $\mathbb{R}$.

Démonstration. On vérifie tout d'abord que $Y := F^{-1}(U)$ est une v.a. réelle bien définie, car U est à valeurs dans $]0,1[$ avec probabilité 1, et $F^{-1} :]0,1[\to \mathbb{R}$. Pour tout $x \in \mathbb{R}$, pour tout $u \in]0,1[$, on a $F^{-1}(u) \leqslant x$ si, et seulement si, $u \leqslant F(x)$. Par conséquent :

$$\mathbb{P}(Y \leqslant x) = \mathbb{P}(F^{-1}(U) \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x)\,,$$

ce qui montre que Y a pour fonction de répartition F. ∎

On peut en fait étendre la proposition 30 au cas où F n'est pas bijective, en introduisant le concept d'inverse généralisé. Ce concept est spécifique aux fonctions de répartition, et ne s'applique pas à des fonctions arbitraires.

Définition 15. Soit F une fonction de répartition. L'*inverse généralisé* de F est la fonction $F^{-1} :]0,1[\to \mathbb{R}$ définie par :

$$F^{-1}(u) := \inf A_u, \quad \text{où } A_u := \{x \in \mathbb{R} \mid F(x) \geqslant u\}\,.$$

La définition est parfaitement sensée car, pour tout $u \in]0,1[$:
(i) $\lim\limits_{x\to+\infty} F(x) = 1$, donc $A_u \neq \varnothing$.
(ii) $\lim\limits_{x\to-\infty} F(x) = 0$, donc A_u est minoré.
Dans le cas où F est bijective, l'inverse généralisé correspond à la fonction réciproque.

Proposition 31. Soient X une v.a. réelle de fonction de répartition F et U une v.a. uniforme sur $[0,1]$. Alors X et $F^{-1}(U)$ ont même loi.

Démonstration. Il suffit de montrer que, pour tout $u \in]0,1[$, pour tout $x \in \mathbb{R}$:

$$F^{-1}(u) \leqslant x \Longleftrightarrow u \leqslant F(x)\,,$$

car alors la démonstration de la proposition 30 s'applique directement.
Supposons que $u \leqslant F(x)$. Alors $x \in A_u$, donc $F^{-1}(u) \leqslant x$.
Supposons que $F^{-1}(u) \leqslant x$.
On peut alors affirmer que, pour tout $n \in \mathbb{N}^*$, on a $F^{-1}(u) < x + 1/n$.

Comme $F^{-1}(u) = \inf A_u$, il existe $t_n \in A_u$ tel que $F^{-1}(u) \leqslant t_n < x + 1/n$. Comme $t_n \in A_u$, on a $F(t_n) \geqslant u$. Comme F est croissante, $F(x + 1/n) \geqslant F(t_n) \geqslant u$. Comme F est continue à droite, on passe à la limite lorsque n tend vers $+\infty$ et l'on trouve que $F(x) \geqslant u$. ■

Exemple. On souhaite simuler une v.a. exponentielle de paramètre $\lambda > 0$. Sa fonction de répartition est donnée par (10), dont la fonction réciproque est :

$$F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u).$$

Donc, si U est une v.a. uniforme sur $[0,1]$, la v.a. $X := -\ln(1-U)/\lambda$ a pour loi la loi exponentielle de paramètre λ. De plus, U et $1-U$ ont la même loi, donc on peut aussi dire que $-\ln(U)/\lambda$ a pour loi la loi exponentielle de paramètre λ.

6.3 Simulation d'une v. a. discrète

Proposition 32.
Soient $E = \{x_1, x_2, \ldots\} \subset \mathbb{R}$ un ensemble discret, avec $x_1 < x_2 < \cdots$, et P la probabilité sur $\big(E, \mathcal{P}(E)\big)$ donnée par $P(\{x_i\}) = p_i$, $i \in \mathbb{N}^*$. Soient U une v.a. uniforme sur $[0,1]$, et X la v.a. discrète à valeurs dans E définie par :

$$X = x_i \text{ si } \sum_{j<i} p_j < U \leqslant \sum_{j\leqslant i} p_j\,.$$

Alors la v.a. X a pour loi P.

Démonstration. Rappelons que la fonction de répartition F de la loi discrète P est constante par morceaux et donnée par la proposition 5 de la page 758. La proposition 32 est alors un corollaire de la proposition 31. En effet, l'inverse généralisé de la fonction de répartition F est précisément la fonction proposée. On peut aussi vérifier directement que, pour tout i :

$$\mathbb{P}(X = x_i) = \mathbb{P}\Big(\sum_{j<i} p_j < U \leqslant \sum_{j\leqslant i} p_j\Big) = \sum_{j\leqslant i} p_j - \sum_{j<i} p_j = p_i\,.$$ ■

Exemple. On souhaite simuler une v.a. de Bernoulli de paramètre p. Cette v.a. est à valeurs dans $\{0,1\}$ et sa loi est donnée par $P(\{0\}) = 1-p$ et $P(\{1\}) = p$. La proposition 32 ci-dessus donne la méthode suivante :

- on tire U selon une loi uniforme sur $[0,1]$,
- on pose $X := 0$ si $U \leqslant p$ et $X := 1$ si $U > p$.

On peut vérifier directement que cet algorithme fournit effectivement une v.a. avec la loi souhaitée. En effet, la v.a. X proposée est simplement $X = \mathbf{1}_{U\leqslant p}$. Comme cette v.a. prend pour valeurs 0 ou 1, il s'agit bien d'une v.a. de Bernoulli, et son paramètre est $\mathbb{P}(U \leqslant p) = p$.

6.4 Simulation d'une v. a. gaussienne

Les v.a. gaussiennes sont très importantes en pratique, mais malheureusement leurs fonctions de répartition n'ont pas d'expression analytique, ainsi que leurs réciproques, et l'on ne peut donc pas utiliser la méthode de la fonction inverse pour elles. La méthode la plus courante pour simuler une v.a. gaussienne est l'algorithme de Box-Muller basé sur la proposition suivante.

Proposition 33. Soient U_1 et U_2 deux v.a. indépendantes de loi uniforme sur $[0,1]$. On pose :

$$X = (-2\ln U_1)^{1/2}\cos(2\pi U_2)\,,$$
$$Y = (-2\ln U_1)^{1/2}\sin(2\pi U_2)\,.$$

Les v.a. X et Y sont indépendantes et suivent des lois $\mathcal{N}(0,1)$.

Démonstration. Soient I et J deux intervalles de $\mathbb{R}$. Notons χ la fonction indicatrice de $I\times J$ sur $\mathbb{R}^2$, donnée par $\chi(x_1,x_2)=\mathbf{1}_I(x_1)\mathbf{1}_J(x_2)$, et calculons :

$$\begin{aligned}\mathbb{P}(X\in I, Y\in J) &= \mathbb{E}\Big[\chi\Big((-2\ln U_1)^{1/2}\cos(2\pi U_2), (-2\ln U_1)^{1/2}\sin(2\pi U_2)\Big)\Big]\\ &= \int_0^1\int_0^1 \chi\Big((-2\ln u_1)^{1/2}\cos(2\pi u_2), (-2\ln u_1)^{1/2}\sin(2\pi u_2)\Big)\,\mathrm{d}u_1\,\mathrm{d}u_2\,.\end{aligned}$$

On effectue le changement de variables $r=(-2\ln u_1)^{1/2}$ et $\theta = 2\pi u_2$:

$$\mathbb{P}(X\in I, Y\in J) = \frac{1}{2\pi}\int_0^{+\infty}\int_0^{2\pi}\chi\left(r\cos(\theta), r\sin(\theta)\right)e^{-r^2/2}r\,\mathrm{d}\theta\,\mathrm{d}r\,.$$

On passe ensuite en coordonnées cartésiennes $x=r\cos(\theta)$, $y=r\sin(\theta)$:

$$\begin{aligned}\mathbb{P}(X\in I, Y\in J) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\chi(x,y)\exp\left(-\frac{x^2}{2}-\frac{y^2}{2}\right)\mathrm{d}x\,\mathrm{d}y\\ &= \int_I g(x)\,\mathrm{d}x\int_J g(y)\,\mathrm{d}y\,,\end{aligned}$$

où $g(x) := 1/\sqrt{2\pi}\exp(-x^2/2)$ est la densité d'une v.a. gaussienne $\mathcal{N}(0,1)$. Comme cette égalité est vraie pour tout couple d'intervalles (I,J), on a bien caractérisé la loi du couple (X,Y). ■

6.5 Méthode du rejet pour une loi uniforme

On souhaite simuler une v.a. X de loi uniforme sur un domaine D de $\mathbb{R}^d$.

Définition 16. Soit D un borélien de $\mathbb{R}^d$ tel que $0<\lambda(D)<+\infty$. On dit que X suit une *loi uniforme* sur D si X a une loi à densité p constante sur D et nulle ailleurs :

$$p(x) = \begin{cases} c & \text{si } x\in D\,,\\ 0 & \text{sinon}\,,\end{cases}\qquad \text{où}\qquad c=\frac{1}{\lambda(D)}\cdot$$

La valeur de c est imposée par la condition $\displaystyle\int_{\mathbb{R}^d} p(x)\,\mathrm{d}x = 1$, qui impose $c\lambda(D) = 1$. On a alors, pour tout $A \in \mathcal{B}(\mathbb{R}^d)$:

$$\mathbb{P}(X \in A) = \int_A p(x)\,\mathrm{d}x = \frac{1}{\lambda(D)} \int_A \mathbf{1}_D(x)\,\mathrm{d}x = \frac{1}{\lambda(D)} \int_{D\cap A} \mathrm{d}x = \frac{\lambda(D\cap A)}{\lambda(D)}\cdot$$

Dans le cas où D est un pavé, il est très facile de simuler une v.a. de loi uniforme sur D.

Lemme 34. Soient $d \in \mathbb{N}^*$ et $a_i < b_i$, $i \in [\![1, d]\!]$.
On considère le pavé $D = \prod_{i=1}^{d} [a_i, b_i]$. Soit $(U_i)_{i\in[\![1,d]\!]}$ une suite de v.a. indépendantes de loi uniforme sur $[0, 1]$. On pose $X_i = a_i + (b_i - a_i)U_i$ pour tout $i \in [\![1, d]\!]$. Alors X_i est une v.a. réelle de loi uniforme sur $[a_i, b_i]$ pour tout $i \in [\![1, d]\!]$, les v.a. $(X_i)_{i\in[\![1,d]\!]}$ sont indépendantes, et $X := (X_1, \ldots, X_d)$ suit une loi uniforme sur D.

Démonstration. La fonction de répartition du vecteur $X = (X_i)_{i\in[\![1,d]\!]}$ est :

$$F_X(x_1, \ldots, x_d) = \mathbb{P}\left(U_1 \leqslant \frac{x_1 - a_1}{b_1 - a_1}, \ldots, U_d \leqslant \frac{x_d - a_d}{b_d - a_d}\right) = \prod_{i=1}^{d} F_U\left(\frac{x_i - a_i}{b_i - a_i}\right),$$

avec F_U la fonction de répartition de la loi uniforme sur $[0, 1]$, donnée par :

$$F_U(x) = \begin{cases} x & \text{si } x \in [0,1] \\ 0 & \text{si } x < 0 \\ 1 & \text{si } x > 1, \end{cases}$$

ce qui permet d'identifier la loi du vecteur X. C'est un vecteur à densité :

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p_i(x_i) = \prod_{i=1}^{d} \frac{1}{b_i - a_i} \mathbf{1}_{]a_i, b_i[}(x_i) = \frac{1}{\lambda(D)} \mathbf{1}_D(x_1, \ldots, x_d). \qquad \blacksquare$$

L'idée pour simuler une v.a. uniforme sur un domaine borélien borné D est la suivante :
(i) On entoure le domaine D par un pavé B.
(ii) On tire M_1, M_2, … indépendants et identiquement distribués selon la loi uniforme sur B (ce que l'on sait faire d'après le lemme 34), jusqu'au premier instant T où M_i tombe dans D.
(iii) Le point M_T est dans D. On va montrer que M_T suit la loi uniforme dans D.
Attention : la loi de M_T n'est pas la loi commune des M_i, car T est aléatoire !
On va en fait montrer un résultat plus général.

Proposition 35. Soit $(M_i)_{i\geqslant 1}$ une suite de v.a. indépendantes et de même loi μ sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Soit D un borélien de $\mathbb{R}^d$ tel que $\mu(D) > 0$. Pour tout $\omega \in \Omega$, on pose :

$$T(\omega) := \inf\{j \geqslant 1 \mid M_j(\omega) \in D\}\,,$$

avec la convention $\inf \varnothing = +\infty$, et l'on définit :

$$M_T(\omega) := \begin{cases} M_{T(\omega)}(\omega) & \text{si } T(\omega) < +\infty\,, \\ 0 & \text{si } T(\omega) = +\infty\,. \end{cases}$$

Alors :
(i) T suit la loi géométrique de paramètre $\mu(D)$.
(ii) M_T suit la loi :

$$\mathbb{P}(M_T \in A) = \frac{\mu(D \cap A)}{\mu(D)}\,, \qquad \forall A \in \mathcal{B}(\mathbb{R}^d)\,.$$

Application : Si μ est la loi uniforme sur un domaine borélien B tel que $D \subset B$, alors $\mu(A) = \lambda(A \cap B)/\lambda(B)$ et la loi de M_T est :

$$\mathbb{P}(M_T \in A) = \frac{\dfrac{\lambda(D \cap A \cap B)}{\lambda(B)}}{\dfrac{\lambda(D \cap B)}{\lambda(B)}} = \frac{\lambda(D \cap A)}{\lambda(D)}\,,$$

qui n'est autre que la loi uniforme sur D. Dans ce cas, l'espérance de T est donnée par :

$$\mathbb{E}[T] = \frac{1}{\mu(B)} = \frac{\lambda(B)}{\lambda(D \cap B)} = \frac{\lambda(B)}{\lambda(D)}\,.$$

Cela montre que, si B est beaucoup plus grand que D, alors il va falloir faire beaucoup de tirages avant d'en trouver un qui tombe dans D. On a donc toujours intérêt à chercher un pavé B englobant D aussi petit que possible pour minimiser le temps de calcul.

Démonstration. Soit $k \in \mathbb{N}^*$. L'événement $\{T = k\}$ se produit si, et seulement si, les $k-1$ premiers points M_j ne tombent pas dans D et si M_k tombe dans D. Par conséquent :

$$\begin{aligned}
\mathbb{P}(T = k) &= \mathbb{P}(M_j \notin D, j = 1, \ldots, k-1, M_k \in D) \\
&= \prod_{j=1}^{k-1} \mathbb{P}(M_j \notin D) \times \mathbb{P}(M_k \in D) = \mathbb{P}(M_1 \notin D)^{k-1}\mathbb{P}(M_k \in D) \\
&= [1 - \mu(D)]^{k-1}\mu(D)\,.
\end{aligned}$$

Donc :

$$\mathbb{P}(T = +\infty) = 1 - \sum_{k=1}^{+\infty} \mathbb{P}(T = k) = 1 - 1 = 0\,,$$

et T suit la loi géométrique de paramètre $\mu(D)$.

Soit $A \in \mathcal{B}(\mathbb{R}^d)$. En utilisant le fait que $\bigcup_{k=1}^{+\infty} \{T = k\} \cup \{T = +\infty\}$ est une partition de Ω :

$$\begin{aligned}
\mathbb{P}(M_T \in A) &= \sum_{k=1}^{+\infty} \mathbb{P}(M_k \in A, T = k) + \mathbb{P}(0 \in A, T = +\infty) \\
&= \sum_{k=1}^{+\infty} \mathbb{P}(M_k \in A, M_k \in D, M_j \notin D, j = 1, \dots k-1) \\
&= \sum_{k=1}^{+\infty} \mathbb{P}(M_k \in A \cap D) \prod_{j=1}^{k-1} \mathbb{P}(M_j \notin D) \\
&= \sum_{k=1}^{+\infty} \mu(A \cap D)[1 - \mu(D)]^{k-1} \\
&= \frac{\mu(A \cap D)}{\mu(D)} .
\end{aligned}$$

Calcul numérique de π. On considère le cercle de rayon 1 et de centre O inscrit dans le carré $[-1, 1]^2$. On tire un grand nombre de points au hasard dans ce carré. On calcule le rapport entre le nombre de points qui sont tombés dans le cercle et le nombre de points total. En multipliant ce rapport par 4 (aire du carré), on obtient une approximation de l'aire du disque, donc de π.
Ce n'est pas la méthode la plus rapide, ni la plus simple, pour évaluer numériquement π !

Exercice type corrigé

IV.3.0 Soit X une v.a. de loi uniforme sur $[a, b]$, avec $0 < a < b$. Donner la fonction de répartition, la densité (si elle existe), l'espérance et la variance de la v.a. $Y = X^2$.
Solution. On calcule la fonction de répartition F_Y de Y. Si $y < 0$, on a :

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = 0 ,$$

car $\mathbb{P}(Y \geqslant 0) = 1$. Si $y \geqslant 0$, alors :

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(X^2 \leqslant y) = \mathbb{P}(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) = \mathbb{P}(X \leqslant \sqrt{y}) = F_X(\sqrt{y}) ,$$

ce qui montre que la fonction de répartition de Y est :

$$F_Y(y) = \begin{cases} 0 & \text{si } y < a^2 , \\ \dfrac{\sqrt{y} - a}{b - a} & \text{si } y \in [a^2, b^2] , \\ 1 & \text{si } y > b^2 . \end{cases}$$

On peut mettre F_Y sous la forme d'une primitive :

$$F_Y(y) = \frac{1}{b-a} \int_{-\infty}^{\sqrt{y}} \mathbf{1}_{]a,b[}(x) \, \mathrm{d}x = \frac{1}{b-a} \int_{-\infty}^{y} \mathbf{1}_{]a^2,b^2[}(u) \frac{\mathrm{d}u}{2\sqrt{u}} ,$$

ce qui montre que Y est une v.a. à densité p_Y donnée par :

$$p_Y(y) = \frac{1}{2(b-a)\sqrt{y}} \mathbf{1}_{]a^2,b^2[}(y)\,.$$

L'espérance, le second moment et la variance de Y sont :

$$\mathbb{E}[Y] = \mathbb{E}[X^2] = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}\,,$$

$$\mathbb{E}[Y^2] = \mathbb{E}[X^4] = \frac{1}{b-a}\int_a^b x^4\,\mathrm{d}x = \frac{b^5-a^5}{5(b-a)} = \frac{a^4+a^3b+a^2b^2+ab^3+b^4}{5}\,,$$

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{b^5-a^5}{5(b-a)} - \frac{(b^3-a^3)^2}{9(b-a)^2} = \frac{4(b-a)^4}{45}\,.$$

Pour aller plus loin

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

IV.3.1 Soit X une v.a. réelle admettant pour densité f et soit $a > 0$.
Montrer que $Y = aX$ est une v.a. réelle admettant pour densité $g(x) = \frac{1}{a} f(\frac{x}{a})$.

IV.3.2 Soient a et b deux réels strictement positifs. Soit X une v.a. réelle de densité :

$$f(x) = a\exp(-b|x|), \text{ pour tout } x \in \mathbb{R}\,.$$

1. Déterminer a en fonction de b afin que f soit bien une densité de probabilité.
2. Calculer l'espérance et la variance de X.
3. Déterminer la loi de $Y = X^2$. Est-elle à densité ? Si oui, quelle est sa densité ?
4. Calculer l'espérance de Y.

IV.3.3

1. Pour quelle valeur de C la fonction $f(x) = \frac{C}{1+x^2}$ est-elle une densité de probabilité ?
   On considère alors X une v.a. réelle admettant cette densité. On dit alors que X suit la loi de Cauchy.
2. Pour quelles valeurs réelles de α le moment $\mathbb{E}[|X|^\alpha]$ est-il fini ?
3. Calculer la fonction de répartition F de X et sa fonction réciproque G. Montrer que si U est une v.a. de loi uniforme sur $[0,1]$, alors $G(U)$ suit la loi de Cauchy.

IV.3.4 Soit X une v.a. réelle dont la fonction de répartition F est définie pour $x \in \mathbb{R}$ par :

$$F(x) = \frac{x^2}{2}\mathbf{1}_{[0,1[}(x) + \frac{x+1}{4}\mathbf{1}_{[1,3[}(x) + \mathbf{1}_{[3,+\infty[}(x)\,.$$

1. Tracer le graphe de F. La fonction associée est-elle continue ? dérivable ? La v.a. X admet-elle une densité ? Si oui, donner l'expression de cette densité.
2. Calculer l'espérance de X.

IV.3.5 Cet exercice illustre la propriété d'absence de mémoire des variables exponentielles.

1. Soit T une v.a. exponentielle de paramètre $\theta > 0$, de densité $p(t) = \theta \exp(-\theta t)\mathbf{1}_{]0,+\infty[}(t)$. Vérifier que T satisfait la propriété suivante :

$$\text{Pour tous } s, t \geqslant 0,\ \mathbb{P}(T > s + t) = \mathbb{P}(T > s)\mathbb{P}(T > t)\,. \tag{24}$$

2. Soit T une v.a. à valeurs dans $\mathbb{R}_+$ vérifiant la propriété (24). On va montrer que, ou bien $\mathbb{P}(T = 0) = 1$, ou bien T est une v.a. exponentielle.
   (a) Montrer que si $\mathbb{P}(T > 0) > 0$, alors pour tout $t \geqslant 0$, $\mathbb{P}(T > t) > 0$.
   (b) Si $\mathbb{P}(T > 0) > 0$, alors on pose $\alpha = \mathbb{P}(T > 1)$. Montrer que $\mathbb{P}(T > t) = \alpha^t$ pour tout $t \in \mathbb{R}_+$ (Montrer-le d'abord pour $\alpha \in \mathbb{N}$, puis $\mathbb{Q}_+$, et enfin $\mathbb{R}_+$). Conclusion ?

IV.3.6 Cet exercice présente quelques inégalités remarquables et leurs conséquences. Soit X une v.a. réelle.

1. Montrer que si φ est convexe et si X et $\varphi(X)$ sont intégrables, alors : $\mathbb{E}[\varphi(X)] \geqslant \varphi(\mathbb{E}[X])$.
2. Déduire du 1. que pour toute suite de réels positifs $a_1, \ldots, a_n$:

$$\sum_{i=1}^{n} a_i^2 \geqslant \frac{1}{n}\left(\sum_{i=1}^{n} a_i\right)^2.$$

3. Déduire du 1. que pour toute suite de réels positifs $a_1, \ldots, a_n$:

$$\frac{1}{n}\sum_{i=1}^{n} a_i \geqslant \left(\prod_{i=1}^{n} a_i\right)^{1/n}.$$

IV.3.7 Soit T l'intérieur d'un triangle du plan délimité par les points $O := (0,0)$, $I := (1,0)$ et $J := (0,1)$ et soit (X,Y) un couple de v.a. de loi uniforme sur le triangle T.

1. Donner la densité du couple (X,Y).
2. Calculer les lois marginales de X et de Y.
3. Les v.a. X et Y sont-elles indépendantes ?
4. Calculer la covariance du couple (X,Y).

IV.3.8 On considère un gâteau circulaire avec une cerise sur le bord. On découpe le gâteau en deux parts en coupant suivant deux rayons choisis au hasard.

1. Avec quelle probabilité la part contenant la cerise est-elle plus petite que l'autre part ?
2. Quelle est la taille (i.e. la longueur angulaire) moyenne de la part contenant la cerise ?

Conclusion pour les gourmands : il faut toujours demander la part avec la cerise, même si vous n'aimez pas les cerises.

IV.3.9 Trois personnes A, B, et C arrivent à un bureau de poste en même temps. Il y a deux guichets ouverts et disponibles qu'occupent immédiatement A et B. C remplace le premier sorti. On note X, Y et Z les temps d'occupation de leur guichet par A, B et C. On suppose que ce sont des v.a. exponentielles indépendantes de même paramètre λ.

1. Calculer la loi de $U := \min(X,Y)$.
   Calculer la loi du couple (U,V) où $V := \max(X,Y) - \min(X,Y)$.

2. Quelle est la probabilité que C sorte le dernier ?
3. Quel est la loi de probabilité du temps total T passé par C à la poste ?

IV.3.10

1. Soit X une v.a. réelle suivant une loi exponentielle de paramètre $\lambda > 0$.
   On pose $Y := X^m$, $m \in \mathbb{N}$, $m \geqslant 2$. Déterminer la fonction de répartition de Y, ainsi que sa densité.
2. Calculer la probabilité que l'équation $t^2 - 2Yt + 1$ ait deux racines réelles distinctes, puis celle qu'elle n'ait pas de racines réelles.
   Quelle est la probabilité que cette équation ait une racine double ?

IV.3.11 Soit X une v.a. réelle suivant une loi normale $\mathcal{N}(m, \sigma^2)$ avec $m \in \mathbb{R}$, $\sigma > 0$.

1. On pose $Y := aX + b$ avec $a > 0$. Quelle est la loi de Y ?
2. Soient Φ la fonction de répartition de la loi normale $\mathcal{N}(0,1)$ et F_X celle de X. Soient a et b deux réels distincts. Montrer que $\frac{a-m}{\sigma} = \Phi^{-1}\big(F_X(a)\big)$ et $\frac{b-m}{\sigma} = \Phi^{-1}\big(F_X(b)\big)$. En déduire m et σ en fonction de a, b, Φ et F_X.

IV.3.12 Soit X une v.a. réelle suivant une loi normale $\mathcal{N}(0, \sigma^2)$ avec $\sigma > 0$. Soit $a \in]1, +\infty[$. Déterminer σ pour que $\mathbb{P}(X \in [a-1, a+1])$ soit maximal.

IV.3.13 Soit $(U_i)_{i \in [\![1,n]\!]}$, un système de v.a. réelles indépendantes suivant toutes la loi uniforme sur $[0,1]$. On note $X := \min\limits_{i \in [\![1,n]\!]} (U_i)$ et $Y := \max\limits_{i \in [\![1,n]\!]} (U_i)$.

1. Déterminer la densité de X, son espérance et sa variance.
2. Même question pour Y.

Théorèmes limites et estimation

IV.4

1 Loi des grands nombres

Cette section a pour but la démonstration d'un résultat intuitif, qui stipule que, lorsqu'on répète un grand nombre de fois une expérience aléatoire dans des conditions identiques et indépendantes, les quantités obtenues par moyennisation sur les résultats de toutes les expériences se stabilisent autour d'une valeur moyenne déterministe. Ce résultat a d'ailleurs été utilisé bien avant sa démonstration complète, et c'est pour cela qu'on l'a appelé « loi », à l'instar d'autres lois de la physique. On va commencer par en montrer une version élémentaire.

1.1 Loi faible des grands nombres

Définition 1. Soient $n \in \mathbb{N}^*$ et $(X_i)_{i\in[\![1,n]\!]}$ une suite de v.a. réelles indépendantes de même loi. La *moyenne empirique* de $(X_i)_{i\in[\![1,n]\!]}$ est la v.a. :

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \,. \tag{1}$$

Supposons que X_1 soit de carré intégrable. Alors la moyenne empirique $\overline{X}_n$ est aussi de carré intégrable. Elle a pour espérance :

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mu \,, \tag{2}$$

avec $\mu := \mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n]$. Ici, on a simplement utilisé la linéarité de la moyenne. La variance de la moyenne empirique est :

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\,(X_i) = \frac{\sigma^2}{n} \,, \tag{3}$$

avec $\sigma^2 := \mathrm{Var}(X_1) = \cdots = \mathrm{Var}(X_n)$. Ici, on a utilisé le fait que la variance d'une somme de v.a. indépendantes est la somme des variances. La quantité $\mathrm{Var}(\overline{X}_n)$ tendant vers 0 lorsque n tend vers $+\infty$, la variable $\overline{X}_n$ se concentre de plus en plus autour de la

constante μ. Voilà le théorème limite le plus simple des probabilités. Le hasard se manifeste par des fluctuations. Mais si l'on répète l'expérience, de manière identique et indépendante, alors la moyenne empirique fluctue de moins en moins et converge vers l'espérance mathématique.
L'inégalité de Bienaymé-Tchebychev (page 769) montre que, pour tout $\varepsilon > 0$:

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2},$$

la probabilité de déviation tend vers 0 lorsque n tend vers $+\infty$. Nous avons montré la loi faible des grands nombres, sous l'hypothèse que X_1 est de carré intégrable :

Proposition 1 (Loi faible des grands nombres). Soit $(X_i)_{i\in\mathbb{N}^*}$ une suite de v.a. indépendantes, de même loi et de carré intégrable. La moyenne empirique $\overline{X}_n$ converge en probabilité vers la constante $\mu := \mathbb{E}[X_1]$, au sens où :

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geqslant \varepsilon\right) \overset{n\to+\infty}{\longrightarrow} 0, \quad \forall \varepsilon > 0. \tag{4}$$

Exemple. Considérons une suite $(X_i)_{i\in\mathbb{N}^*}$ de v.a. indépendantes et identiquement distribuées selon la loi de Bernoulli de paramètre p. L'événement $X_i = 1$ représente un « succès » lors de la $i^{ème}$ expérience. La probabilité de succès est p. La somme $\sum_{j=1}^{n} X_j$ suit une loi binomiale de paramètres (n, p) et représente le nombre de succès lors des n premières expériences. La moyenne empirique $\overline{X}_n$ représente la proportion de succès, elle a pour moyenne p et variance $p(1-p)/n$. La loi (faible) des grands nombres affirme que $\overline{X}_n$ doit tendre vers p lorsque n tend vers $+\infty$.

1.2 Loi forte des grands nombres

Voilà une autre formulation, un peu plus forte, du même résultat.

Proposition 2 (Loi forte des grands nombres). Soit $(X_i)_{i\in\mathbb{N}^*}$ une suite de v.a. réelles indépendantes intégrables et de même loi. La moyenne empirique $\overline{X}_n$ converge avec probabilité 1 vers la constante $\mu := \mathbb{E}[X_1]$, au sens où :

$$\mathbb{P}\big(\lim_{n\to+\infty} \overline{X}_n \text{ existe et vaut } \mu\big) = 1. \tag{5}$$

La proposition 1 exprime que pour tout $\varepsilon > 0$, la v.a. $\overline{X}_n$ est comprise entre $\mu - \varepsilon$ et $\mu + \varepsilon$ avec une probabilité tendant vers 1 lorsque n tend vers $+\infty$.
La proposition 2 exprime quant à elle, qu'avec probabilité 1, on peut pour tout $\varepsilon > 0$ trouver un n_ε tel que $|\overline{X}_n - \mu| < \varepsilon$ si $n > n_\varepsilon$; ce n_ε dépend de ε et de la réalisation, i.e. du tirage des X_i.

Démonstration. On donne une preuve sous l'hypothèse plus forte que $\mathbb{E}[X_1^4] < +\infty$. Cette démonstration a le mérite d'être élémentaire, alors que la preuve sous l'hypothèse $\mathbb{E}[|X_1|] < +\infty$ nécessite des résultats fins de la théorie de l'intégration. Quitte à

considérer $X_i - \mu$ au lieu de X_i, on peut supposer que les X_i sont d'espérance nulle. Par indépendance, $\mathbb{E}[X_i X_j^3] = \mathbb{E}[X_i]\mathbb{E}[X_j^3] = 0$ pour tout $i \neq j$, et donc :

$$\begin{aligned}\mathbb{E}[\overline{X}_n^4] &= \frac{1}{n^4} \sum_{i_1,i_2,i_3,i_4=1}^{n} \mathbb{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}] = \frac{1}{n^4}\left(\sum_{i\neq j} \mathbb{E}[X_i^2 X_j^2] + \sum_{i=1}^{n} \mathbb{E}[X_i^4]\right) \\ &= \frac{1}{n^4}\left(n(n-1)\mathbb{E}[X_1^2]^2 + n\mathbb{E}[X_1^4]\right) = \mathrm{O}\left(\frac{1}{n^2}\right).\end{aligned}$$

Cela montre que la série à termes positifs $\mathbb{E}[\overline{X}_n^4]$ est convergente. En utilisant l'inégalité de Markov (proposition 11 de la page 769), on trouve que, pour tous $N, M \in \mathbb{N}^*$:

$$\mathbb{P}\left(\sum_{n=1}^{N} \overline{X}_n^4 > M\right) \leqslant \frac{1}{M}\mathbb{E}\left[\sum_{n=1}^{N} \overline{X}_n^4\right] = \frac{1}{M}\sum_{n=1}^{N} \mathbb{E}\left[\overline{X}_n^4\right] \leqslant \frac{C}{M},$$

avec $C = \sum_{n=1}^{+\infty} \mathbb{E}\left[\overline{X}_n^4\right]$. Or la somme de la série positive $S := \sum_{n=1}^{+\infty} \overline{X}_n^4$ est une v.a. à valeurs dans $[0, +\infty]$ qui vérifie :

$$\{S \leqslant M\} = \bigcap_{N\in\mathbb{N}^*} \left\{\sum_{n=1}^{N} \overline{X}_n^4 \leqslant M\right\}.$$

Comme, pour M fixé, les événements du membre de droite forment une suite décroissante en N :

$$\mathbb{P}(S \leqslant M) = \lim_{N\to+\infty} \mathbb{P}\left(\sum_{n=1}^{N} \overline{X}_n^4 \leqslant M\right) = 1 - \lim_{N\to+\infty} \mathbb{P}\left(\sum_{n=1}^{N} \overline{X}_n^4 > M\right) \geqslant 1 - \frac{C}{M}\cdot$$

On a aussi :

$$\{S < +\infty\} = \bigcup_{M\in\mathbb{N}^*} \{S \leqslant M\}.$$

Comme les événements du membre de droite forment une suite croissante en M :

$$\mathbb{P}(S < +\infty) = \lim_{M\to+\infty} \mathbb{P}(S \leqslant M) = 1.$$

Avec probabilité 1, la somme de la série S est finie, donc le terme général positif $\overline{X}_n^4$ converge vers 0 lorsque n tend vers $+\infty$, ce qui est le résultat voulu. ∎

Exemple. On répète indépendamment une même expérience aléatoire n fois, et l'on s'intéresse à la fréquence empirique d'apparition d'un événement A, c'est-à-dire le rapport $F_n(A) := N(A)/n$ où $N(A)$ est le nombre de fois où l'événement A s'est réalisé. Posons $X_i := 1$ si A est réalisé lors de la $i^{ème}$ répétition, $X_i := 0$ sinon. Les variables aléatoires $(X_i)_{i\geqslant 1}$ sont indépendantes et identiquement distribuées selon la loi de Bernoulli de paramètre $\mathbb{P}(A) = \mathbb{E}[X_i]$. La loi forte des grands nombres affirme que la fréquence empirique :

$$F_n(A) := \frac{1}{n}\sum_{i=1}^{n} X_i$$

converge avec probabilité 1 vers $\mathbb{E}[X_1]$ qui n'est autre que la probabilité $\mathbb{P}(A)$:

$$\mathbb{P}\left(\lim_{n\to+\infty} F_n(A) = \mathbb{P}(A)\right) = 1\,.$$

Il suffit d'effectuer une seule suite infinie de répétitions de notre expérience aléatoire, pour découvrir empiriquement la valeur de $\mathbb{P}(A)$. Notons que la loi faible des grands nombres ne donne pas un tel résultat. Nous avons ainsi démontré dans notre cadre théorique, ce résultat qui est conforme à notre expérience et intuition de la notion de probabilité.

Remarque. La loi forte des grands nombres est bien plus forte que la loi faible. D'une part, les hypothèses de la loi forte sont plus faibles que celles de la loi faible, car elle réclame seulement que X_1 soit intégrable, alors que la loi faible réclame que X_1 soit de carré intégrable. D'autre part, le résutat de la loi forte est plus fort, car on peut montrer que (5) implique (4) avec une extension de théorème de convergence dominée qu'on a vu dans le module Intégration de ce livre (le théorème dans ce module est restreint aux fonctions intégrables au sens de Lebesgue). Historiquement, la loi faible a été montrée avant la loi forte, la preuve de cette dernière par Kolmogorov date de 1929.

2 Théorème de la limite centrale

Dans ce paragraphe on considère une suite de v.a. indépendantes et identiquement distribuées $(X_n)_{n\geqslant 1}$ de carré intégrable. Pour la moyenne empirique $\overline{X}_n$, la loi des grands nombres affirme que :

$$\overline{X}_n \overset{n\to+\infty}{\longrightarrow} \mu := \mathbb{E}[X_1]\,,$$

avec probabilité 1. Pour affiner ce résultat, on étudie ici les fluctuations de $\overline{X}_n$ autour de sa limite. Ces fluctuations sont de moyenne nulle, car, d'après (2), $\mathbb{E}[\overline{X}_n - \mu] = 0$. La « taille moyenne » des fluctuations peut s'estimer en évaluant la variance de $\overline{X}_n$, qui représente l'écart quadratique moyen, et qui est donnée par (3) : $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$, où l'on a noté $\sigma^2 := \mathrm{Var}(X_1)$. Nous considérerons (en nous restreignant au cas $\sigma > 0$) :

$$\frac{\overline{X}_n - \mu}{\sqrt{\mathrm{Var}\left(\overline{X}_n\right)}} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}\,,$$

qui est une v.a. de moyenne nulle et de variance 1. Pour un n donné, cette v.a. possède une loi qui peut être compliquée et qui dépend évidemment de la loi des X_n. Une propriété remarquable est que la distribution de ces fluctuations approche, quand n tend vers $+\infty$, la loi gaussienne, quelle que soit la loi commune des X_n (sous la seule réserve que les X_n soient de carré intégrable). On va d'abord considérer le cas où les variables $(X_n)_{n\geqslant 1}$ suivent des lois de Bernoulli.

2.1 Approximation normale de la loi binomiale

La proposition qui suit (théorème de De Moivre-Laplace) a été établie pour la première fois dans le cas particulier $p = 1/2$ par De Moivre en 1733 et généralisée à toute valeur de p par Laplace en 1812.

Proposition 3. Soit $(X_n)_{n\geqslant 1}$ une suite de v.a. indépendantes de même loi de Bernoulli de paramètre $p \in]0,1[$. Soit $S_n = X_1 + \cdots + X_n$.
(i) On a, pour tout $x \in \mathbb{R}$:

$$\mathbb{P}\left(S_n = [np + \sqrt{n}x]\right) \overset{n\to+\infty}{\sim} \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{x^2}{2p(1-p)}\right),$$

où $[\cdot]$ désigne la partie entière.
(ii) Soit $\overline{X}_n = S_n/n$ la moyenne empirique. Pour tout intervalle I de $\mathbb{R}$, on a :

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \in I\right) \overset{n\to+\infty}{\longrightarrow} \int_I g(x)\,\mathrm{d}x\,,$$

où $g(x) := (2\pi)^{-1/2}\exp(-x^2/2)$ est la densité de la loi gaussienne $\mathcal{N}(0,1)$.

Pour tout n, la v.a. S_n suit la loi binomiale de paramètres (n,p). Cette proposition décrit donc comment on peut approcher, lorsque n est grand, la loi binomiale de paramètres (n,p) par la loi gaussienne de moyenne np et de variance $np(1-p)$ (qui sont les moyenne et variance de la loi $\mathcal{B}(n,p)$). La preuve est élémentaire mais calculatoire et elle utilise la formule de Stirling $n! \sim (n/e)^n\sqrt{2\pi n}$ lorsque $n \to +\infty$.

Démonstration. On considère d'abord le cas symétrique $p = 1/2$ et l'on note P la loi de probabilité de S_n. On supposera que n est pair, le cas impair se traitant de même. Soit m un entier compris entre $n/2+1$ et n. On a :

$$P(\{m\}) = \frac{n!2^{-n}}{m!(n-m)!} = \frac{n!2^{-n}}{(\frac{n}{2})!(\frac{n}{2})!} \times \frac{\frac{n}{2}\left(\frac{n}{2}-1\right)\cdots\left[\frac{n}{2}-\left(m-\frac{n}{2}-1\right)\right]}{m(m-1)\cdots\left(\frac{n}{2}+1\right)}.$$

On applique la formule de Stirling aux trois termes factoriels du membre de droite pour obtenir l'équivalent :

$$\frac{n!2^{-n}}{(\frac{n}{2})!(\frac{n}{2})!} \sim \frac{\sqrt{2}}{\sqrt{\pi n}}\,,$$

et l'on divise chacun des autres termes (au dénominateur et au numérateur) par $n/2$:

$$P(\{m\}) \sim \frac{\sqrt{2}}{\sqrt{\pi n}} \times \frac{\left(1-\frac{2}{n}\right)\cdots\left[1-\left(m-\frac{n}{2}-1\right)\frac{2}{n}\right]}{\left(1+\frac{2}{n}\right)\cdots\left[1+\left(m-\frac{n}{2}\right)\frac{2}{n}\right]}.$$

On évalue cette identité en $m = k + n/2$, où k est un entier compris entre 1 et $n/2$:

$$P\left(\{\frac{n}{2}+k\}\right) \sim \frac{\sqrt{2}}{\sqrt{\pi n}} \times \frac{\left(1-\frac{2}{n}\right)\cdots\left[1-(k-1)\frac{2}{n}\right]}{\left(1+\frac{2}{n}\right)\cdots\left[1+k\frac{2}{n}\right]}.$$

On utilise l'inégalité $|\exp(-x) - (1-x)| \leqslant x^2/2$ valable pour tout $x \in [0,1]$ pour écrire que, pour tout $j \in [\![1, n/2]\!]$:

$$\left|\exp\left(-\frac{2j}{n}\right) - \left(1 - j\frac{2}{n}\right)\right| \leqslant \frac{2j^2}{n^2}\,,$$

et donc :

$$\left|\left(1-\frac{2}{n}\right)\cdots\left[1-(k-1)\,\frac{2}{n}\right]-\exp\left(-\frac{2}{n}-\cdots-(k-1)\,\frac{2}{n}\right)\right|\leqslant\frac{2k^3}{n^2}\cdot$$

Si k est d'ordre $\sqrt{n}$, alors k^3/n^2 est d'ordre $1/\sqrt{n}$ et l'on trouve l'équivalent :

$$\begin{aligned}P(\{\frac{n}{2}+k\})&\sim\frac{\sqrt{2}}{\sqrt{\pi n}}\times\exp\left\{-\frac{2}{n}-\cdots-(k-1)\,\frac{2}{n}-\left[\frac{2}{n}+\cdots+k\frac{2}{n}\right]\right\}\\&\sim\frac{\sqrt{2}}{\sqrt{\pi n}}\times\exp\left\{-\frac{4}{n}\left[\sum_{j=1}^{k-1}j\right]-k\frac{2}{n}\right\}=\frac{\sqrt{2}}{\sqrt{\pi n}}\times\exp\left\{-\frac{2k^2}{n}\right\},\end{aligned}$$

donc, lorsque $k=[\sqrt{n}x]$, $x>0$, on obtient :

$$P\left(\{\frac{n}{2}+[\sqrt{n}x]\}\right)\sim\frac{\sqrt{2}}{\sqrt{\pi n}}\times\exp\left(-2x^2\right)\,.$$

Le cas asymétrique $p\neq 1/2$ se traite de façon similaire, avec des calculs un peu plus compliqués. On note $q=1-p$. On suppose que np et nq sont entiers pour ne pas avoir à prendre des parties entières (c'est juste un détail technique), et l'on calcule, pour $k>np$:

$$P(\{k\})=\frac{n!p^kq^{n-k}}{k!(n-k)!}=\frac{n!p^kq^{n-k}}{(np)!(nq)!}\times\frac{nq\,(nq-1)\cdots[nq-(k-np-1)]}{k(k-1)\cdots(np+1)}\cdot$$

En utilisant la formule de Stirling, on obtient au point $np+k$:

$$P(\{np+k\})\sim\frac{1}{\sqrt{2\pi npq}}\times\frac{\left(1-\frac{1}{nq}\right)\cdots\left(1-\frac{k-1}{nq}\right)}{\left(1+\frac{1}{np}\right)\cdots\left(1+\frac{k}{np}\right)}\,,$$

lorsque $n\to+\infty$. En utilisant comme ci-dessus l'approximation $\exp(x)=1+x+O(x^2)$, on obtient que, si k est au plus d'ordre $\sqrt{n}$:

$$\begin{aligned}P(\{np+k\})&\sim\frac{1}{\sqrt{2\pi npq}}\times\exp\left\{-\frac{1}{nq}-\cdots-\frac{k-1}{nq}-\left[\frac{1}{np}+\cdots+\frac{k}{np}\right]\right\}\\&\sim\frac{1}{\sqrt{2\pi npq}}\times\exp\left\{-\frac{k^2}{2n}\left(\frac{1}{p}+\frac{1}{q}\right)\right\}=\frac{1}{\sqrt{2\pi npq}}\times\exp\left\{-\frac{k^2}{2npq}\right\},\end{aligned}$$

ce qui donne le résultat énoncé dans le premier point de la proposition lorsque $k=[\sqrt{n}x]$. Si $I=]a,b]$ est un intervalle de $\mathbb{R}$, alors :

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n-p}{\sqrt{p(1-p)}}\in I\right)=\sum_{k=k_a}^{k_b}\mathbb{P}(S_n=k)\,,$$

où $k_a=[np+\sqrt{np(1-p)}a]+1$ et $k_b=[np+\sqrt{np(1-p)}b]$. On utilise le premier point de la proposition et l'on reconnaît une somme de Riemann :

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n-p}{\sqrt{p(1-p)}}\in I\right)\overset{n\to+\infty}{\longrightarrow}\int_{\sqrt{p(1-p)}a}^{\sqrt{p(1-p)}b}\frac{1}{\sqrt{2\pi p(1-p)}}\exp\left(-\frac{x^2}{2p(1-p)}\right)\mathrm{d}x\,.$$

En effectuant le changement de variable $x:=\sqrt{p(1-p)}y$, on obtient le résultat voulu. ■

En raffinant la preuve précédente, on peut donner des majorations fines de l'écart entre les probabilités à n fixé et leur approximation gaussienne asymptotique. Cet écart est d'ordre $1/\sqrt{n}$, et l'on peut en fait appliquer en pratique l'approximation $\mathcal{B}(n,p) \simeq \mathcal{N}\big(np, np(1-p)\big)$ dès que $np \geqslant 10$ et $n(1-p) \geqslant 10$.
On peut aussi refaire le même genre de calculs dans d'autres cas, par exemple dans le cas où les X_i sont indépendantes et suivent une loi de Poisson $\mathcal{P}(\lambda)$. On sait alors (voir page 746) que la somme $S_n := n\overline{X}_n$ suit une loi de Poisson de paramètre $n\lambda$, et on montre par des développements asymptotiques (en utilisant la formule de Stirling) que $\mathcal{P}(n\lambda) \simeq \mathcal{N}(n\lambda, n\lambda)$ lorsque n tend vers $+\infty$ dans le sens où :

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}} \in I\right) \overset{n\to+\infty}{\longrightarrow} \int_I g(x)\,\mathrm{d}x\,.$$

L'approximation $\mathcal{P}(n\lambda) \simeq \mathcal{N}(n\lambda, n\lambda)$ est en pratique valable dès que $n\lambda \geqslant 15$.
L'approximation d'une somme de v.a. indépendantes par une loi normale est en fait quasi-universelle, comme on va le voir dans le paragraphe suivant.

2.2 Énoncé du théorème

Théorème 4 (de la limite centrale).
Soit $(X_n)_{n\geqslant 1}$ une suite de v.a. indépendantes de même loi de carré intégrable, de moyenne $\mu \in \mathbb{R}$ et de variance σ^2, $\sigma \in \,]0,+\infty[$. Alors, en notant $\overline{X}_n$ la moyenne empirique, on a pour tout intervalle $I \subset \mathbb{R}$:

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \in I\right) \overset{n\to+\infty}{\longrightarrow} \int_I g(x)\,\mathrm{d}x\,,$$

où $g(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ est la densité de la loi gaussienne $\mathcal{N}(0,1)$.

Ce théorème montre que la loi normale décrit les fluctuations autour du comportement déterministe prédit par la loi des grands nombres. Sa démonstration nécessite au préalable de développer plus en avant la notion de convergence en loi qu'on n'abordera pas dans le cadre de ce cours.
On a vu dans le paragraphe précédent qu'on pouvait obtenir une preuve de manière calculatoire dans le cas particulier où la loi des X_n est une loi de Bernoulli. On a aussi vu dans la proposition 23 de la page 779 que le résultat du théorème est trivial lorsque les v.a. X_n sont indépendantes et suivent des lois gaussiennes de moyenne μ et de variance σ^2, car alors $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ suit une loi gaussienne de moyenne nulle et de variance 1 pour tout $n \in \mathbb{N}^*$.

Exercice 1.
Soit X_n une v.a. de loi χ^2_n, pour $n \geqslant 1$. En appliquant le théorème de la limite centrale, montrer que la loi de $n^{-1/2}(X_n - n)$ peut être approchée, lorsque n tend vers $+\infty$, par la loi $\mathcal{N}(0,2)$.
Solution. Si X_n suit la loi χ^2_n, alors il existe des v.a. gaussiennes centrées réduites in-

dépendantes Y_i telles que X_n et $\sum_{i=1}^{n} Y_i^2$ suivent la même loi. Les v.a. Y_i^2 sont de carrés intégrables, i.e. $\mathbb{E}[Y_i^4] = 3 < +\infty$ (voir l'exercice de la page 784), donc on peut appliquer le théorème de la limite centrale. On obtient que la loi de :

$$\sqrt{n}\frac{X_n/n - \mathbb{E}[Y_1^2]}{\sqrt{\mathrm{Var}(Y_1^2)}}$$

peut être approchée par la loi normale centrée réduite quand n tend vers $+\infty$.
Comme $\mathbb{E}[Y_1^2] = 1$ et $\mathrm{Var}(Y_1^2) = 2$, on obtient le résultat annoncé.

Il y a trois hypothèses dans le théorème de la limite centrale : l'indépendance, l'équidistribution et le second moment fini. L'hypothèse d'indépendance est celle que l'on retient facilement, et qui est la clé du résultat. L'hypothèse d'équidistribution n'est pas vraiment nécessaire, et il existe des versions du théorème de la limite centrale avec des hypothèses plus larges. L'hypothèse selon laquelle les v.a. que l'on additionne doivent être de carré intégrable est très importante, pour la preuve et pour le résultat lui-même. En effet, il est facile de construire des exemples de sommes de v.a. indépendantes et de même loi, qui ne se laissent pas approcher par une loi normale. La somme peut ne pas converger du tout, mais elle peut aussi converger vers une loi qui n'est pas la loi normale, comme le montre l'exercice suivant.

Exercice 2.

1. Soient X_1 et X_2 deux v.a. indépendantes de densités respectives :

$$p_{X_j}(x) := \frac{\sigma_j}{\pi}\frac{1}{\sigma_j^2 + x^2}\,,$$

où $\sigma_1, \sigma_2 > 0$. Ce sont des lois dites de Cauchy. Montrer que les v.a. X_j ne sont pas de carré intégrable. Montrer que $S := X_1 + X_2$ est une v.a. à densité :

$$p_S(x) = \frac{\sigma}{\pi}\frac{1}{\sigma^2 + x^2}\,,$$

avec $\sigma = \sigma_1 + \sigma_2$.

2. Soit $(X_i)_{i\in[\![1,n]\!]}$ une suite de de n v.a. indépendantes de loi de Cauchy de même paramètre σ. Montrer que la loi de la somme $\sum_{i=1}^{n} X_i$ est une loi de Cauchy de paramètre $n\sigma$. En déduire que la loi de la moyenne empirique $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ est une loi de Cauchy de paramètre 1.

Solution. 1. Comme $p_{X_j}(x) \sim x^{-2}$ à l'infini, les v.a. X_j ne sont pas de carré intégrable, ni même intégrables. La proposition 22 de la page 778 montre que la v.a. S est à densité p_S donnée par :

$$p_S(s) = \frac{\sigma_1\sigma_2}{\pi^2}\int_{\mathbb{R}} \frac{1}{(\sigma_1^2 + x^2)(\sigma_2^2 + (s-x)^2)}\,\mathrm{d}x\,.$$

Le problème est donc de calculer l'intégrale d'une fraction rationnelle. On met cette fraction sous forme irréductible. Soit $s \neq 0$ fixé. On a :

$$\frac{1}{(\sigma_1^2+x^2)(\sigma_2^2+(s-x)^2)} = \frac{\alpha x+\beta}{\sigma_1^2+x^2} + \frac{\gamma(x-s)+\delta}{\sigma_2^2+(x-s)^2}\,,$$

avec

$$\frac{\alpha}{2s} = \frac{1}{(s^2+(\sigma_1+\sigma_2)^2)(s^2+(\sigma_1-\sigma_2)^2)}\,, \quad \beta = \frac{\alpha}{2s}(s^2+\sigma_2^2-\sigma_1^2)\,,$$

$$\gamma = -\alpha\,, \quad \delta = \frac{\alpha}{2s}(s^2-\sigma_2^2+\sigma_1^2)\,.$$

On trouve alors (il faut faire un peu attention, car $x \mapsto x/(\sigma^2+x^2)$ n'est pas intégrable sur $\mathbb{R}$) :

$$\begin{aligned}
p_S(s) &= \frac{\sigma_1\sigma_2}{\pi^2} \lim_{M\to+\infty} \int_{-M}^{M} \frac{\alpha x+\beta}{\sigma_1^2+x^2} + \frac{\gamma(x-s)+\delta}{\sigma_2^2+(x-s)^2}\,\mathrm{d}x \\
&= \frac{\sigma_1\sigma_2}{\pi^2} \lim_{M\to+\infty} \Big\{\alpha \int_{-M}^{M} \frac{x}{\sigma_1^2+x^2}\,\mathrm{d}x + \beta \int_{-M}^{M} \frac{1}{\sigma_1^2+x^2}\,\mathrm{d}x \\
&\quad + \gamma \int_{-M}^{M} \frac{x-s}{\sigma_2^2+(x-s)^2}\,\mathrm{d}x + \delta \int_{-M}^{M} \frac{x-s}{\sigma_2^2+(x-s)^2}\,\mathrm{d}x\Big\}\,.
\end{aligned}$$

Chacune des quatre intégrales a une limite quand $M \to +\infty$. En effet :

i) $\displaystyle\int_{-M}^{M} \frac{x}{\sigma_1^2+x^2}\,\mathrm{d}x = 0$ pour tout M car la fonction intégrée est impaire.

ii) $\displaystyle\int_{-M}^{M} \frac{1}{\sigma_1^2+x^2}\,\mathrm{d}x \to \int_{-\infty}^{\infty} \frac{1}{\sigma_1^2+x^2}\,\mathrm{d}x = \frac{\pi}{\sigma_1}$ quand $M \to +\infty$.

iii)
$$\begin{aligned}
\int_{-M}^{M} \frac{x-s}{\sigma_2^2+(x-s)^2}\,\mathrm{d}x &= \int_{-M-s}^{M-s} \frac{x}{\sigma_2^2+x^2}\,\mathrm{d}x \\
&= \int_{-M-s}^{M+s} \frac{x}{\sigma_2^2+x^2}\,\mathrm{d}x - \int_{M-s}^{M+s} \frac{x}{\sigma_2^2+x^2}\,\mathrm{d}x \\
&= -\int_{M-s}^{M+s} \frac{x}{\sigma_2^2+x^2}\,\mathrm{d}x
\end{aligned}$$

est majorée (en valeur absolue) par $2|s|/(M-|s|)$, donc tend vers 0 quand $M \to +\infty$.

iv) $\displaystyle\int_{-M}^{M} \frac{1}{\sigma_2^2+(x-s)^2}\,\mathrm{d}x \to \int_{-\infty}^{\infty} \frac{1}{\sigma_2^2+x^2}\,\mathrm{d}x = \frac{\pi}{\sigma_2}$ quand $M \to +\infty$.

On obtient finalement :

$$p_S(s) = \frac{\sigma_1\sigma_2}{\pi^2}\Big(\frac{\pi\beta}{\sigma_1} + \frac{\pi\delta}{\sigma_2}\Big) = \frac{\sigma_1+\sigma_2}{\pi}\frac{1}{s^2+(\sigma_1+\sigma_2)^2}\,,$$

ce qui démontre que S est v.a. de loi de Cauchy de paramètre $\sigma_1+\sigma_2$.

Il existe une preuve moins calculatoire utilisant la notion de fonction caractérisque de loi de probabilité qu'on ne verra pas dans ce cours.

2. On montre par récurrence sur n que $\sum_{i=1}^{n} X_i$ est une v.a. de loi de Cauchy de paramètre $n\sigma$. En effet, c'est vrai pour $n = 1$. Supposons qu'on ait montré que $\sum_{i=1}^{n-1} X_i$ est une v.a. de loi de Cauchy de paramètre $(n-1)\sigma$. Alors $\sum_{i=1}^{n-1} X_i$ et X_n sont indépendantes et en appliquant le résultat de la première question, on trouve que $\sum_{i=1}^{n} X_i$ est une v.a. de loi de Cauchy de paramètre $n\sigma$. Le dernier résultat sur la loi de la moyenne empirique $\overline{X}_n$ s'obtient en regardant la fonction de répartition de $\overline{X}_n$. On a pour tout $s \in \mathbb{R}$:

$$\mathbb{P}(\overline{X}_n \leqslant s) = \mathbb{P}\Big(\sum_{i=1}^{n} X_i \leqslant ns\Big) = \frac{n\sigma}{\pi}\int_{-\infty}^{ns} \frac{1}{n^2\sigma^2 + x^2}\,\mathrm{d}x \overset{x=ny}{=} \frac{\sigma}{\pi}\int_{-\infty}^{s} \frac{1}{\sigma^2 + y^2}\,\mathrm{d}y\,,$$

ce qui montre que $\overline{X}_n$ est une v.a. de loi de Cauchy de paramètre σ.

La loi de Gauss des erreurs de mesure fournit une application importante de ces idées. L'erreur de mesure $\Delta X := X - x$ commise en mesurant une grandeur, qui est égale à la différence entre la mesure X de cette grandeur et sa valeur exacte x, est souvent due en pratique à un grand nombre de causes distinctes. En outre, il est souvent naturel d'admettre que ces causes d'erreurs sont indépendantes entre elles. D'autre part, si nous éliminons de nos considérations les erreurs « systématiques » qui agissent dans un sens déterminé, pour ne retenir que les « erreurs dues au hasard », nous sommes amenés à considérer les erreurs comme des v.a. réelles centrées (c'est-à-dire *grosso modo* comme pouvant être aussi bien positives que négatives). Nous sommes ainsi conduits à penser que l'erreur de mesure ΔX est la somme d'un grand nombre de v.a. centrées indépendantes. Cette erreur suit donc, d'après ce qui précède, approximativement une loi normale. De nombreuses variables courantes d'intérêt sont distribuées approximativement selon une loi gaussienne. Ce phénomène peut être souvent vérifié expérimentalement pour des grandeurs biologiques (par exemple, taille, poids,...), pour la variabilité de la production de chaînes industrielles, pour des erreurs de mesures en physique expérimentale, etc. Ce phénomène peut être aussi justifié théoriquement dans certains cas, à l'aide du théorème 4 ou de ses généralisations. La loi normale occupe donc une place importante dans la modélisation de phénomènes aléatoires. Le graphe de la densité g de la loi normale centrée réduite est la célèbre « courbe en cloche » représentée sur la figure IV.4.1 et il est très utile pour les applications de retenir au moins que :

$$\int_{-c}^{c} g(x)\,\mathrm{d}x = \begin{cases} 0{,}95 & \text{pour } c \simeq 1{,}96\,, \\ 0{,}99 & \text{pour } c \simeq 2{,}6\,. \end{cases}$$

Le tableau IV.4.1 de la page 807 donne des valeurs de la fonction de répartition de la loi gaussienne centrée réduite.

x	0,00	0,010	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

x	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
$\Phi(x)$	0,99865	0,99903	0,99931	0,99952	0,99966	0,99977	0,999841	0,999928	0,999968	0,999997

TABLE IV.4.1 – Valeurs de la fonction de répartition $\Phi(x) = \displaystyle\int_{-\infty}^{x} g(y)\,\mathrm{d}y$ de la loi normale $\mathcal{N}(0,1)$. Si on cherche la valeur de $\Phi(x = k \times 0{,}1 + j \times 0{,}01)$, il faut regarder dans le gros tableau supérieur, la valeur située à l'intersection de la ligne $k+1$ et de la colonne $j+1$. Le tableau donne ainsi toutes les valeurs de $\Phi(x)$ pour x allant de 0 à $2{,}99$ par pas de $0{,}01$. Le petit tableau inférieur donne les valeurs de $\Phi(x)$ pour x allant de 3 à $4{,}5$ par pas variable. Enfin, si on souhaite trouver une valeur de $\Phi(x)$ pour un x négatif, alors on utilise le fait que $\Phi(x) = 1 - \Phi(-x)$.

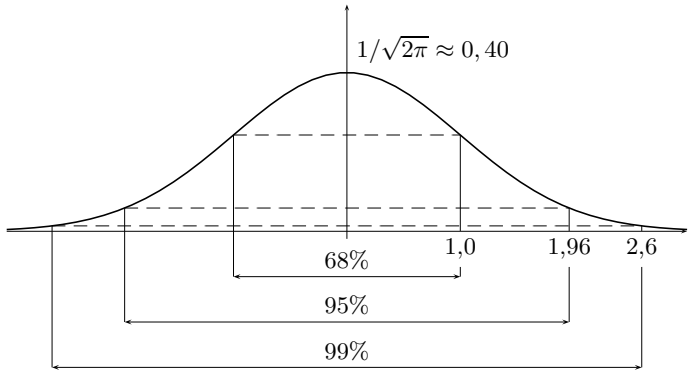


FIGURE IV.4.1 – La densité g de la loi gaussienne $\mathcal{N}(0,1)$.

Exemple.
Après $n = 400$ parties de pile ou face (la probabilité d'obtenir pile étant $p = 1/2$), la proportion $P_{400} := N_{400}/400$ de piles obtenus est une v.a. dont les valeurs probables sont voisines de $1/2$ et pour laquelle le théorème de la limite centrale montre que :

$$\mathbb{P}\left(\sqrt{n}\frac{|P_n - p|}{\sigma} < c\right) \simeq \int_{-c}^{c} g(x)\,\mathrm{d}x\,,$$

avec $\sigma := \sqrt{p(1-p)}$. Ici, $\sqrt{n} = 20$, $p = 1/2$ et $\sigma = 1/2$, ce qui donne :

$$\mathbb{P}\left(40\,|\,P_{400} - 0{,}5| < c\right) \simeq \int_{-c}^{c} g(x)\,\mathrm{d}x\,.$$

Comme on l'a vu ci-dessus, cette probabilité vaut $0{,}95$ (respectivement $0{,}99$) pour $c = 1{,}96$ (respectivement pour $c = 2{,}6$). Ainsi, lors d'une expérience de 400 lancers, la probabilité de trouver que :

$$P_{400} \in [0{,}45,\ 0{,}55]\,, \quad \text{respectivement } P_{400} \in [0{,}435,\ 0{,}565]\,,$$

est de 95 %, respectivement 99 %. Lorsque n varie, le théorème de la limite centrale montre que la loi de $\sqrt{n}(P_n - 0{,}5)$ varie lentement, en $\sqrt{n}$. Pour resserrer la fourchette $[0{,}45,\ 0{,}55]$ ci-dessus d'un facteur 10, il faut multiplier n par le facteur $10^2 = 100$; autrement dit c'est seulement pour n plus grand que $40\,000$ qu'il est vrai que $P_n \in [0{,}495,\ 0{,}505]$ avec une probabilité d'au moins $0{,}95$.

Exemple. Le cahier des charges d'un central téléphonique desservant $1\,000$ abonnés prévoit que si, à un instant donné, chaque abonné a la probabilité $4\ \%$ d'être en ligne, le trafic doit être entièrement assuré dans $97{,}5\ \%$ des cas. Le nombre minimum ν de circuits à prévoir doit donc être tel que si N désigne le nombre d'abonnés en ligne à un instant donné :

$$\mathbb{P}(N \leqslant \nu) \geqslant 0{,}975\,. \tag{6}$$

D'après ce qui précède, il y a en moyenne $1\,000 \times 0{,}04 = 40$ abonnés en ligne à chaque instant. On pourrait donc s'attendre à ce que 40 lignes suffisent. Cependant, le nombre d'abonnés en ligne est un nombre aléatoire, avec des fluctuations qu'il faut prendre en

compte dans la conception du central. Il est naturel de supposer que les abonnés se comportent indépendamment les uns des autres, donc on modélise N comme une v.a. binomiale de paramètres $n = 1\,000$ et $p = 0{,}04$ pour laquelle le théorème de la limite centrale montre que :

$$\mathbb{P}\big(N - np > c\sqrt{np(1-p)}\big) \simeq 0{,}975 \text{ pour } c = 1{,}96\,.$$

Donc le ν le plus petit qui vérifie (6) est le premier entier supérieur à $np+1{,}96\sqrt{np(1-p)}$, soit :

$$\nu = \Big[np + 1{,}96\sqrt{np(1-p)}\Big] + 1 = 53\,.$$

Le ν trouvé est nettement supérieur au nombre moyen $\mathbb{E}[N] = np = 40$ d'abonnés en ligne ; on remarquera que $\mathbb{P}(N > 40) \sim 1/2$ et qu'en prenant $\nu = 40$ on serait donc loin de remplir les conditions du cahier des charges.

Exercice 3.
Dans le contexte précédent, serait-il plus avantageux de construire deux centraux desservant chacun 500 abonnés ?

Solution. Pour $j = 1, 2$, on note N_j le nombre d'abonnés en ligne à un instant donné sur chacun des deux centraux. On cherche le nombre total de circuits ν (répartis sur les deux centraux de manière égale) tel que :

$$\mathbb{P}(N_1 \leqslant \nu/2,\ N_2 \leqslant \nu/2) = 0{,}975\,.$$

En utilisant l'indépendance des deux événements, qui résulte de l'indépendance de N_1 et N_2 :

$$\mathbb{P}(N_1 \leqslant \nu/2) = \sqrt{0{,}975} \simeq 0{,}9874\,.$$

Or le nombre N_1 d'abonnés connectés au premier central parmi les $n/2 = 500$ abonnés à ce central vérifie :

$$\mathbb{P}\big(N_1 - (n/2)p > c\sqrt{(n/2)p(1-p)}\big) = 0{,}9874 \text{ pour } c = 2{,}24\,.$$

Donc il faut prendre :

$$\nu = 2\left\{\Big[(n/2)p + 2{,}24\sqrt{(n/2)p(1-p)}\Big] + 1\right\} = 60\,.$$

Il faut prévoir plus de lignes dans le schéma à deux centraux que dans le schéma à un central unique. On a donc intérêt à regrouper les centraux, car on réduit ainsi les fluctuations.

---

3 Estimation

Les théorèmes limites des deux sections précédentes vont nous permettre d'aborder des questions fondamentales de statistique, autour de l'estimation.

3.1 But de l'estimation

Le but de l'estimation statistique est le suivant. On observe des réalisations d'un phénomène aléatoire (tirages du loto, pluviométrie journalière, ...) modélisé par une v.a. réelle X. On suppose que la loi de X dépend d'un ou de plusieurs paramètres inconnus. On souhaite déterminer, ou plus exactement estimer, la valeur de ce(s) paramètre(s). On notera génériquement θ le paramètre, que l'on suppose appartenir à un ensemble Θ connu. Pour cela on

dispose d'un certains nombre n de réalisations indépendantes de la v.a. X : les observations.

Exemple. La durée de vie d'une ampoule électrique produite par une usine est correctement modélisée par une v.a. exponentielle de paramètre λ. Pour estimer λ, on laisse allumées $n := 1\,000$ ampoules jusqu'à ce qu'elles grillent, on observe donc $n = 1\,000$ durées de vie, à partir desquelles on essaye d'inférer une estimation de λ.

D'une manière générale, un modèle statistique est défini de manière similaire à un espace de probabilité, à la différence essentielle près que le dernier élément du triplet n'est pas une probabilité, mais une famille de probabilités.

Définition 2. Un *modèle statistique* est un triplet $(\Omega, \mathcal{A}, \mathcal{P})$ où Ω est l'espace fondamental, $\mathcal{A}$ est une tribu sur Ω, et $\mathcal{P}$ est une *famille* de probabilités sur $(\Omega, \mathcal{A})$. Un modèle statistique est dit *paramétrique* s'il existe un entier p et un ensemble $\Theta \subset \mathbb{R}^p$ tels que la famille de probabilités $\mathcal{P}$ soit de la forme :

$$\mathcal{P} = \left\{ \mathbb{P}_\theta \;\middle|\; \theta \in \Theta \right\} ,$$

où, pour tout $\theta \in \Theta$, $\mathbb{P}_\theta$ est une probabilité sur $(\Omega, \mathcal{A})$.

Exemple. Dans l'exemple précédent, on considère sur l'espace $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ l'ensemble des lois exponentielles :

$$\mathcal{P} = \{ P_\lambda \mid \lambda > 0 \} ,$$

où P_λ est la probabilité sur $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ de densité $p_\lambda(x) = \lambda e^{-\lambda x} \mathbf{1}_{]0,+\infty[}(x)$.

Définition 3. Un *n-échantillon* d'une v.a. réelle X est une famille $(X_i)_{i \in [\![1,n]\!]}$ de n v.a. indépendantes et identiquement distribuées de même loi que X.

On observe un échantillon, c'est-à-dire un vecteur $(X_1, \ldots, X_n)$ de n répliques indépendantes de la v.a. X. On souhaite construire à partir de cet échantillon une quantité $f(X_1, \ldots, X_n)$ qui s'approche du paramètre inconnu θ. Une telle quantité est appelée estimateur de θ et est souvent notée $\hat{\theta}_n$. Un estimateur ne dépend que de l'échantillon $(X_1, \ldots, X_n)$, et est donc entièrement caractérisé par la fonction déterministe $f : \mathbb{R}^n \to \Theta$. L'objectif de cette section est de proposer et de discuter les propriétés d'estimateurs classiques.

Exemple. On souhaite connaître la proportion de garçons à la naissance. La v.a. X qui modèle ce problème est une v.a. de Bernoulli de paramètre inconnu $p \in [0,1]$ (avec $\{X = 1\} =$« l'enfant est un garçon »). Pour estimer p, on observe toutes les naissances dans une maternité, et l'on note les résultats $X_1, \ldots, X_n$. Un estimateur intuitif de p est la proportion empirique de garçons, c'est-à-dire le nombre de garçons observés, divisé par le nombre total d'enfants :

$$\hat{p}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \,.$$

On a donc ici $f(x_1, \ldots, x_n) = (1/n) \sum_{i=1}^{n} x_i$.

L'exemple précédent est particulièrement simple, alors que l'exemple sur la durée de vie d'une ampoule est un peu plus complexe et l'on a moins d'intuition. De plus, il s'agit de savoir si l'estimateur proposé est « bon ». Pour répondre à cette question, il faut d'abord éclaircir ce que nous entendons par « bon estimateur ».

3.2 Qualités d'un estimateur

Souvent, le nombre d'observations n est grand au point qu'il est judicieux d'étudier les propriétés limites, ou asymptotiques, des estimateurs, quand n tend vers $+\infty$. C'est ici qu'interviennent les théorèmes limites introduits dans ce module. Dans ce paragraphe, on examine les qualités d'une suite d'estimateurs $(\hat{\theta}_n)_{n\in\mathbb{N}^*}$ définis par une suite de fonctions $f_n : \mathbb{R}^n \to \Theta$.

Exemple. On peut prendre l'exemple du nombre de garçons à la naissance et examiner la suite $\hat{p}_n = f_n(X_1, \ldots, X_n)$ définie par la suite de fonctions :

$$f_n(x_1, \ldots, x_n) := \frac{1}{n}\sum_{i=1}^{n} x_i\,.$$

Définition 4 (Consistance). On dit que $\hat{\theta}_n$ est *consistant* si, pour tout $\theta \in \Theta$:

$$\mathbb{P}_\theta\left(\hat{\theta}_n \overset{n\to+\infty}{\longrightarrow} \theta\right) = 1\,.$$

La propriété de consistance est essentielle. Elle dit que, si la taille de l'échantillon est suffisamment grande, alors la valeur donnée par l'estimateur devient proche de θ lorsque l'échantillon est tiré avec la loi $\mathbb{P}_\theta$. Il est important de réclamer que cette propriété soit vérifiée pour tous les $\theta \in \Theta$, car on ne connaît pas *a priori* la vraie valeur de θ.

Exemple. Dans l'exemple précédent, l'application de la loi forte des grands nombres prouve la consistance de l'estimateur proposé :

$$\mathbb{P}_p\left(\lim_{n\to+\infty} \hat{p}_n = p\right) = \mathbb{P}_p\left(\lim_{n\to+\infty} \frac{1}{n}\sum_{j=1}^{n} X_j = p\right) = 1\,.$$

Définition 5 (Biais).

- Un estimateur est dit *non biaisé* si $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$ pour tout $\theta \in \Theta$ et pour tout $n \in \mathbb{N}^*$.
- Un estimateur est dit *asymptotiquement non biaisé* si $\forall\theta \in \Theta \lim_{n\to+\infty} \mathbb{E}_\theta[\hat{\theta}_n] = \theta$.

On note ici et dans la suite $\mathbb{E}_\theta$ l'espérance prise sous la loi $\mathbb{P}_\theta$. Un estimateur non biaisé est évidemment asymptotiquement non biaisé.

Exemple. Dans l'exemple précédent, on utilise successivement la linéarité de l'espérance et le fait que l'espérance d'une v.a. de Bernoulli est égale à son paramètre pour écrire :

$$\mathbb{E}_p[\hat{p}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_p[X_i] = \mathbb{E}_p[X_1] = p\,.$$

L'estimateur $\hat{p}_n$ est donc non biaisé.

Cette qualité est souvent recherchée, mais insuffisante. Pour comparer deux estimateurs, on utilise le risque quadratique moyen défini comme suit.

Définition 6 (Risque quadratique moyen). Le *risque quadratique moyen* est la moyenne quadratique des écarts à la valeur à estimer :

$$\mathrm{RQM}_\theta(\hat{\theta}_n) := \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2]\,.$$

On peut donner une expression différente du risque quadratique moyen en développant le carré :

$$\begin{aligned}\mathrm{RQM}_\theta(\hat{\theta}_n) &= \mathbb{E}_\theta[(\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n] + \mathbb{E}_\theta[\hat{\theta}_n] - \theta)^2] \\ &= \mathbb{E}_\theta[(\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n])^2] + 2\mathbb{E}_\theta[\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n]](\mathbb{E}_\theta[\hat{\theta}_n] - \theta) + \Big(\mathbb{E}_\theta[\hat{\theta}_n] - \theta\Big)^2\,.\end{aligned}$$

Le second terme du membre de droite est nul car $\mathbb{E}_\theta[\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n]] = \mathbb{E}_\theta[\hat{\theta}_n] - \mathbb{E}_\theta[\hat{\theta}_n] = 0$. Le risque quadratique moyen se décompose donc en une partie « variance » et une partie « biais » :

$$\mathrm{RQM}_\theta(\hat{\theta}_n) = \mathrm{Var}_\theta(\hat{\theta}_n) + \Big(\mathbb{E}_\theta[\hat{\theta}_n] - \theta\Big)^2\,. \tag{7}$$

Si un estimateur est sans biais, son risque quadratique moyen est égal à sa variance. Entre deux estimateurs, on choisira celui au risque le plus faible. Ainsi, de deux estimateurs non biaisés, on choisira celui de variance plus faible.

Exemple.

Dans l'exemple précédent, on utilise la proposition 24 de la page 738 et l'on obtient :

$$\mathrm{RQM}_p[\hat{p}_n] = \mathrm{Var}_p\,(\hat{p}_n) = \frac{1}{n}\mathrm{Var}_p(X_1) = \frac{p(1-p)}{n}\,.$$

3.3 Estimateurs usuels

Les quantités les plus recherchées en statistique sont la moyenne et la variance. On va proposer et discuter dans la suite les estimateurs usuels de ces quantités.

Proposition 5 (Estimateur de la moyenne).
Soit $(X_i)_{i\in[\![1,n]\!]}$ un n-échantillon d'une v.a. réelle de moyenne μ et de variance σ^2.
La moyenne empirique $\overline{X}_n$ définie par :

$$\overline{X}_n := \frac{1}{n}\sum_{j=1}^{n} X_j\,,$$

est un estimateur de la moyenne qui satisfait les propriétés suivantes :
(i) $\overline{X}_n$ est non biaisé. *(ii)* $\mathrm{RQM}(\overline{X}_n) = \sigma^2/n$. *(iii)* $\overline{X}_n$ est consistant.

Démonstration. Il n'y a rien de nouveau dans cette proposition qui collecte des assertions déjà montrées dans ce qui précède. ∎

D'après la formule de Huygens, la variance d'une v.a. réelle X est donnée par :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

En copiant ce qu'on a fait pour l'estimateur de la moyenne, on peut proposer comme estimateur de $\mathbb{E}[X]$ la moyenne empirique $\overline{X}_n$, et comme estimateur du second moment $\mathbb{E}[X^2]$ la somme $\frac{1}{n}\sum_{j=1}^{n} X_j^2$. On obtient alors un estimateur de la variance, noté $\tilde{V}_n$:

$$\tilde{V}_n = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^2, \tag{8}$$

qu'on appelle variance empirique. Cette somme peut se réécrire différemment.

Proposition 6 (Estimateur de la variance). Soit $(X_i)_{i\in[\![1,n]\!]}$ un n-échantillon d'une v.a. réelle X de moyenne μ et de variance σ^2. La variance empirique de l'échantillon est définie par :

$$\tilde{V}_n := \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 . \tag{9}$$

Cet estimateur de la variance vérifie les propriétés suivantes.
(i) L'estimateur est biaisé :

$$\mathbb{E}[\tilde{V}_n] = \frac{n-1}{n}\sigma^2 .$$

(ii) L'estimateur est consistant.
(iii) Si la v.a. réelle X a un quatrième moment fini, alors le risque quadratique moyen de $\tilde{V}_n$ est d'ordre $1/n$:

$$\mathrm{RQM}(\tilde{V}_n) = \frac{\mu_4 - \sigma^4}{n} + \mathrm{O}\left(\frac{1}{n^2}\right) \cdot$$

où $\mu_4 = \mathbb{E}[X^4] - 4\mathbb{E}[X]\mathbb{E}[X^3] + 6\mathbb{E}[X^2]\mathbb{E}[X]^2 - 3\mathbb{E}[X]^4$.

Le nombre μ_4 est le moment centré d'ordre 4, dont la définition générale est, pour $p \in \mathbb{N}^*$:

$$\mu_p = \mathbb{E}\left[(X - \mathbb{E}[X])^p\right] .$$

En développant le polynôme $(X - \mathbb{E}[X])^p$, on peut aussi écrire que :

$$\mu_p = \sum_{j=0}^{p}(-1)^{p-j}\binom{p}{j}\mathbb{E}[X^j]\mathbb{E}[X]^{p-j} .$$

Noter que le second moment centré μ_2 est égal à la variance σ^2. Donc $\mu_4 - \sigma^4$ est la variance de la v.a. réelle $Y := (X - \mathbb{E}[X])^2$:

$$\mu_4 - \sigma^4 = \mathrm{Var}(Y) = \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] .$$

Démonstration. Pour montrer que (8) et (9) sont équivalents, on part de la seconde expression, et l'on développe les carrés :

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 &= \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \frac{2}{n}\sum_{j=1}^{n} X_j\overline{X}_n + \overline{X}_n^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - 2\overline{X}_n^2 + \overline{X}_n^2 \\ &= \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}_n^2 = \tilde{V}_n .\end{aligned}$$

On va maintenant calculer les deux premiers moments de cet estimateur. On commence par introduire les v.a. $\tilde{X}_i := X_i - \mu$, qui sont indépendantes et identiquement distribuées, de moyenne 0 et de variance σ^2. Alors :

$$\tilde{V}_n = \frac{1}{n}\sum_{j=1}^{n}(\tilde{X}_j + \mu)^2 - \left(\frac{1}{n}\sum_{j=1}^{n}\tilde{X}_j + \mu\right)^2 = \frac{1}{n}\sum_{j=1}^{n}\tilde{X}_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n}\tilde{X}_j\right)^2 .$$

Le $p^{ème}$ moment de $\tilde{V}_n$ se développe en somme d'espérances de la forme $\mathbb{E}[\tilde{X}_{i_1}\cdots\tilde{X}_{i_{2p}}]$. Il faut prendre soin de distinguer les indices i_j qui sont égaux, car, si une variable $\tilde{X}_{i_k}$ apparaît une seule fois dans le produit $\tilde{X}_{i_1}\cdots\tilde{X}_{i_{2p}}$, alors, par indépendance des variables $\tilde{X}_i$, l'espérance $\mathbb{E}[\tilde{X}_{i_1}\cdots\tilde{X}_{i_{2p}}] = \mathbb{E}[\tilde{X}_{i_k}]\mathbb{E}[\prod_{j\neq k}\tilde{X}_{i_j}]$ est nulle car $\mathbb{E}[\tilde{X}_{i_k}] = 0$. Ainsi, pour les deux premiers moments, on obtient :

$$\begin{aligned}
\mathbb{E}[\tilde{V}_n] &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\tilde{X}_j^2] - \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}[\tilde{X}_i\tilde{X}_j] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\tilde{X}^2] - \frac{1}{n^2}\left(\sum_{i\neq j}\mathbb{E}[\tilde{X}]^2 + \sum_{i=1}^{n}\mathbb{E}[\tilde{X}^2]\right) \\
&= \mathbb{E}[\tilde{X}^2] - \frac{1}{n}\mathbb{E}[\tilde{X}^2] = \frac{n-1}{n}\sigma^2 ,
\end{aligned}$$

et
$$\begin{aligned}
\mathbb{E}[\tilde{V}_n^2] &= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}[\tilde{X}_i^2\tilde{X}_j^2] - \frac{2}{n^3}\sum_{i,j,k=1}^{n}\mathbb{E}[\tilde{X}_i\tilde{X}_j\tilde{X}_k^2] + \frac{1}{n^4}\sum_{i,j,k,l=1}^{n}\mathbb{E}[\tilde{X}_i\tilde{X}_j\tilde{X}_k\tilde{X}_l] \\
&= \frac{1}{n}\left(\mathbb{E}[\tilde{X}^4] + (n-1)\mathbb{E}[\tilde{X}^2]^2\right) - \frac{2}{n^2}\left(\mathbb{E}[\tilde{X}^4] + (n-1)\mathbb{E}[\tilde{X}^2]^2\right) \\
&\quad + \frac{1}{n^3}\left(\mathbb{E}[\tilde{X}^4] + 3(n-1)\mathbb{E}[\tilde{X}^2]^2\right) \\
&= \frac{(n-1)^2}{n^3}\mathbb{E}[\tilde{X}^4] + \frac{(n-1)(n^2-2n+3)}{n^3}\mathbb{E}[\tilde{X}^2]^2,
\end{aligned}$$

qui peut être développé lorsque n tend vers $+\infty$:

$$\mathbb{E}[\tilde{V}_n^2] = \mathbb{E}[\tilde{X}^2]^2 + \frac{1}{n}\left(\mathbb{E}[\tilde{X}^4] - 3\mathbb{E}[\tilde{X}^2]^2\right) + \mathrm{O}\left(\frac{1}{n^2}\right) .$$

On peut alors calculer le risque quadratique moyen :

$$\mathrm{RQM}(\tilde{V}_n) = \mathbb{E}[\tilde{V}_n^2] - 2\sigma^2\mathbb{E}[\tilde{V}_n] + \sigma^4 = \frac{\mathbb{E}[\tilde{X}^4] - \mathbb{E}[\tilde{X}^2]^2}{n} + \mathrm{O}\left(\frac{1}{n^2}\right) .$$

En tenant compte du fait que $\tilde{X} = X - \mu$, on obtient le résultat annoncé. Comme on peut le constater, la preuve est calculatoire... Il existe une démonstration beaucoup moins calculatoire, mais celle-ci se fonde sur l'application d'une forme vectorielle du théorème central limite qui dépasse le cadre de ce cours.

Pour montrer la consistance, il faut se servir de l'expression (8) et utiliser la loi forte des grands nombres sur chacun des deux termes du membre de droite :

$$\tilde{V}_n = \frac{1}{n}\sum_{j=1}^{n}X_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n}X_j\right)^2 \overset{n\to+\infty}{\longrightarrow} \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 \text{ avec probabilité } 1 . \qquad \blacksquare$$

Il est facile de proposer un estimateur sans biais de la variance, il suffit de prendre la variance empirique $\tilde{V}_n$ et de la multiplier par $n/(n-1)$ (si $n \geqslant 2$). On obtient alors l'estimateur appelé variance empirique non biaisée dont les propriétés sont décrites dans la proposition suivante.

Proposition 7. Soit $(X_i)_{i\in[\![1,n]\!]}$ un n-échantillon (avec $n \geqslant 2$) d'une v.a. réelle X de moyenne μ et de variance σ^2. La variance empirique non biaisée V_n est définie par :

$$V_n := \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 \,. \tag{10}$$

C'est un estimateur de la variance qui satisfait les propriétés suivantes.
(i) V_n est non biaisé : pour tout $n \geqslant 2$, $\mathbb{E}[V_n] = \sigma^2$.
(ii) V_n est consistant.
(iii) Si X est de quatrième moment fini, alors le risque quadratique moyen de la variance empirique est d'ordre $1/n$:

$$\mathrm{RQM}(V_n) = \frac{\mu_4 - \sigma^4}{n} + \mathrm{O}\left(\frac{1}{n^2}\right) \cdot$$

Noter que les risques quadratiques moyens de V_n et de $\tilde{V}_n$ sont les mêmes à l'ordre $1/n$. C'est normal, car la différence entre ces deux estimateurs se situe au niveau du biais. Or le risque quadratique moyen d'ordre $1/n$ mesure une erreur d'ordre $1/\sqrt{n}$, il est donc complètement insensible au biais qui est plus petit, d'ordre $1/n$.

Démonstration.

Le premier point découle de la linéarité de l'espérance et du fait que $V_n = \frac{n}{n-1}\tilde{V}_n$.

Pour montrer le deuxième point, il faut utiliser l'expression :

$$V_n = \frac{n}{n-1} \times \left[\frac{1}{n}\sum_{j=1}^{n} X_j^2\right] - \frac{n}{n-1} \times \left[\frac{1}{n}\sum_{j=1}^{n} X_j\right]^2 .$$

On applique la loi forte des grands nombres à chacun des crochets. Avec probabilité 1, le premier crochet converge vers $\mathbb{E}[X^2]$, le second crochet converge vers $\mathbb{E}[X]$. Comme $n/(n-1)$ converge vers 1, on obtient que V_n converge vers $\mathbb{E}[X^2] - \mathbb{E}[X]^2$, ce qui donne la consistance.

Le troisième et dernier point découle du calcul suivant :

$$\begin{aligned}\mathrm{RQM}(V_n) &= \mathbb{E}[V_n^2] - 2\sigma^2\mathbb{E}[V_n] + \sigma^4 = \frac{n^2}{(n-1)^2}\mathbb{E}[\tilde{V}_n^2] - \sigma^4 \\ &= \frac{\mathbb{E}[X^4] - 4\mathbb{E}[X^3]\mathbb{E}[X] - \mathbb{E}[X^2]^2 + 8\mathbb{E}[X^2]\mathbb{E}[X]^2 - 4\mathbb{E}[X]^4}{n} + \mathrm{O}\left(\frac{1}{n^2}\right) .\end{aligned}$$

■

4 Intervalles de confiance

4.1 Intervalle de confiance et estimation

L'estimation d'un paramètre, même dans le cas d'un estimateur consistant, donnera presque sûrement une valeur différente de la vraie valeur inconnue. Ce que l'on peut dire, c'est que cette valeur inconnue est proche de la valeur estimée, mais tout l'art du statisticien est de quantifier cette erreur par nature aléatoire. Pour répondre rigoureusement au problème de l'estimation d'un paramètre, il est agréable de pouvoir construire, à partir de l'échantillon observé, un intervalle tel que le paramètre inconnu en fasse partie avec une grande probabilité donnée. On parlera alors d'intervalle de confiance I au niveau de confiance α si I ne dépend que de l'échantillon (et pas explicitement de θ) et si $\mathbb{P}_\theta(\theta \in I) \geqslant \alpha$. Pour obtenir des intervalles de confiance sur la moyenne et la variance d'un échantillon, on a besoin de connaître certaines propriétés statistiques des estimateurs obtenus à partir d'échantillons, telles la moyenne empirique, la variance empirique, la moyenne empirique renormalisée par la variance empirique, etc. Il est établi au paragraphe précédent que, avec probabilité 1 :

$$\overline{X}_n \overset{n\to+\infty}{\longrightarrow} \mu \text{ et } V_n \overset{n\to+\infty}{\longrightarrow} \sigma^2 ,$$

c'est-à-dire que la moyenne et la variance empiriques sont des estimateurs consistants de l'espérance μ et de la variance σ^2. Cela veut dire *grosso modo* que plus l'échantillon est grand, plus on a de chances d'obtenir une bonne estimation de μ ou σ^2.

4.2 Échantillons gaussiens

La proposition suivante caractérise la distribution statistique des moyenne et variance empiriques dans le cas d'un échantillon à statistique gaussienne.

Proposition 8.
Soit n un entier, avec $n \geqslant 2$. Soit $(X_i)_{i\in[\![1,n]\!]}$ un n-échantillon de loi $\mathcal{N}(\mu,\sigma^2)$, avec $\sigma > 0$.
Les v.a. réelles $\overline{X}_n$, V_n et le vecteur aléatoire $\hat{X}$ de coordonnées $\hat{X}_i = (X_i - \overline{X}_n)/\sqrt{V_n}$ pour $i \in [\![1,n]\!]$ sont indépendants. De plus, on a :

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1) , \tag{11}$$

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sqrt{V_n}} \sim T_{n-1} , \tag{12}$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2_n , \tag{13}$$

$$(n-1)\frac{V_n}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \sim \chi^2_{n-1} , \tag{14}$$

et la loi de $\hat{X}$ ne dépend pas de μ et σ.

Les lois χ^2_n et T_n sont décrites pages 782 et 785. Cette proposition donne donc les lois des estimateurs empiriques de la moyenne et de la variance pour tout n. Mais elle donne bien plus. Notons $\mathbb{S}^n := \{x \in \mathbb{R}^n, \sum_{i=1}^n x_i^2 = n-1\}$. L'application :

$$(X_1, \ldots, X_n) \in \mathbb{R}^n \mapsto (\overline{X}_n, V_n, \hat{X}) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{S}^n$$

est une bijection, donc toute l'information disponible dans l'échantillon est aussi dans le triplet $(\overline{X}_n, V_n, \hat{X})$. Or la proposition nous dit que $\hat{X}$ est indépendant de $(\overline{X}_n, V_n)$ et ne dépend pas des paramètres μ et σ. Il n'y a donc rien à en tirer du point de vue de l'estimation (de la moyenne et de la variance), le vecteur $\hat{X}$ est juste du « bruit ».

Démonstration. Introduisons les v.a. $Z_j = (X_j - \mu)/\sigma$.
Ce sont des v.a. gaussiennes indépendantes centrées réduites, et l'on peut exprimer le n-échantillon $(X_i)_{i \in [\![1,n]\!]}$ comme $X_j = \mu + \sigma Z_j$. D'après la proposition 23 de la page 779, la v.a. réelle :

$$Y_1 := \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{j=1}^n Z_j \tag{15}$$

a une loi gaussienne. Sa moyenne est 0 car les Z_j sont centrées, et sa variance est 1 car les Z_j sont indépendantes et de variance 1.
On trouve semblablement :

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \sum_{j=1}^n Z_j^2 \,,$$

ce qui montre que cette v.a. réelle suit une loi χ^2_n.
Considérons maintenant la variance empirique renormalisée (14) :

$$\begin{aligned}(n-1)\frac{V_n}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^n \left((\mu + \sigma Z_i) - (\mu + \frac{\sigma}{\sqrt{n}} Y_1) \right)^2 = \sum_{i=1}^n \left(Z_i - \frac{Y_1}{\sqrt{n}} \right)^2 \\ &= \sum_{i=1}^n Z_i^2 - Y_1^2 \,.\end{aligned}$$

Donnons-nous maintenant une base $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ de $\mathbb{R}^n$ dont le premier vecteur est le vecteur dont toutes les coordonnées valent $1/\sqrt{n}$, les autres vecteurs orthonormaux de la base étant choisis arbitrairement (par exemple, par le procédé d'orthogonalisation de Gram-Schmidt, voir le module « Espaces préhilbertiens » du livre [L2]). Appelons U la matrice $n \times n$ orthogonale de vecteurs lignes données par les $\mathbf{e}_j$, et Y le vecteur gaussien donné par $Y = UZ$. Noter que la première coordonnée Y_1 est bien donnée par (15). Le vecteur Y est un vecteur gaussien de moyenne nulle et de matrice de covariance $C = UU^t = I_n$, la matrice identité. D'après la proposition 27 de la page 781, la loi de Y a pour densité $p(y) = (2\pi)^{-n/2} \exp(-\|y\|^2/2)$. Cela veut dire que les Y_j sont des v.a. gaussiennes centrées réduites indépendantes. En utilisant une nouvelle fois le fait que

la matrice U est orthogonale, on a $\sum_{i=1}^{n} Z_i^2 = \sum_{j=1}^{n} Y_j^2$, et donc :

$$(n-1)\frac{V_n}{\sigma^2} = \sum_{j=2}^{n} Y_j^2\,,$$

où la somme va bien de 2 à n. La loi de cette v.a. réelle est donc un χ^2_{n-1}.

Le vecteur $\hat{X}$ a pour coordonnées :

$$\hat{X}_i = \frac{X_i - \overline{X}_n}{\sqrt{V_n}} = \frac{\mu + \sigma Z_i - (\mu + \frac{\sigma}{\sqrt{n}} Y_1)}{\frac{\sigma}{\sqrt{n-1}}\sqrt{\sum_{j=2}^{n} Y_j^2}} = \sqrt{n-1}\frac{Z_i - Y_1(\mathbf{e}_1)_i}{\sqrt{\sum_{j=2}^{n} Y_j^2}}\,,$$

donc le vecteur $U\hat{X}$ a pour coordonnées :

$$(U\hat{X})_i = \frac{\sqrt{n-1}}{\sqrt{\sum_{j=2}^{n} Y_j^2}}\big((UZ)_i - Y_1(U\mathbf{e}_1)_i\big) = \frac{\sqrt{n-1}}{\sqrt{\sum_{j=2}^{n} Y_j^2}}\big(Y_i - Y_1(U\mathbf{e}_1)_i\big)\,.$$

Or le vecteur $U\mathbf{e}_1$ a pour coordonnées $\langle \mathbf{e}_1, \mathbf{e}_j\rangle = 1$ si $j = 1$ et 0 sinon, donc

$$(U\hat{X})_i = \sqrt{n-1}\begin{cases} \dfrac{Y_i}{\sqrt{\sum_{j=2}^{n} Y_j^2}} & \text{si } i \geqslant 2\,, \\ 0 & \text{sinon}\,. \end{cases}$$

La loi de $U\hat{X}$ ne dépend donc pas de μ et σ. De plus, d'après la proposition 28 de la page 782, le vecteur $U\hat{X}$ et la v.a. $\sum_{j=2}^{n} Y_j^2$ sont indépendants.

Comme $\hat{X} = U^t U\hat{X}$ et V_n est une fonction de $\sum_{j=2}^{n} Y_j^2$, cela montre que $\hat{X}$ et V_n sont indépendants. De plus, ils ne dépendent que de $(Y_2, \ldots, Y_n)$, alors que $\overline{X}_n$ ne dépend que de Y_1. Donc $\overline{X}_n$, V_n et $\hat{X}$ sont indépendants.

Enfin, comme $(n-1)\sigma^{-2}V_n \sim \chi^2_{n-1}$, et $\sqrt{n}\sigma^{-1}(\overline{X}_n - \mu) \sim \mathcal{N}(0,1)$ sont indépendantes, la v.a. réelle :

$$\sqrt{n-1}\frac{\sqrt{n}\sigma^{-1}(\overline{X}_n - \mu)}{\sqrt{n-1}\sigma^{-1}\sqrt{V_n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{V_n}}$$

suit une loi T_{n-1}. ∎

Corollaire 9. Soit $(X_i)_{i\in[\![1,n]\!]}$ un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. Les risques quadratiques moyens des estimateurs empiriques de la moyenne et de la variance sont :

$$\mathrm{RQM}(\overline{X}_n) = \frac{\sigma^2}{n}\,, \qquad \mathrm{RQM}(V_n) = \frac{2\sigma^4}{n-1}\,.$$

Démonstration. Le risque quadratique moyen de la moyenne empirique est déjà connu par la proposition 5. Celui de la variance empirique se calcule à partir de la caractérisation (14)

et du fait qu'une v.a. suivant la loi χ_p^2 a pour variance $2p$ (voir l'exercice 4 de la page 784).

■

4.2.1 Estimation par intervalle de confiance de la moyenne

On observe un n-échantillon d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, et l'on cherche à estimer la moyenne μ. On peut se servir de la moyenne empirique $\overline{X}_n$, dont on sait qu'elle va donner une estimation bonne lorsque n est grand. En pratique, on cherche un intervalle de confiance, c'est-à-dire un intervalle qui est construit à partir de l'échantillon dont on puisse dire qu'il contient la valeur inconnue μ avec une probabilité que l'on se fixe.
Supposons dans un premier temps que l'écart-type σ de la loi soit connu. Soit un niveau de confiance α donné (en général $\alpha = 0{,}9$, $0{,}95$, ou $0{,}99$). Soit t_α le nombre tel que si $Z \sim \mathcal{N}(0,1)$, alors $\mathbb{P}(-t_\alpha \leqslant Z \leqslant t_\alpha) = \alpha$. On note toujours Φ la fonction de répartition normale centrée réduite. Par parité de la densité de la gaussienne, $\Phi(x) = 1 - \Phi(-x)$ et la condition sur t_α s'écrit $\Phi(t_\alpha) - \Phi(-t_\alpha) = 2\Phi(t_\alpha) - 1 = \alpha$ soit $\Phi(t_\alpha) = (1+\alpha)/2$. On cherche alors t_α dans une table de la loi normale ou par un logiciel grâce à la fonction quantile de la loi normale standard. D'après la proposition 8 :

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \in [-t_\alpha, t_\alpha]\right) = \alpha\,.$$

En réécrivant l'événement en question :

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \in [-t_\alpha, t_\alpha] \Longleftrightarrow \mu \in \left[\overline{X}_n - \frac{\sigma t_\alpha}{\sqrt{n}}, \overline{X}_n + \frac{\sigma t_\alpha}{\sqrt{n}}\right]\,,$$

on obtient l'intervalle de confiance I_α au niveau α :

$$I_\alpha = \left[\overline{X}_n - \frac{\sigma t_\alpha}{\sqrt{n}}, \overline{X}_n + \frac{\sigma t_\alpha}{\sqrt{n}}\right]\,,$$

dont on sait que $\mathbb{P}(\mu \in I_\alpha) = \alpha$. Bien sûr :

- plus on se fixe un niveau de confiance α élevé, plus l'intervalle est large. Si on demande d'être absolument sûr, i.e. $\alpha = 1$, alors $I_\alpha = \mathbb{R}$.
- plus la v.a. sous-jacente a une dispersion élevée (i.e. plus σ est grand), plus l'intervalle est grand.
- plus l'échantillon est grand, plus l'intervalle est petit.

La largeur de l'intervalle de confiance est proportionnelle à $1/\sqrt{n}$. Donc, à niveau α fixé, pour avoir un intervalle de confiance 2 fois plus petit, il faut 4 fois plus d'observations.
Les calculs précédents sont intéressants, mais en pratique, il est rare que l'on connaisse σ. Dans le cas général où l'on ne connaît pas σ, on utilise une estimation de celui-ci, donnée par la variance empirique sans biais V_n.
Soit un niveau de confiance α donné. Soit t_α le nombre tel que si $Z \sim T_{n-1}$, alors $\mathbb{P}(-t_\alpha \leqslant Z \leqslant t_\alpha) = \alpha$. D'après la proposition 8 :

$$\mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n - \mu}{\sqrt{V_n}} \in [-t_\alpha, t_\alpha]\right) = \alpha\,.$$

En réécrivant l'événement en question, on obtient l'intervalle de confiance I_α au niveau α :

$$I_\alpha = \left[\overline{X}_n - \frac{\sqrt{V_n}t_\alpha}{\sqrt{n}}, \overline{X}_n + \frac{\sqrt{V_n}t_\alpha}{\sqrt{n}}\right],$$

dont on sait que $\mathbb{P}(\mu \in I_\alpha) = \alpha$.
L'intervalle de confiance obtenu ainsi a la même comportement qualitatif que celui décrit dans le cas σ connu. Il n'y a rien d'étonnant à cela, car on sait que la variance empirique converge avec probabilité 1 vers la variance théorique. Cela se traduit aussi par le fait que, lorsque n est grand, la loi de Student à n degrés de libertés devient très proche de la loi gaussienne centrée réduite. En pratique, on utilise des tables de loi T_n (voir la table IV.4.2 de la page suivante) pour $n \leqslant 30$, et, lorsque $n > 30$, on utilise l'approximation gaussienne.

Exemple. On mesure les durées de vie de $n = 10$ ampoules. On obtient :

$$1\,864,\ 1\,934,\ 2\,033,\ 1\,890,\ 1\,997,\ 1\,974,\ 1\,837,\ 1\,903,\ 2\,009,\ 1\,950.$$

On cherche la durée de vie moyenne μ d'une ampoule, en supposant que la loi décrivant cette durée de vie est une gaussienne. La moyenne et la variance empirique sont :

$$\bar{X}_{10} \simeq 1\,939, \qquad V_{10} \simeq 4\,244.$$

Un intervalle de confiance pour la moyenne au niveau $95\ \%$ est donc :

$$I_{0,95} = \left[\bar{X}_{10} - \frac{\sqrt{V_{10}}t_{0,95}}{\sqrt{10}}, \bar{X}_{10} + \frac{\sqrt{V_{10}}t_{0,95}}{\sqrt{10}}\right],$$

où $t_{0,95}$ est tel que $\mathbb{P}(-t_{0,95} \leqslant Z \leqslant t_{0,95}) = 0{,}95$, avec $Z \sim T_9$.
En utilisant une table de la loi T_9, on trouve $t_{0,95} \simeq 2{,}26$, et l'on obtient ainsi un intervalle de confiance au niveau 95% de la valeur inconnue μ :

$$I_{0,95} = [1\,892,\ 1\,986].$$

4.2.2 Estimation par intervalle de confiance de la variance

Les raisonnements que l'on vient d'appliquer pour l'estimation de la moyenne peuvent être repris pour l'estimation de la variance σ^2, lorsque la moyenne μ est connue, puis lorsqu'elle ne l'est pas.
Commençons par le cas où la moyenne μ de l'échantillon est connue. Donnons-nous le niveau de confiance α. On commence par choisir $t_{\alpha,0}$ et $t_{\alpha,1}$ tels que, si $Z \sim \chi^2_n$, alors $\mathbb{P}(t_{\alpha,0} \leqslant Z \leqslant t_{\alpha,1}) = \alpha$. La loi χ^2_n n'est pas symétrique et il y a une infinité de couples possibles $(t_{\alpha,0}, t_{\alpha,1})$. On choisit en général les deux extrémités $t_{\alpha,0}$ et $t_{\alpha,1}$ de sorte que le risque soit également réparti à gauche et à droite, i.e. $\mathbb{P}(Z < t_{\alpha,0}) = (1-\alpha)/2$ et $\mathbb{P}(Z > t_{\alpha,1}) = (1-\alpha)/2$. Mais on pourrait parfaitement proposer un intervalle asymétrique. On utilise en pratique des tables de la loi χ^2_n (voir la table IV.4.3 de la page 823) ou des logiciels de calcul scientifique. Lorsque la moyenne μ est connue, l'estimateur empirique de la variance est :

$$V_n^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

$n \backslash \alpha$	0,40	0,25	0,10	0,05	0,025	0,01	0,005	0,0005
1	0,324920	1,000000	3,077684	6,313752	12,70620	31,82052	63,65674	636,6192
2	0,288675	0,816497	1,885618	2,919986	4,30265	6,96456	9,92484	31,5991
3	0,276671	0,764892	1,637744	2,353363	3,18245	4,54070	5,84091	12,9240
4	0,270722	0,740697	1,533206	2,131847	2,77645	3,74695	4,60409	8,6103
5	0,267181	0,726687	1,475884	2,015048	2,57058	3,36493	4,03214	6,8688
6	0,264835	0,717558	1,439756	1,943180	2,44691	3,14267	3,70743	5,9588
7	0,263167	0,711142	1,414924	1,894579	2,36462	2,99795	3,49948	5,4079
8	0,261921	0,706387	1,396815	1,859548	2,30600	2,89646	3,35539	5,0413
9	0,260955	0,702722	1,383029	1,833113	2,26216	2,82144	3,24984	4,7809
10	0,260185	0,699812	1,372184	1,812461	2,22814	2,76377	3,16927	4,5869
11	0,259556	0,697445	1,363430	1,795885	2,20099	2,71808	3,10581	4,4370
12	0,259033	0,695483	1,356217	1,782288	2,17881	2,68100	3,05454	4,3178
13	0,258591	0,693829	1,350171	1,770933	2,16037	2,65031	3,01228	4,2208
14	0,258213	0,692417	1,345030	1,761310	2,14479	2,62449	2,97684	4,1405
15	0,257885	0,691197	1,340606	1,753050	2,13145	2,60248	2,94671	4,0728
16	0,257599	0,690132	1,336757	1,745884	2,11991	2,58349	2,92078	4,0150
17	0,257347	0,689195	1,333379	1,739607	2,10982	2,56693	2,89823	3,9651
18	0,257123	0,688364	1,330391	1,734064	2,10092	2,55238	2,87844	3,9216
19	0,256923	0,687621	1,327728	1,729133	2,09302	2,53948	2,86093	3,8834
20	0,256743	0,686954	1,325341	1,724718	2,08596	2,52798	2,84534	3,8495
21	0,256580	0,686352	1,323188	1,720743	2,07961	2,51765	2,83136	3,8193
22	0,256432	0,685805	1,321237	1,717144	2,07387	2,50832	2,81876	3,7921
23	0,256297	0,685306	1,319460	1,713872	2,06866	2,49987	2,80734	3,7676
24	0,256173	0,684850	1,317836	1,710882	2,06390	2,49216	2,79694	3,7454
25	0,256060	0,684430	1,316345	1,708141	2,05954	2,48511	2,78744	3,7251
26	0,255955	0,684043	1,314972	1,705618	2,05553	2,47863	2,77871	3,7066
27	0,255858	0,683685	1,313703	1,703288	2,05183	2,47266	2,77068	3,6896
28	0,255768	0,683353	1,312527	1,701131	2,04841	2,46714	2,76326	3,6739
29	0,255684	0,683044	1,311434	1,699127	2,04523	2,46202	2,75639	3,6594
30	0,255605	0,682756	1,310415	1,697261	2,04227	2,45726	2,75000	3,6460
$+\infty$	0,253347	0,674490	1,281552	1,644854	1,95996	2,32635	2,57583	3,2905

TABLE IV.4.2 – Table de la loi T_n.
On reporte dans ce tableau les valeurs t_α pour lesquelles $\mathbb{P}(T_n \geqslant t_\alpha) = \alpha$.
Pour $n > 30$, on adopte l'approximation gaussienne $T_n \sim \mathcal{N}(0,1)$.

L'estimateur V_n^* est sans biais, consistant, et nV_n^*/σ^2 suit la loi χ_n^2 d'après la proposition 8. On peut donc affirmer que :

$$\mathbb{P}\left(\frac{nV_n^*}{\sigma^2} \in [t_{\alpha,0}, t_{\alpha,1}]\right) = \alpha\,.$$

En réécrivant l'événement en question, on trouve un intervalle de confiance au niveau α de la variance :

$$I_\alpha = \left[\frac{nV_n^*}{t_{\alpha,1}}, \frac{nV_n^*}{t_{\alpha,0}}\right]\,,$$

dont on sait que $\mathbb{P}(\sigma^2 \in I_\alpha) = \alpha$.

Exemple. Dans l'exemple précédent (durée de vie d'une ampoule), imaginons que le fabricant ait fait des mesures extensives, et qu'il indique sur la boîte la durée de vie moyenne : $\mu = 1\,920$. On recherche la variance σ^2 inconnue de la loi de durée de vie. L'estimateur empirique de la variance est :

$$V_{10}^* = \frac{1}{10}\sum_{i=1}^{10}(X_i - 1\,920)^2 \simeq 4\,184\,.$$

On cherche un intervalle de confiance au niveau $\alpha = 95\ \%$ de la variance à partir de l'échantillon de 10 ampoules observées. Cet intervalle est :

$$I_{0,95} = \left[\frac{10V_{10}^*}{t_{0,95,1}}, \frac{10V_{10}^*}{t_{0,95,0}}\right]\,.$$

Il faut évaluer les deux nombres $t_{0,95,0}$ et $t_{0,95,1}$. Pour cela, on consulte une table de la fonction de répartition de la loi χ_{10}^2 et l'on cherche les niveaux $t_{\alpha,0}$ et $t_{\alpha,1}$ tels que, si $Z \sim \chi_{10}^2$, alors $\mathbb{P}(Z < t_{\alpha,0}) = (1-\alpha)/2 = 0{,}025$ et $\mathbb{P}(Z < t_{\alpha,1}) = 1 - (1-\alpha)/2 = 0{,}975$. On trouve $t_{\alpha,0} = 3{,}25$ et $t_{1,\alpha} = 20{,}48$. On obtient alors l'intervalle de confiance :

$$I_{0,95} = [2\,043,\ 12\,875]\,.$$

Noter que l'intervalle de confiance pour la variance est beaucoup plus large que pour l'espérance. Il est en effet plus difficile d'estimer la variance que la moyenne.

Dans le cas d'un échantillon de taille n supérieure à 30, on peut admettre que, si $Z \sim \chi_n^2$, alors la v.a. $\sqrt{2}\sqrt{Z} - \sqrt{2n-1}$ suit la loi normale centrée réduite. Ce résultat est très proche de celui donné par l'exercice de la page 803, mais des calculs plus précis montrent que la première approximation est en fait plus fine, la seconde nécessitant des n plus grands pour être valable avec la même précision. Pour des échantillons de taille n supérieure à 30, on utilise cette approximation gaussienne pour déterminer les quantiles de la loi χ_n^2. C'est pourquoi les tables de la loi χ_n^2 s'arrêtent en général à $n = 30$. Bien sûr, si on utilise un logiciel de mathématiques ou de statistique, on peut trouver des valeurs approchées des quantiles de la loi χ_n^2 valables pour des n plus grands, avec une précision arbitraire.

$n \backslash \alpha$	0,995	0,990	0,975	0,950	0,050	0,025	0,010	0,005
1	0,00004	0,00016	0,00098	0,00393	3,84146	5,02389	6,63490	7,87944
2	0,01003	0,02010	0,05064	0,10259	5,99146	7,37776	9,21034	10,59663
3	0,07172	0,11483	0,21580	0,35185	7,81473	9,34840	11,34487	12,83816
4	0,20699	0,29711	0,48442	0,71072	9,48773	11,14329	13,27670	14,86026
5	0,41174	0,55430	0,83121	1,14548	11,07050	12,83250	15,08627	16,74960
6	0,67573	0,87209	1,23734	1,63538	12,59159	14,44938	16,81189	18,54758
7	0,98926	1,23904	1,68987	2,16735	14,06714	16,01276	18,47531	20,27774
8	1,34441	1,64650	2,17973	2,73264	15,50731	17,53455	20,09024	21,95495
9	1,73493	2,08790	2,70039	3,32511	16,91898	19,02277	21,66599	23,58935
10	2,15586	2,55821	3,24697	3,94030	18,30704	20,48318	23,20925	25,18818
11	2,60322	3,05348	3,81575	4,57481	19,67514	21,92005	24,72497	26,75685
12	3,07382	3,57057	4,40379	5,22603	21,02607	23,33666	26,21697	28,29952
13	3,56503	4,10692	5,00875	5,89186	22,36203	24,73560	27,68825	29,81947
14	4,07467	4,66043	5,62873	6,57063	23,68479	26,11895	29,14124	31,31935
15	4,60092	5,22935	6,26214	7,26094	24,99579	27,48839	30,57791	32,80132
16	5,14221	5,81221	6,90766	7,96165	26,29623	28,84535	31,99993	34,26719
17	5,69722	6,40776	7,56419	8,67176	27,58711	30,19101	33,40866	35,71847
18	6,26480	7,01491	8,23075	9,39046	28,86930	31,52638	34,80531	37,15645
19	6,84397	7,63273	8,90652	10,11701	30,14353	32,85233	36,19087	38,58226
20	7,43384	8,26040	9,59078	10,85081	31,41043	34,16961	37,56623	39,99685
21	8,03365	8,89720	10,28290	11,59131	32,67057	35,47888	38,93217	41,40106
22	8,64272	9,54249	10,98232	12,33801	33,92444	36,78071	40,28936	42,79565
23	9,26042	10,19572	11,68855	13,09051	35,17246	38,07563	41,63840	44,18128
24	9,88623	10,85636	12,40115	13,84843	36,41503	39,36408	42,97982	45,55851
25	10,51965	11,52398	13,11972	14,61141	37,65248	40,64647	44,31410	46,92789
26	11,16024	12,19815	13,84390	15,37916	38,88514	41,92317	45,64168	48,28988
27	11,80759	12,87850	14,57338	16,15140	40,11327	43,19451	46,96294	49,64492
28	12,46134	13,56471	15,30786	16,92788	41,33714	44,46079	48,27824	50,99338
29	13,12115	14,25645	16,04707	17,70837	42,55697	45,72229	49,58788	52,33562
30	13,78672	14,95346	16,79077	18,49266	43,77297	46,97924	50,89218	53,67196

TABLE IV.4.3 – Table de la loi χ^2_n.
On reporte dans ce tableau les valeurs t_α pour lesquelles $\mathbb{P}(\chi^2_n \geqslant t_\alpha) = \alpha$.
Pour $n > 30$, on adopte l'approximation gaussienne $\sqrt{2\chi^2_n} - \sqrt{2n-1} \sim \mathcal{N}(0,1)$.

Supposons maintenant que la moyenne μ soit inconnue. On commence par choisir $t_{\alpha,0}$ et $t_{\alpha,1}$ tels que, si $Z \sim \chi^2_{n-1}$, alors $\mathbb{P}(t_{\alpha,0} \leqslant Z \leqslant t_{\alpha,1}) = \alpha$. Lorsque la moyenne μ est inconnue, l'estimateur empirique de la variance est :

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 .$$

V_n est sans biais, consistant, et $(n-1)V_n/\sigma^2$ suit la loi χ^2_{n-1} d'après la proposition 8. Un intervalle de confiance au niveau α est donc :

$$I_\alpha = \left[\frac{(n-1)V_n}{t_{\alpha,1}}, \frac{(n-1)V_n}{t_{\alpha,0}}\right] .$$

Exemple. On reprend l'exemple sur les ampoules, mais sans supposer qu'on connaît l'espérance μ. L'estimateur empirique de la variance est alors :

$$V_{10} = \frac{1}{9}\sum_{i=1}^{10}(X_i - \bar{X}_{10})^2 \simeq 4\,244 .$$

On cherche un intervalle de confiance au niveau $\alpha = 95\ \%$:

$$I_{0,95} = \left[\frac{9V_{10}}{t_{0,95,1}}, \frac{9V_{10}}{t_{0,95,0}}\right] .$$

On commence par consulter une table de la fonction de répartition de la loi χ^2_9 et l'on cherche les niveaux $t_{0,95,0}$ et $t_{0,95,1}$ tels que, si $Z \sim \chi^2_9$, alors

$$\mathbb{P}(Z < t_{0,95,0}) = (1 - 0{,}95)/2 = 0{,}025$$
$$\mathbb{P}(Z < t_{0,95,1}) = 1 - (1 - 0{,}95)/2 = 0{,}975.$$

On trouve $t_{0,95,0} = 2{,}70$ et $t_{1,0,95} = 19{,}02$. On obtient alors l'intervalle de confiance :

$$I_{0,95} = [2\,008,\ 14\,147] .$$

Comme on pouvait s'y attendre, l'intervalle est un peu plus grand que dans le cas où l'on connaît l'espérance.

4.3 Échantillons non gaussiens

4.3.1 Résultats asymptotiques

Dans le cas non gaussien, la proposition 8 de la page 816 n'est plus vraie, mais en vertu du théorème de la limite centrale, on peut obtenir des résultats similaires sous forme asymptotique. On obtient les distributions de la moyenne et de la variance empirique « pour n assez grand ». Dans la pratique, on admettra ces résultats lorsque $n \geqslant 30$. Des résultats fins permettent de contrôler l'erreur ainsi commise, mais ils sortent du cadre de ce cours. Nous nous contenterons de donner une application particulièrement importante de cette méthode pour les intervalles de confiance pour une proportion, c'est-à-dire les sondages.

4.3.2 Sondages

À la veille du second tour d'une élection présidentielle, on effectue un sondage afin de déterminer la proportion p de votes pour le candidat H. On pourrait considérer plus généralement tout problème de sondage avec réponses binaires. Le sondage porte sur $n = 2\,500$ individus choisis au hasard dans le corps électoral. On note $X_i = 1$ si le $i^{\text{ème}}$ individu interrogé vote pour H, $X_i = 0$ sinon. En pratique, on évite d'interroger deux fois un même individu (tirage sans remise), de sorte que les X_i sont dépendants. Mais lorsque la taille n de l'échantillon est faible comparée à la population totale, il y a peu de différence avec le tirage avec remise, et l'on se placera dans cette hypothèse simplificatrice (voir la section 4.1 à la page 727). Les v.a. X_i, $i \in [\![1, n]\!]$, sont supposées indépendantes et de même loi de Bernoulli $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$. Nous nous trouvons dans le problème statistique suivant : comment estimer le paramètre p inconnu au vu des observations $X_1, \ldots, X_n$? Intuitivement, aucune information sur p n'est contenue dans l'ordre des réponses, et il suffit de résumer le sondage $X_1, \ldots, X_n$ par le nombre d'intentions de vote $S_n := X_1 + \cdots + X_n$. Ce raisonnement heuristique peut d'ailleurs se quantifier par le résultat qui suit.

Exercice 4.

Soit $k \in [\![0, n]\!]$. Pour tout $(x_1, \ldots, x_n) \in \{0, 1\}^n$, on a :

$$\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) | S_n = k) = \begin{cases} \frac{1}{\binom{n}{k}} & \text{si } (x_1, \ldots, x_n) \in E_{n,k}\,, \\ 0 & \text{sinon}\,, \end{cases}$$

où $E_{n,k}$ représente l'ensemble des suites $(x_1, \ldots, x_n)$ comportant k uns et $n - k$ zéros. On dit que la loi conditionnelle de $(X_1, \ldots, X_n)$ sachant $S_n = k$ est la probabilité uniforme sur les $\binom{n}{k}$ suites $(x_1, \ldots, x_n)$ comportant k uns et $n - k$ zéros.

Solution. Soit $(x_1, \ldots, x_n) \in \{0, 1\}^n$. Par définition d'une probabilité conditionnelle :

$$\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) | S_n = k) = \frac{\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \cap S_n = k)}{\mathbb{P}(S_n = k)}\,.$$

Comme S_n suit une loi binomiale de paramètres (n, p) :

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1 - p)^{n-k}\,.$$

Si $(x_1, \ldots, x_n) \notin E_{n,k}$, alors la somme des x_i ne vaut pas k, et donc :

$$\mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \cap S_n = k) = 0\,.$$

Si $(x_1, \ldots, x_n) \in E_{n,k}$, alors la somme des x_i vaut k, et :

$$\begin{aligned} \mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \cap S_n = k) &= \mathbb{P}((X_1, \ldots, X_n) = (x_1, \ldots, x_n)) \\ &= p^k (1 - p)^{n-k}\,. \end{aligned}$$

En injectant dans la probabilité conditionnelle, on obtient le résultat cherché.

L'exercice précédent montre que la loi conditionnelle des observations $(X_1, \ldots, X_n)$ sachant $S_n = k$ ne dépend pas du paramètre inconnu p. Par conséquent la valeur de S_n

contient toute l'information sur p contenue dans le sondage. On cherche alors à estimer p par une fonction $\hat{p}_n$ de S_n, et le choix naturel est la proportion $\hat{p}_n$ de vote pour H dans l'échantillon :

$$\hat{p}_n = \frac{S_n}{n} \cdot$$

Cet estimateur est sans biais $\mathbb{E}[\hat{p}_n] = p$, c'est-à-dire qu'il est « bon » en moyenne. Cet estimateur est consistant, c'est-à-dire qu'il est « bon » lorsque l'échantillon est de grande taille. En effet, la loi des grands nombres nous apprend que :

$$\hat{p}_n \overset{n\to+\infty}{\longrightarrow} p\,,$$

en probabilité et avec probabilité 1 : l'estimateur converge vers le paramètre à estimer. Le sondage donne 1 300 intentions de votes pour H, et 1 200 pour son adversaire. L'estimateur $\hat{p}_n$ prend la valeur $0{,}52$ mais est-on « sûr » pour autant que H sera élu ? Plus précisément, cette valeur est-elle significativement supérieure à $0{,}5$? La réponse dépend de l'amplitude des fluctuations de $\hat{p}_n$ autour de p, et nous utilisons l'approximation gaussienne pour quantifier ces fluctuations (théorème de la limite centrale de la page 803) :

$$\mathbb{P}\left(|\hat{p}_n - p| \leqslant a\sigma\right) \simeq \int_{-a}^{a} g(x)\,\mathrm{d}x\,.$$

L'erreur commise $\hat{p}_n - p$ ne dépassera pas, avec probabilité $95\ \%$ [respectivement $99\ \%$], le seuil :

$$\frac{1{,}96\sqrt{p(1-p)}}{\sqrt{n}}, \quad \text{respectivement} \quad \frac{2{,}6\sqrt{p(1-p)}}{\sqrt{n}}\,.$$

Ce seuil dépend malencontreusement du paramètre p inconnu, mais la fonction $\sqrt{p(1-p)}$ est majorée par sa valeur maximale $1/2$, de sorte qu'en remplaçant le facteur $\sqrt{p(1-p)}$ par $1/2$ dans les seuils précédents, on ne fera qu'augmenter le niveau de confiance. On peut maintenant reformuler le résultat précédent sous sa forme la plus couramment utilisée (en remplaçant $1{,}96$ par 2).

Proposition 10 (Intervalle de confiance pour l'estimation de p).
Dès que n est assez grand (np et $n(1-p) \geqslant 10$ en pratique) :

$$\mathbb{P}\left(p \in \left[\hat{p}_n - \frac{1}{\sqrt{n}}\,,\ \hat{p}_n + \frac{1}{\sqrt{n}}\right]\right) = 0{,}95\,,$$

$$\mathbb{P}\left(p \in \left[\hat{p}_n - \frac{1{,}3}{\sqrt{n}}\,,\ \hat{p}_n + \frac{1{,}3}{\sqrt{n}}\right]\right) = 0{,}99\,.$$

Dans notre exemple, l'intervalle de confiance à $95\ \%$ pour p est $[0{,}50, 0{,}54]$. Les instituts de sondage annoncent d'ailleurs leurs résultats ainsi (sous forme de fourchettes).

4.3.3 Calcul d'intégrale par la méthode de Monte-Carlo

Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction continue par morceaux intégrable. Lorsque le calcul analytique de $I := \displaystyle\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x$ n'est pas possible, il existe diverses méthodes d'intégration

numérique. Des méthodes de quadrature déterministe sont couramment utilisées. Nous décrivons ici une méthode probabiliste basée sur la représentation suivante de I. Soit p une densité de probabilité sur $\mathbb{R}^d$, dont le support contient celui de f ($f(x) \neq 0 \Rightarrow p(x) > 0$). Alors :

$$I = \int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \frac{f(x)}{p(x)} p(x)\,\mathrm{d}x\,,$$

et I s'écrit comme :

$$I = \mathbb{E}[\psi(X)] \text{ avec } \psi(x) := \begin{cases} f/p(x) \text{ si } p(x) > 0 \\ 0 \text{ si } p(x) = 0 \end{cases}, \ X \text{ de densité } p\,.$$

Si p est la densité d'une v.a. facilement simulable, on sait générer des v.a. $X_1, \ldots, X_n$ indépendantes et de densité p. La quantité :

$$\hat{I}_n := \frac{1}{n}\left[\psi(X_1) + \cdots + \psi(X_n)\right]$$

constitue une approximation, en fait, une estimation au sens statistique du terme, de I. Cette méthode de calcul (approximatif) est appelée méthode de Monte-Carlo. Le nom de la méthode fait référence aux jeux de hasard pratiqués à Monte-Carlo.

L'estimateur $\hat{I}_n$ est en fait la moyenne empirique de la v.a. réelle $\psi(X)$, qui a pour moyenne I. La proposition 5 de la page 812 donne les propriétés importantes de cet estimateur : il est sans biais (par linéarité de la moyenne), consistant (par application de la loi des grands nombres), son risque quadratique moyen est $\mathrm{RQM}(\hat{I}_n) = \sigma^2/n$, où $\sigma^2 := \mathrm{Var}\big(\psi(X)\big)$, qui est fini lorsque $\mathbb{E}[\psi(X)^2] < +\infty$. Cette variance est donnée par :

$$\begin{aligned} \sigma^2 = \mathbb{E}[\psi(X)^2] - \mathbb{E}[\psi(X)]^2 &= \int_{\mathbb{R}^d} \psi(x)^2 p(x)\,\mathrm{d}x - I^2 \\ &= \int_{\mathbb{R}^d} \frac{f(x)^2}{p(x)}\,\mathrm{d}x - \left(\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x\right)^2 . \end{aligned} \tag{16}$$

Un point remarquable de la méthode est que la variance σ^2, et donc le risque quadratique moyen, dépend explicitement du choix de la densité p. Comme il y a une infinité de choix possibles pour la densité p, tout l'art du praticien est de bien choisir la densité pour « réduire la variance ». Les techniques de réduction de variance pour les méthodes de Monte Carlo font l'objet d'intenses recherches.

Le théorème de la limite centrale nous renseigne, lorsque $\mathbb{E}[\psi(X)^2] < +\infty$, sur la distribution de l'erreur $\hat{I}_n - I$. En procédant comme dans le paragraphe précédent, on obtient un intervalle de confiance pour I à la quasi-certitude $95\,\%$ de la forme :

$$\left[\hat{I}_n - \frac{1{,}96\sigma}{\sqrt{n}}\,,\ \hat{I}_n + \frac{1{,}96\sigma}{\sqrt{n}}\right].$$

Ici, la variance σ^2 définie par (16) est le plus souvent inconnue, de sorte qu'il faut estimer sa valeur elle aussi. On définit alors :

$$\hat{\sigma}_n := \left(\frac{1}{n}\sum_{i=1}^{n} \psi(X_i)^2 - \hat{I}_n^2\right)^{1/2},$$

qui vérifie $\hat{\sigma}_n \to \sigma$ quand n tend vers $+\infty$ d'après la loi forte des grands nombres, et l'on utilise alors l'intervalle :

$$\left[\hat{I}_n - \frac{1{,}96\hat{\sigma}_n}{\sqrt{n}}\ ,\ \hat{I}_n + \frac{1{,}96\hat{\sigma}_n}{\sqrt{n}}\right]\ , \tag{17}$$

qui est un intervalle de confiance asymptotique pour I à $95\,\%$ (i.e. valable pour n suffisamment grand). L'intervalle (17) indique l'erreur de la méthode de Monte-Carlo. La méthode s'applique aussi en dimension supérieure, i.e. si $f : \mathbb{R}^d \to \mathbb{R}^k$, mais la formule (17) se complique alors. Comparée aux méthodes d'intégration numériques usuelles, par quadrature, la méthode de Monte-Carlo est surtout intéressante en dimension d grande.

Exercice type corrigé

IV.4.0 Cet exercice est classique et s'appelle le collecteur de coupons.
Votre petit frère collectionne les images des joueurs de la coupe du monde que l'on trouve dans les tablettes de chocolat. On suppose qu'il existe N images différentes, et qu'elles sont équitablement distribuées dans l'ensemble des tablettes (que l'on supposera infini). On note T_k le nombre de tablettes qu'il faut acheter pour avoir k images différentes. On s'intéresse en particulier à T_N.

1. Montrer que les accroissements $\tau_k := T_k - T_{k-1}$, $k \geqslant 1$ sont indépendants et déterminer leur loi (on note $T_0 = 0$). Calculer $\mathbb{E}[\tau_k]$, $\mathrm{Var}(\tau_k)$. En déduire $\mathbb{E}[T_N]$ et $\mathrm{Var}(T_N)$.
2. En utilisant l'inégalité de Bienaymé-Tchebychev et le fait que $\sum_{j=1}^{N} 1/j$ est équivalent à $\ln N$, montrer que pour tout $\delta > 0$, on a :

$$\mathbb{P}\left(\left|\frac{T_N}{N\ln N} - 1\right| > \delta\right) \overset{N\to+\infty}{\longrightarrow} 0\,.$$

Solution.

1. On a $\tau_1 = T_1 = 1$. τ_2 est le nombre de tirages nécessaire pour tirer une des $N-1$ images qui n'ont pas été tirées au premier coup. C'est donc une v.a. géométrique de paramètre $(N-1)/N$. On raisonne de proche en proche : τ_k est le nombre de tirages nécessaire pour tirer une des $N-(k-1)$ images qui n'ont pas été tirées lors des T_{k-1} premiers tirages. τ_k est donc indépendante de T_{k-1} et suit une loi géométrique de paramètre $(N-k+1)/N$. On montre ainsi que les τ_k sont indépendantes et suivent des lois géométriques de paramètres $(N-k+1)/N$, $k = 1, \ldots, N$. D'après les formules donnant l'espérance et la variance d'une v.a. géométrique :

$$\mathbb{E}[\tau_k] = \frac{N}{N-k+1}\,, \qquad \mathrm{Var}(\tau_k) = \frac{(k-1)N}{(N-k+1)^2}\,.$$

On a $T_N = \sum_{k=1}^{N} \tau_k$. Par linéarité de l'espérance :

$$\mathbb{E}[T_N] = \sum_{k=1}^{N} \mathbb{E}[\tau_k] = N\sum_{k=1}^{N} \frac{1}{N-k+1} = N\sum_{j=1}^{N} \frac{1}{j}\,.$$

Par indépendance des τ_k :

$$\mathrm{Var}(T_N) = \sum_{k=1}^{N} \mathrm{Var}(\tau_k) = N \sum_{k=1}^{N} \frac{k-1}{(N-k+1)^2} = N \sum_{j=1}^{N} \frac{N-j}{j^2}\,.$$

2. D'une part, on a :

$$\mathbb{E}[T_N] = N \sum_{j=1}^{N} \frac{1}{j} \simeq N \ln N + o(N \ln N)\,, \qquad N \to +\infty.$$

Soit $\delta > 0$. Pour N assez grand,

$$\Big|\frac{\mathbb{E}[T_N]}{N \ln N} - 1\Big| \leqslant \frac{\delta}{2}\,.$$

Comme :

$$\mathbb{P}\left(\Big|\frac{T_N}{N \ln N} - 1\Big| \geqslant \delta\right) \leqslant \mathbb{P}\left(\Big|\frac{T_N}{N \ln N} - \frac{\mathbb{E}[T_N]}{N \ln N}\Big| \geqslant \delta - \Big|\frac{\mathbb{E}[T_N]}{N \ln N} - 1\Big|\right)\,,$$

on trouve que, pour N assez grand,

$$\mathbb{P}\left(\Big|\frac{T_N}{N \ln N} - 1\Big| \geqslant \delta\right) \leqslant \mathbb{P}\left(\Big|\frac{T_N}{N \ln N} - \frac{\mathbb{E}[T_N]}{N \ln N}\Big| \geqslant \frac{\delta}{2}\right)\,.$$

D'autre part :

$$\mathrm{Var}(T_N) = N \sum_{j=1}^{N} \frac{N-j}{j^2} \leqslant N^2 \sum_{j=1}^{N} \frac{1}{j^2} \leqslant K N^2\,,$$

où $K = \sum_{j=1}^{+\infty} j^{-2}$ est fini car c'est la somme d'une série convergente (en fait, $K = \pi^2/6$). En utilisant l'inégalité de Bienaymé-Tchebychev :

$$\mathbb{P}\left(|T_N - \mathbb{E}[T_N]| \geqslant \frac{\delta}{2} N \ln N\right) \leqslant \frac{4\mathrm{Var}(T_N)}{\delta^2 N^2 \ln^2 N} \leqslant \frac{4K}{\delta^2 \ln^2 N}\,,$$

qui tend vers 0 quand $N \to +\infty$.
On a donc montré que :

$$\mathbb{P}\left(\Big|\frac{T_N}{N \ln N} - 1| \geqslant \delta\right) \overset{N \to +\infty}{\longrightarrow} 0\,.$$

Pour aller plus loin

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

IV.4.1

1. On effectue 10 mesures d'une quantité qui suit une loi gaussienne de moyenne μ et de variance σ^2 inconnues. On obtient :
$$4{,}81\,,\ 5{,}72\,,\ 4{,}41\,,\ 7{,}18\,,\ 4{,}86\,,\ 5{,}11\,,\ 6{,}07\,,\ 5{,}06\,,\ 4{,}90\,,\ 4{,}17\,.$$
Déterminer un intervalle de confiance pour la moyenne et la variance à $90\,\%$.
2. On suppose maintenant qu'on a effectué 100 mesures. Les moyenne et variance empiriques sont $\bar{X}_{100} = 4{,}88$ et $V_{100} = 0{,}914$. Déterminer un intervalle de confiance pour la moyenne et la variance à $90\,\%$.

IV.4.2

Un livre de $1\,000$ pages contient $1\,000$ erreurs, réparties au hasard. On ouvre le livre et l'on compte le nombre d'erreurs X contenues dans une page.

1. Identifier la loi de X.
2. Quelle loi pourrait approcher la loi de X ?
3. Vérifier l'approximation précédente en calculant $\mathbb{P}(X = 0)$ et $\mathbb{P}(X = 1)$ et en comparant avec les mêmes probabilités pour la loi approchée.

IV.4.3

Le but de cet exercice est de démontrer de manière probabiliste le théorème de Weierstrass : « toute fonction continue sur $[0,1]$ est limite uniforme de polynômes ».
Soit $f : [0,1] \to \mathbb{R}$ continue. Soit S_n une v.a. binomiale de loi $\mathcal{B}(n,x)$, où $x \in [0,1]$. On note pour $\delta > 0$:
$$A_{n,\delta} := \left\{ \left| \frac{S_n}{n} - x \right| > \delta \right\}.$$

1. En utilisant l'inégalité de Markov, montrer que :
$$\mathbb{P}(A_{n,\delta}) \leqslant \delta^{-2}\mathbb{E}\left[\left(\frac{S_n}{n} - x\right)^2\right].$$
En déduire que :
$$\mathbb{P}(A_{n,\delta}) \leqslant \frac{1}{4n\delta^2}\,.$$
2. On note $Z = f(x) - f(S_n/n)$. En écrivant $Z = Z\mathbf{1}_{A_{n,\delta}} + Z\mathbf{1}_{A^c_{n,\delta}}$, majorer $|\mathbb{E}[Z]|$. En déduire que pour toute fonction continue :
$$\lim_{n\to+\infty} \sup_{x\in[0,1]} \left| f(x) - \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k} \right| = 0\,.$$
3. Soit $f : \mathbb{R}_+ \to \mathbb{R}$ continue bornée. Montrer que pour tout $x \in \mathbb{R}_+$:
$$\lim_{n\to+\infty} \left| f(x) - \sum_{k=0}^{+\infty} e^{-nx} \frac{(nx)^k}{k!} f\left(\frac{k}{n}\right) \right| = 0\,.$$
Si l'on suppose f uniformément continue, la convergence ci-dessus est-elle uniforme en x ? Pour montrer que ce n'est pas le cas, on considérera l'exemple où $f(x) := \cos(x)$ et l'on étudiera la convergence de la suite de fonctions aux points $x_n = 2\pi n$.

IV.4.4 On souhaite effectuer un sondage d'opinion sur les intentions de vote d'un second tour des élections présidentielles (où il ne reste plus que deux candidats). On note p_A la proportion de votes pour le candidat A dans la population, et n le nombre de personnes interrogées.

1. Quelle est la loi de $X :=$ le nombre de personnes interrogées déclarant vouloir voter pour A ?
2. Proposer une loi simplifiée qui approche la loi de X.
3. Combien de personnes faut-il interroger pour avoir une fourchette sur les intentions de vote de $\pm 1\ \%$ avec le niveau de confiance $95\ \%$?

IV.4.5 Des élections à un tour opposent deux candidats, disons B et G. La proportion d'électeurs qui votent pour B vaut p, paramètre inconnu, qui selon sa position par rapport à $50\ \%$ décide du sort de l'élection. Pour connaître le gagnant sans attendre le décompte de toutes les voix, on procède par sondage en interrogeant une petite partie des électeurs (n personnes). M_n d'entre elles ont voté pour B et $n - M_n$ pour G.

1. On veut déterminer un intervalle de confiance pour p au niveau α.
   En clair : on se donne $\alpha \in]0, 1[$, et l'on cherche des fonctions f et g telles que :
   $$\mathbb{P}_p\left(f(M_n) \leqslant p \leqslant g(M_n)\right) \geqslant 1 - \alpha\,.$$
   Si n est assez grand pour admettre l'approximation gaussienne, déterminer les fonctions f et g.
   Application : $\alpha = 5\ \%$.
2. À partir de quelle valeur m_0 (dépendant de n) de M_n peut-on déclarer B vainqueur avec probabilité de se tromper inférieure à $5\ \%$?
   Application numérique : $n = 1\,000$.
3. Supposons qu'un homme (respectivement une femme) vote B avec probabilité p_1 (respectivement p_2), et qu'il y autant d'hommes que de femmes. Par des calculs de variance, expliquer pourquoi il vaut mieux interroger $n/2$ hommes et $n/2$ femmes que n personnes choisies au hasard.

IV.4.6 Cet exercice est un problème qui fait intervenir différents résultats vus dans ce module.

Partie A : identité de Wald.

Soit X une v.a. entière de fonction génératrice G. On considère une suite $(X_i)_{i>0}$ de v.a. indépendantes et de même loi, la loi de X, supposée de carré intégrable.

1. Préliminaire : montrer que $G(0) = \mathbb{P}(X = 0)$ et que $G(1) = 1$.
2. Soient $S_0 := 0$ et, pour n entier strictement positif, $S_n := X_1 + X_2 + \cdots + X_n$.
   Montrer que la fonction génératrice de S_n est donnée par $G_{S_n}(s) = (G(s))^n$.
3. Soit, plus généralement, N une v.a. entière de carré intégrable indépendante des X_i et de fonction génératrice G_N. Soit S_N la v.a. définie par : $S_N := X_1 + X_2 + \cdots + X_N$ (avec la convention $S_N(\omega) = 0$ si $N(\omega) = 0$). Noter que le nombre de termes de la somme est aléatoire.
   (a) Montrer que la loi de S_N est donnée par :
   $$\mathbb{P}(S_N = k) = \sum_{n=0}^{+\infty} \mathbb{P}(S_n = k)\mathbb{P}(N = n)\,.$$
   (b) En déduire que la fonction génératrice de S_N est $G_{S_N} := G_N \circ G$ (identité de Wald).

4. Montrer, en utilisant l'identité de Wald, que :

$$\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X] \text{ et } \mathrm{Var}(S_N) = \mathbb{E}[N]\mathrm{Var}(X) + \mathbb{E}[X]^2\mathrm{Var}(N). \tag{18}$$

5. Exemple : on suppose que la loi des X_i est une loi de Bernoulli $\mathcal{B}(p)$ de paramètre p et que la loi de N est une loi de Poisson de paramètre λ. Montrer, en utilisant l'identité de Wald, que la loi de S_N est une loi de Poisson de paramètre λp et vérifier sur cet exemple les formules (18).

Partie B : intermède technique : une suite récurrente.
Soit g une application continue et croissante de $[0,1]$ vers $[0,1]$.
On considère la suite récurrente (a_n) définie par : $a_0 := 0$ et pour $n \geqslant 0$, $a_{n+1} := g(a_n)$.

1. Représenter graphiquement un tel exemple de suite (a_n) dans le carré $[0,1] \times [0,1]$.
2. Démontrer que la suite (a_n) est croissante. Montrer qu'elle est convergente. Montrer que sa limite a est un point fixe de g (i.e. $g(a) = a$).
3. Soit a_c le plus petit point fixe de G. Montrer que $a = a_c$.

Partie C : un processus généalogique. On étudie la transmission du nom « Tartempion » porté à l'origine par un seul homme. On modélise ce processus de la façon suivante :

- M. Tartempion forme la génération 0.
- M. Tartempion donne naissance à un nombre aléatoire $X_{0,1}$ de fils qui vont porter son nom. Ces hommes forment la génération 1.
- Chacun de ces $X_{0,1}$ hommes a des enfants. L'homme numéro i, pour $i \in \{1, \ldots, X_{0,1}\}$, donne naissance à un nombre aléatoire $X_{1,i}$ d'enfants mâles, qui forment la génération 2.
- D'une manière générale, chaque homme de la $n^{ème}$ génération donne naissance à un nombre aléatoire $X_{n,i}$ de fils et ce indépendamment des autres individus.
- On suppose en outre que les v.a. entières $X_{n,i}$ sont de carrés intégrables, indépendantes et de même loi, de fonction génératrice G (loi du nombre de fils par individu mâle).

On s'intéresse dans la suite à la probabilité d'extinction de ce processus. Soit Z_n le nombre d'individus mâles de la $n^{ème}$ génération et G_n la fonction génératrice de Z_n. On a donc en particulier $Z_0 = 1$ et $Z_1 = X_{0,1}$.

1. Exemple : représenter (par un arbre) les générations 0, 1 et 2 de l'arbre généalogique dans le cas où $X_{0,1}(\omega) = 2$, $X_{1,1}(\omega) = 3$, $X_{1,2}(\omega) = 4$. Préciser les valeurs de $Z_i(\omega)$ dans ce cas, pour $i = 0, 1, 2$.
2. Montrer que $Z_{n+1} = X_{n,1} + X_{n,2} + \cdots + X_{n,Z_n}$.
3. En déduire, en utilisant la partie A, que $G_{n+1}(s) = G_n\big(G(s)\big)$ et donc que $G_n(s) = G^n(s)$. (G^n désigne ici la composée $G \circ G \circ \cdots \circ G$).
4. Soit A_n l'événement « le processus est éteint à la $n^{ème}$ génération » et A_∞ l'événement « le processus s'éteint ».
   (a) Exprimer A_n, puis A_∞ comme événements liés aux v.a. Z_n. Montrer que la suite d'événements $(A_n)_{n \geqslant 0}$ est croissante.
   (b) Soit a_n la probabilité de A_n. Montrer que la suite (a_n) est convergente. On appelle a sa limite. Quelle est la probabilité de A_∞ ?

(c) Prouver que $a_n = G_n(0)$. En utilisant le résultat de la partie B, montrer que a est le plus petit point fixe de G dans $[0,1]$.

5. Que se passe-t-il si $\mathbb{P}(X_{0,1} = 0) = 1$? Que se passe-t-il si $\mathbb{P}(X_{0,1} = 0) = 0$?
On exclut ces cas dans la suite.

6. On note $m = \mathbb{E}[X_{0,1}]$ le nombre moyen de fils par individu.

   (a) Donner un point fixe évident de G.

   (b) Montrer que G est convexe. En vous aidant éventuellement d'un dessin, montrer que G a au plus deux points fixes sur $[0,1]$.

   (c) Que vaut $G'(1)$ (la dérivée à gauche en 1 de G) ?

   (d) Montrer que la probabilité d'extinction est 1 si $m \leqslant 1$, et strictement inférieure à 1 si $m > 1$.

7. Application : calculer le nombre moyen de fils de la $n^{ème}$ génération $\mathbb{E}[Z_n]$ et trouver la probabilité d'extinction a lorsque la loi du nombre de fils par individu est :

   (a) uniforme sur $\{0,1\}$,

   (b) uniforme sur $\{0,1,2\}$,

   (c) uniforme sur $\{0,1,2,3\}$.

Bibliographie

[Arnold] **V. Arnold**

[Ak] **N. I. Akhiezer.** *The Classical Moment Problem*, édition anglaise, Oliver and Boyd, 1965.

[Baker] **A. Baker.** *A Comprehensive Course in Number Theory*, Cambridge U.P., 2012.

[Bart] **R. G. Bartle.** *A Modern Theory of Integration*, Graduate Studies in Mathematics, Volume 32, American Mathematical Society, 2001.

[BCh] **Z.I. Borevitch, I.R. Chafarevitch.** Théorie des nombres, Gauthier-Villars, Paris 1967.

[Berger] **M. Berger.** *Géométrie*, Éditions Nathan, 1990

[Dei] **P. Deift.** *Orthogonal Polynomials and Random Matrices : a Riemann-Hilbert Approach*, American Math. Soc., 2000.

[De] **J.P. Demailly.** *Théorie élémentaire de l'intégration : l'intégrale de Kurzweil-Henstock.*
`http://www-fourier.ujf-grenoble.fr/~demailly/manuscripts/kurzweil.pdf`

[Demazure] **M. Demazure.** *Cours d'algèbre. Primalité, divisibilité, codes*, Cassini, 1997.

[DW2] **C. Deschamps, A. Warusfel.** *Mathématiques* 2ème *année*, Dunod, 2001.

[D1] **J. Dieudonné.** *Éléments d'analyse, Tome I, fondements de l'analyse moderne*, Gauthier-Villars, 1969.

[D2] **J. Dieudonné.** *Calcul infinitésimal*, Collection Méthodes, Hermann, 1968.

[D3] **J. Dieudonné.** *Sur les groupes classiques*, Actualités Scientifiques et industrielles, Hermann Paris, 1948.

[G2] **R. Godement.** *Analyse mathématique II*, Springer-Verlag, 1998.

[HW] **G.H. Hardy and E.M. Wright.** *An Introduction to the Theory of Numbers*, Oxford U.P.

[Hindry] **M. Hindry.** *Arithmétique*, Calvage et Mounet, 2008.

[Knuth2] **D. E. Knuth.** *The Art of Computer Programming, vol. 2*, Addison-Wesley.

[PCP] **J.-Ph. Pérez, R. Carles, O. Pujol.** *Quantique*, De Boeck, Bruxelles 2013.

[Perrin] **D. Perrin,** *Cours d'algèbre*, CAPES/AGREG Mathématiques, ellipses, 1996.

[R3] **E. Ramis, C. Deschamps, J. Odoux.** *Cours de Mathématique spéciales, tome* 3 *: topologie et éléments d'analyse*, Masson, 1976.

[Ru] **W. Rudin.** *Analyse réelle et complexe* Dunod, 2009.

[Serre] **J.-P. Serre.** *Cours d'arithmétique*, PUF.

[L1] **J.P. Ramis, A. Warusfel.** *Mathématiques, Tout-en-un pour la Licence, niveau L1*, Dunod Paris, 2005.

[L2] **J.P. Ramis, A. Warusfel.** *Mathématiques, Tout-en-un pour la Licence, niveau L2,*, Dunod Paris, 2013.

[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 : Arithmétique

I.1.1 On trouvera *quatre* couples (q, r) possibles.

I.1.2 Dans les deux cas, prouver que si u, v sont dans l'anneau, l'élément q de l'anneau tel que ses composantes approchent le mieux celles de u/v vérifie $N(u/v - q) < 1$.

I.1.3 Si $g(x)$ est minimum, diviser 1 par x.

I.1.4 (i) Utiliser la multilinéarité du discriminant.
(ii) Calculer les $F_k(D)$ pour la matrice D de forme réduite de la proposition.

I.1.5 Le système admet une solution entière si, et seulement si, $2 \mid b - 4a$, $2 \mid c - a$ et $174 \mid -3a - 2b + 12c$.

I.1.6 Écrire l'algorithme en style récursif.

I.1.7 Si $a \in A$, l'idéal $<a, X>$ est principal si, et seulement si, $a = 0$ ou a inversible.

I.1.8 Considérer l'exemple de 2 et X dans $\mathbb{Z}[X]$.

I.1.9 Démontrer que $\sqrt{|f| + |g|}$ divise f et g. Considérer l'idéal des fonctions nulles en $x \in I$.

I.1.10 Montrer que la réunion est un idéal (s'inspirer de la preuve du théorème de Krull dans le module « Compléments d'algèbre » du livre [L2]) et en considérer un générateur.

I.1.11 Poser
$I_n := \{a \in A \mid \exists P \in I \ : \ P = aX^n + \cdots\}$
et utiliser l'exercice I.1.10.

I.1.12 Prendre $A := \mathbb{Z}[X]/<8X>$, $x := X \pmod{<8X>}$ et $y := 3x$.

I.1.13 Factoriser un élément non nul d'un tel idéal.

I.1.14 Pour compter les $m = 1, \ldots, n$ tels que $v_p(m) = k$, commencer par compter ceux tels que $v_p(m) \geqslant k$.

I.1.15 Montrer que $v_p(x) \geqslant 0$ pour tout $p \in P$.

I.1.16 (i) Réduire modulo p l'égalité $P = QR$.
(ii) Appliquer le critère d'Eisenstein au polynôme $P(X + 1)$.

I.1.17 Écrire $F = c\tilde{F}$ et $G = d\tilde{G}$ (contenu et polynôme primitif).

I.1.18 Si $L = \mathbb{Q}$ ou $K(X)$ selon le cas, partir d'un irréductible de $L[T]$, chasser les dénominateurs et diviser les coefficients par leur pgcd.

I.1.19 (i) Appliquer l'exercice II.1.30 du module « Compléments d'algèbre » de [L1].
(ii) Si $\mathcal{P} \neq \{0\}$, soit $F \in \mathbb{Z}[X]$ un générateur primitif de $\mathcal{P}\mathbb{Q}[X]$; alors $\mathcal{P} = <F>$.

I.1.20 Si $f, g \in I$, appliquer la théorie du résultant pour éliminer une variable.

I.1.21 La liste des facteurs premiers de n et leurs exposants sont bornés par la connaissance de m.

I.1.22 (i) Pour que $G_1 \times G_2$ soit cyclique, il faut, et il suffit, que G_1 et G_2 le soient et que leurs ordres soient premiers entre eux.

I.1.23 (i) Si $m \wedge n = 1$, tout diviseur de mn est de la forme de pour un unique couple (d, e) avec $d \mid m$, $e \mid n$.

I.1.24 Prendre pour K le corps des quaternions (exercices à la fin du module « Compléments d'algèbre » du livre [L1]).

I.1.25 Le groupe $< a > \cap < b >$ est trivial.

I.1.26 Utiliser le lemme chinois.

I.1.27 (ii) Compter les éléments tels que $x^2 = 1$.

I.1.28 $\varphi(\varphi(p^r)) = (p^{r-1} - p^{r-2})\varphi(p-1)$.

I.1.29 (ii) C'est le ppcm des $\varphi(p_i^{r_i})$.

I.1.30 Montrer que $a^{p-1} = 1 + px$ avec $x \in \mathbb{Z} \setminus p\mathbb{Z}$.

I.1.31 Tester séparément les congruences modulo p et q.

I.1.32 (ii) r est la dimension du $\mathbb{F}_p$-espace vectoriel K.
(iii) Automorphisme : voir ce qui concerne la caractéristique dans le livre [L1].

I.1.33 Utiliser le fait que K^* est cyclique.

I.1.34 Un groupe cyclique admet un élément d'ordre 4 si, et seulement si, son cardinal est multiple de 4 ; et, dans $\mathbb{F}_p^*$, un tel élément est la même chose qu'une racine carrée de -1.

I.1.35 On a $\mathbb{Z}[j] \simeq \mathbb{Z}[X]/ < X^2 + X + 1 >$.

I.1.36 Pour l'implication non triviale, considérer le morphisme de groupes surjectif $(\mathbb{Z}/p^r\mathbb{Z})^* \to (\mathbb{Z}/p\mathbb{Z})^*$ et sa restriction aux sous-groupes des carrés.

I.1.37 Prendre $b := a^{(k+1)/2}$.

I.1.38 (ii) Utiliser la loi de réciprocité quadratique.
(iii) On trouvera une condition de congruence modulo 60.

I.1.39 Si $N(\pi) = p$ nombre premier décomposé, on trouvera $\mathbb{F}_p$; si $\pi = q$ nombre premier inerte, on trouvera un corps à q^2 éléments.

I.1.40 Utiliser les exercices I.1.2 de la page 68 et I.1.35 de la page 71.

I.1.41 (i) Notant $E(n) := \{u \in \mathbb{Z}[\mathrm{i}] \mid N(u) = n\}$, considérer l'application $(u, u') \mapsto uu'$ de $E(n) \times E(n')$ dans $E(nn')$.

I.1.42 La fonction $n \mapsto \sum \left(\dfrac{-1}{m}\right)$ est multiplicative (exercice I.1.23 de la page 70).

I.1.43 $44! \equiv 34 \pmod{89}$.

I.1.44 $9^2 + 32^2 + 1 = 14 \times 79$.

I.1.45 (i) Écrire $F_m = u_{m+1}/u_m$, où $u_m := 2^{2^m} - 1$.

I.1.46 Les nombres premiers $p \leqslant n$ divisent $n!$.

I.1.47 Partir de :

$$(Dn/\ln(Dn))\ln(Dn/\ln(Dn)) > k \ln k.$$

I.1.48 De façon générale, $u_n \sim v_n/\ln v_n$ équivaut à $v_n \sim u_n \ln u_n$ (si les deux suites tendent vers $+\infty$).

I.1.49 Les deux expressions valent $\pi(n)$.

I.1.50 La taille de M_p est essentiellement proportionnelle à p, celle de F_n à 2^n.

I.1.51 On peut obtenir le nombre de facteurs premiers $p \leqslant \sqrt{n}$ de chaque $m \leqslant n$, voire les valeurs de l'indicatrice d'Euler.

I.1.52 Il faut invoquer la formule de Bayes.

I.1.53 Il faut utiliser la loi de réciprocité quadratique et le fait que $\left(\dfrac{m}{n}\right) = \left(\dfrac{m'}{n}\right)$ si $m \equiv m' \pmod{n}$.

I.1.54 Cette condition porte sur les coefficients binomiaux $\binom{n}{k}$.

I.1.55 On trouvera des racines primitives (pour servir de témoins) sur la page : `http://fr.wikipedia.org/wiki/Racine_primitive_modulo_n`

I.1.56 Factorisations : $561 = 3 \times 11 \times 17$; $1105 = 5 \times 13 \times 17$; $1729 = 7 \times 13 \times 19$; $41041 = 7 \times 11 \times 13 \times 41$.

I.1.57 Calculer la classe de F_n modulo 5 et invoquer la loi de réciprocité quadratique.

I.1.58 Les $\mathrm{i}^{-k} u_k$ et $\mathrm{i}^{-k} v_k$ satisfont une relation de récurrence bien connue.

I.1.59 On a une suite définie par $r_0 := a$ et $r_{n+1} := a + 1/r_n$, et une suite définie par $r_0 := a$ et $r_{n+1} := a/(1 + r_n)$. Dans le cas général, on trouve des points fixes $\ell \neq \ell'$ et l'on étudie la suite des $(r_n - \ell')/(r_n - \ell)$; si $\ell = \ell'$, on étudie $1/(r_n - \ell)$.

I.1.60 Le cas le pire est celui où $q_i = \varepsilon$ pour tout i.

I.1.61 Utiliser la formule donnant $r_i - r_{i-1}$.

I.1.62 $u_i v_{i-1} - v_i u_{i-1}$ est un déterminant.

I.1.63 $R_i = X_0 + \dfrac{1|}{|X_1} + \cdots + \dfrac{1|}{|X_i}$
et $R_i - R_{i-1} = \dfrac{(-1)^{i-1}}{V_i V_{i-1}}$.

I.1.64 Majorer $|\alpha - r_n| + |\alpha - r_{n+1}|$
par $\dfrac{1}{v_n v_{n+1}} \leqslant \dfrac{1}{2v_n^2} + \dfrac{1}{2v_{n+1}^2}$.

I.1.65 Faire varier x de 0 à $+\infty$ dans l'expression $\begin{bmatrix} u_i & u_{i-1} \\ v_i & v_{i-1} \end{bmatrix}(x)$.

I.1.66 Les u_i et les v_i satisfont une récurrence linéaire.

I.1.67 (i)

```
let rec fracrat n x =
if n = 0 then [] else
let a = (floor x) in
let y = (1./. (x -. a))
in ((int_of_float a) ::
(fracrat (n - 1) y));;
```

I.1.68 Commandes :

```
f:= exp(x):
convert(f,confrac,x);
```

(l'ordre du développement est fixé par la variable `Order`).
Puis `plot({f,g,f-g},x=0..10);`.

I.2 : Compléments sur les formes bilinéaires alternées et sur les formes quadratiques

I.2.4 Faire d'abord l'exercice I.2.4. 3 a)
Utiliser plusieurs fois l'égalité :
$\Phi(x+y) = \Phi(x) + \Phi(y) + B_\Phi(x,y)$.

I.2.8 3. $\kappa = 1/\sqrt{3}$.

I.2.9 3. Le noyau de σ est $\{0,1\}$. 4. a) Les formes Φ et Ψ sont équivalentes si, et seulement si, a est dans l'image de l'application σ. 4 b) Les formes Φ' et Ψ' sont équivalentes.

7. Utiliser le changement de base $e_1' := e_1/\sqrt{b}$ et $e_1' := \alpha e_1/\sqrt{b} + e_2\sqrt{b}$.

I.2.11 Montrer que la conique a un point dans $P^2(K)$. On a $|(C)| = q+1$.

I.2.13 3. Distinguer deux cas, selon la parité de $(q-1)/2$.

I.2.16 On peut s'inspirer du cas du plan *euclidien* traité dans le livre [L2].

I.2.23 2. On a $\varepsilon_\infty(\Phi) = (-1)^{\frac{q(q-1)}{2}}$.

II.1 : Surfaces

II.1.1 $M := (0,0,0)$ est le seul point multiple.

II.1.2 Le support de la nappe est la sphère de centre $(0,0,0)$ et de rayon 1, privée du point $(0,0,1)$. Tous les points sont simples et réguliers.

II.1.3 Les deux nappes ont même support : un paraboloïde de révolution. Elles ne représentent pas la même nappe géométrique puisque $\mathbf{f}_1$ est une nappe simple tandis que $\mathbf{f}_2$ ne l'est pas.

II.1.4 Non : c'est un « ruban de Moebius ». On pourra utiliser le difféomorphisme $\varphi : (t,u) \mapsto (-t, u+\pi)$.

II.1.5

1. Les points singuliers de $\mathbf{f}_1$ sont les points $\big((r,t), M\big)$ avec $r = 0$ et $M = (0,0,0)$.
2. Le seul point singulier de $\mathbf{f}_2$ est le point $\big((u,v), O\big)$ avec $u = v = 0$ et $O := (0,0,0)$.
3. Les points singuliers de $\mathbf{f}_3$ sont les points $\big((u,v), M\big)$ avec $0 \mod \pi/2$ ou $v = 0 \mod \pi/2$ et $M := \mathbf{f}_3(u,v)$. Le point M appartient alors à l'une des trois courbes paramétrées $u \in \mathbb{R} \mapsto (\cos^3 u, \sin^3 u, 0)$, $v \in \mathbb{R} \mapsto (0, \cos^3 v, \sin^3 v)$ ou $v \in \mathbb{R} \mapsto (\cos^3 v, 0, \sin^3 v)$.
4. Tous les points sont réguliers.

II.1.6

1. Le plan tangent au point :
$$(x_0, y_0, z_0) := \mathbf{f}_1(u_0, v_0)$$
est le plan d'équation :
$$a(x-x_0)+b(y-y_0)+c(z-z_0) = 0$$
avec $a := 6u_0v_0^2$, $b := -3v_0^2$ et $c := (1-2u_0)$.
2. Le plan tangent au point :
$$(x_0, y_0, z_0) := \mathbf{f}_2(x_0, y_0)$$
est le plan d'équation :
$$(z-z_0) = a(x-x_0) + b(y-y_0)$$
avec $a := 4x_0(x_0^2+y_0^2-4)$ et $b := 4y_0(x_0^2+y_0^2-4)$.

II.1.7

1. Les points de la droite $\{x = 0 \text{ et } z = 1\}$ et ceux de la droite $\{y = 0 \text{ et } z = -1\}$.
2. Les points $(-1,1,1)$, $(1,-1,1)$ et $(1,1,-1)$.
3. Les points $(1,0,0)$ et $(-1,0,0)$.
4. Il n'y a pas de point singulier.

II.1.8

1. $\overrightarrow{\operatorname{grad}} F(M_0) = \frac{4\sqrt{3}}{3}(1,1,1)$.
2. L'équation du plan tangent est $x+y+z = \sqrt{3}-1$.
3. La surface traverse le plan tangent au voisinage de M_0 et l'intersection avec le plan tangent est composé de deux arcs qui se coupent en M_0 avec des tangentes distinctes.

II.1.11 C'est le tore de révolution de l'exemple de la page 153.

II.1.12 Observer que (x,y,z) appartient à la surface si, et seulement si, (tx, ty, z) appartient à la surface pour tout $t \in \mathbb{R}$.

II.1.13 La quadrique est un hyperboloïde à deux nappes. Le centre est $(-1/8, 1/4, 3/4)$ et la nouvelle constante $7/8$. Il y a trois valeurs propres : 1 (évidente), et les solutions de $u^2+u-8 = 0$.

II.1.14 1. $\frac{X^2}{2} + Y^2 + \frac{3Z^2}{4} = 1$; ellipsoïde.

2. $12X^2 + 18Y^2 - 36Z^2 = 1$; hyperboloïde à une nappe.

II.1.15 1. C'est la moitié du paraboloïde hyperbolique d'équation $z = 4x^2 - y^2$.

2. Il s'agit de l'ellipsoïde d'équation $\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} + \dfrac{x_3^2}{\alpha_3^2} = 1$.

3. C'est l'hyperboloïde à une nappe d'équation $\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} - \dfrac{x_3^2}{\alpha_3^2} = 1$.

4. C'est la moitié de l'hyperboloïde à deux nappes d'équation $\dfrac{x_1^2}{\alpha_1^2} + \dfrac{x_2^2}{\alpha_2^2} - \dfrac{x_3^2}{\alpha_3^2} = -1$.

III.1 : Intégration

III.1.1 3. On utilise une subdivision produit de $\mathbf{P}$ induisant des subdivisions adaptées sur X et Y.

III.1.3 Soient $\mathbf{P} \subset \mathbb{R}^n$ le pavé fermé de centre a et de rayon $r > 0$ et $\mathbf{P}' \subset \mathbb{R}^n$ le pavé fermé de centre a et de rayon $r2^{1/n}$. On vérifie qu'il existe un pavé fermé $\mathbf{P}''$ dont les sommets sont des nombres rationnels et tel que $\mathbf{P} \subset \mathbf{P}'' \subset \mathbf{P}'$.

III.1.4 2. On se ramène au cas où $A \subset \mathbf{P} \subset U$, où $\mathbf{P}$ est un pavé fermé.

III.1.5 1. À x fixé il n'existe qu'un nombre fini de points y tels que $f(x,y) \neq 0$. 2. Si φ et ψ sont des fonctions en escalier sur $[0,1]^2$ telles que $\psi \leqslant f \leqslant \varphi$, on a $\displaystyle\int_{[0,1]^2} \varphi = 1$ et $\displaystyle\int_{[0,1]^2} \psi = 0$.

III.1.7 1. On utilise l'exercice 13 de la page 665 du module *Intégration* du livre [L2]. On définit une fonction $g : [0,1] \to \mathbb{R}$ par $g(x) := 1/q_x$ si $x \in \mathbb{Q}$ et $g(x) = 0$ sinon. On utilise ensuite $f(x,y) \leqslant g(x) + g(y)$.
3. On définit, pour $x \in \mathbb{Q}$ fixé, une fonction $h : [0,1] \to \mathbb{R}$ par, $h(y) := 1/q_x$ si $q \in \mathbb{Q}$ et $h(y) := 0$ sinon. On note $f_x(y) := f(x,y)$. On a $f_x = h + g$, g est réglée et h n'est pas $\mathscr{R}$-intégrable.

III.1.8 1. On montre que l'ensemble

$$W := \{\lambda + \mu\sqrt{2}, \lambda - \mu\sqrt{2} \mid \lambda, \mu \in \mathbb{Q}\}$$

est un sous $\mathbb{Q}$-espace vectoriel dense de $\mathbb{R}^2$. On en déduit la densité de A dans $[0,1]^2$. On raisonne ensuite comme dans l'exercice III.1.5, question 2.

III.1.9 2. On remarque :

$$f(x,y) = 1 + 2(y - 1/2)1_{\mathbb{R}\setminus\mathbb{Q}}(x).$$

III.1.10 1. Utiliser un raisonnement par récurrence. 3. Appliquer 2. à la suite $|f - f_n|$.

III.1.11 2. Utiliser le critère de Darboux. 3. Utiliser le théorème d'Arzela (*cf.* le module *Intégration* du livre [L2]).

III.1.12 L'intégrale vaut $\ln(4/3)$.

III.1.13 L'aire vaut πab.

III.1.14 1. L'intégrale vaut $1/72$.

2. L'intégrale vaut $2 - 5/e$.

III.1.16 1. Utiliser $\cos(2n+3)$. 3. Si $p \in \mathbb{N}^*$ est impair, $p^2 - 1 = 2n$. On a $\cos(n+1)^2 = \cos n^2 \cos p^2 - \sin n^2 \sin p^2$. 4. On a $T_9(-1/2) = 1 \neq -1/2$.

III.1.17 On a :

$$\int_{K_n} f = \frac{\pi}{2} - \left(\arctan\frac{1}{n} + \arctan n\right) = 0$$

$$\int_{K'_n} f = \frac{\pi}{4} - \arctan\frac{1}{n^2} - \arctan n + \arctan\frac{1}{n}\cdot$$

On a $\lim\limits_{n\to+\infty} \int_{K''_n} f = \pi/4$.

III.1.19 2. On montre $I = \sum\limits_{k\geqslant 0} \frac{1}{(2n+1)^2}$.

3. Le jacobien de h est égal à :

$$1 - \frac{\sin^2\varphi\sin^2\psi}{\cos^2\varphi\cos^2\psi} = 1 - x^2y^2.$$

4. $\operatorname{vol} T = \pi^2/8$ donc $\frac{3}{4}\zeta(2) = \frac{\pi^2}{8}$ et $\zeta(2) = \frac{\pi^2}{6}$.

III.1.20 1. On a $\sum\limits_{k\in\mathbb{N}^*} \frac{1}{(k+1/2)^2} = 3\,\zeta(2)$.

2. Utiliser $\frac{1}{(1-xy)\sqrt{xy}} = \sum\limits_{k\in\mathbb{N}} (xy)^{k-1/2}$.

3. $J(g)(\xi,\eta) = 4\,\frac{\xi\eta(1-\xi^2\eta^2)}{(1+\xi^2)(1+\eta^2)}$.

III.1.21 1. Intégrer en x, puis en y et utiliser $y = u^2$. On obtient $I = \pi^2/2$. Intégrer en x, puis en y et utiliser $y = u^2$. On obtient $I = \displaystyle\int_0^{+\infty} \frac{\ln x}{x^2-1}\,\mathrm{d}x$.

2. On montre :

$$\int_0^{+\infty} \frac{\ln x}{x^2-1}\,\mathrm{d}x = 2\int_0^{+\infty} \frac{\ln x}{x^2-1}\,\mathrm{d}x.$$

III.1.22 1. $A = \pi^2/32$.

3. $B+C = D$, $B-C = -2D$, $D = -\frac{3}{4}\zeta(2)$, $B = -D/2 = \frac{3}{8}\zeta(2)$.

4. $A = \frac{1}{2}B$, donc $\pi^2/32 = \frac{3}{16}\zeta(2)$ et $\zeta(2) = \pi^2/6$.

III.1.24 Si $J \subset [0,1]$ est un intervalle ouvert non vide, f n'est pas bornée sur J.

III.1.25 2. Choisir une jauge δ « forçant », pour toute subdivision δ-fine, 1 et tous les c_k n'appartenant pas à l'intérieur du dernier segment de subdivision à être des points de marquage.

3. On a $\displaystyle\int_{c_{k-1}}^{c_k} \big|f(x)\big|\,\mathrm{d}x = |a_k|$, donc $\displaystyle\lim_{k\to+\infty} \int_{c_{k-1}}^{c_k} \big|f(x)\big|\,\mathrm{d}x = +\infty$.

Utiliser la série harmonique alternée.

4. Montrer que $f-g$ est $\mathscr{HK}$-intégrable.

III.1.26 La fonction f est intégrable au sens de Riemann généralisé sur $[0,+\infty[$.

On a $\displaystyle\int_0^{+\infty} f = S$.

III.1.27 On utilise le « Squeeze Theorem », proposition 80 de la page 266.

III.1.28 1. On a $\displaystyle\int_{\mathbb{R}} f_n = a_n$.

2. Utiliser $h := \sup\limits_{n\in\mathbb{N}} f_n$.

3. Utiliser des suites extraites de $(a_n)_{n\in\mathbb{N}}$ telles que la série correspondante converge.

III.1.29 2. Utiliser la suite $(h_n := f_n 1_{\mathbb{R}_+})_{n\in\mathbb{N}^*}$.

III.1.31 Considérer $\mathbb{R}\setminus\mathbb{Q}$.

III.1.32 1. La suite $\big(\|f_k\|_1\big)_{k\in\mathbb{N}}$ tend vers 0.

2. Pour tout $n\in\mathbb{N}$ il existe $k', k'' > n$ tels que $f_{k'}(x) = 1$ et $f_{k''}(x) = 0$.

III.1.33 Soit $c\in\mathbb{R}$, on a $c = \inf\{a\in\mathbb{Q} \mid c < a\}$.

III.1.34 2. Soit $\delta := \inf_{x \in F} d(x, \mathbb{R} \setminus U)$. On a $\delta > 0$. La fonction $\varphi : x \mapsto \left(1 - \delta^{-1} d(x, F)\right)^+$ répond à la question.

III.1.38 1. On utilise des bijections $\mathbb{N} \to \mathbb{N} \times \mathbb{N}$ et $\mathbb{R} \to \mathbb{N} \times \mathbb{R}$. Le cardinal de $\mathbb{R}^{\mathbb{N}}$ est celui de $\mathbb{R}$. Le cardinal de $\mathbb{R}^{\mathbb{R}}$ est celui de $2^{\mathbb{R}}$.

5. Considérer l'ensemble des parties de l'ensemble de Cantor.

III.1.41 Utiliser une *bijection* $\varphi : \mathbb{N} \to \mathbb{N}^2$ et les résultats de l'exercice III.1.18.

III.1.43 1.a) On utilise la formule de Fubini pour l'intégrale de Riemann.

1.b) On utilise la formule de Chasles pour l'intégrale de Riemann.

1.c) La notion de suite exhaustive de compacts est invariante par homéomorphisme.

III.1.44 On a $\operatorname{vol}(U \times V) = \operatorname{vol} U \times \operatorname{vol} V$ et $\mu(K \times L) = \mu(K) \times \mu(L)$.

III.1.45 On utilise l'exercice précédent.

On a $\mu(K^n) = (\log K)^n > 0$, donc K^n n'est pas un ensemble négligeable, mais K^n est d'intérieur vide.

III.2 : Introduction aux équations aux dérivées partielles

III.2.1 On pourra utiliser la formule de d'Alembert (84) de la page 409.

III.2.2 On pourra appliquer la proposition 5 de la page 390.

III.2.3 On pourra reprendre et adapter les calculs de l'exercice précédent.

III.2.8 Il faudra utiliser les résultats classiques (du livre [L1]) sur la résolution des équations différentielles ordinaires d'ordre 2 à coefficients constants.

III.3 : Constructions de R et introduction aux nombres p-adiques

III.3.4 (i) b) C'est la question difficile de l'exercice. Choisir k minimal en un sens convenable et distinguer les cas $a_k = (v_k)_k$ et $a_k \neq (v_k)_k$. Dans le premier cas on montre $a = (v_k)_k$ et dans le second on montre qu'il existe $p > k$ tel que $a = (v_p)_p$.

(iv) Utiliser les suites $(s_n := 1 + 1 + \frac{1}{2} + \cdots + \frac{1}{n!})$, $\left(t_n := \frac{1}{n \times n!}\right)$ et $(u_n) := (s_n) + (t_n)$.

III.3.6 On trouve une écriture périodique à partir du troisième chiffre : $24/17 = \ldots \underline{011202122110201010}$.

III.3.7 $x^{-1} = \ldots 5650335$.

III.3.10 On a : $5^6 - 1 = 15624 = 9 \times 1736$, $1/9 = \ldots 023421023421\underline{023421}$.

III.3.12 Les suites (n), $(1/n)$ et (q^n) divergent.

La suite (p^n) tend vers 0.

III.3.14 C'est une généralisation de la suite récurrente de l'exercice 13 de la page 459.

III.3.20 Les rayons critiques r figurent parmi l'ensemble des solutions des équations $r^{j-i} = |a_i|_p / |a_j|_p$, avec $0 \leqslant i < j \leqslant m$.

III.3.24 2 Utiliser la proposition et 36 de la page 466, et le corollaire 35 de la page 466.

III.3.25 3, 4 Le point délicat est la composition.
On pourra utiliser l'exercice 16.

III.3.27 3. Utiliser $\mathbb{U}_1 = \varprojlim \mathbb{U}_1/\mathbb{U}_n$. 4. Remplacer $a = 1 + p$ de la question 2 par $a \in \mathbb{U}_2 \setminus \mathbb{U}_3$.

III.3.35 L'élément neutre de l'addition de $\mathcal{S}$ n'est pas celui de $\mathbb{R}$.

III.3.39 2. Utiliser $|x+1|^m$ $(m \in \mathbb{N}^*)$ et montrer
$|x+1| \leqslant \sqrt[m]{m+1}\max(1, |x|)$.

III.3.44 Pour montrer l'inégalité ultramétrique, majorer et minorer f sur la sphère unité de L pour une norme sup et utiliser la question 1 de l'exercice III.3.39.

III.3.45 1. Procéder par récurrence : $H_{n-1} \cup H_2 \Rightarrow H_n$. 2. Chercher x sous la forme $x = u_1x_1 + \cdots + u_nx_n$, avec $u_i = z_i^n/(1+z_i^n)$.

III.4 : Polynômes orthogonaux

III.4.1 b) On a

$$\nu_n = a^{-n}\sum_{k=0}^{n}\binom{n}{k}(-b)^{n-k}\mu_k$$

III.4.2 On pose $p_n := \gamma_n u_n$. On montre $(tu_{n-1} \mid u_{n-1}) = (tu_{n-1} - u_n \mid u_{n-1})$ et

$$(tu_{n-1} - u_n \mid u_{n-1}) = \left(\frac{\gamma_n'}{\gamma_n} - \frac{\gamma_{n-1}'}{\gamma_{n-1}}\right)(t^{n-1} \mid u_{n-1}).$$

III.4.5 a) On a $J_0 = \frac{\pi}{2}$, $J_1 = 0$, et pour $n \geqslant 2$, $J_n = \frac{n-1}{n}J_{n-2}$. Donc $J_{2p} = \frac{(2p)!}{4^p(p!)^2}\frac{\pi}{2}$ et $J_{2p+1} = 0$.
b) On a $\mu_n = 2(J_n - J_{n+1})$. Donc $\mu_{2p} = \frac{(2p)!}{4^p(p!)^2}\pi$ et $\mu_{2p+1} = -\frac{(2p+2)!}{4^p((p+1)!)^2}\pi$.
d) On utilise $\frac{\cos\left((2n+1)\frac{\arccos t}{2}\right)}{\sin\frac{\arccos t}{2}}$.

III.4.6 Indication. Supposons qu'il existe $g \in \mathcal{C}([\alpha, \beta])$, telle que $\mu_g = \delta_c$. Soit $d \in]\alpha, \beta[$, $d \neq c$, alors $g(d) = 0$.

III.4.7 Si $P \in \mathcal{U}_n$, $P = \frac{T_n}{2^{n-1}} - Q$, où T_n est le polynôme de Tchebychev de degré n et $Q \in \mathbb{R}_{n-1}[T]$. Donc :

$$\alpha_n = d^2\left(\frac{T_n}{2^{n-1}}, \mathbb{R}_{n-1}[T]\right) = \left(\frac{1}{2^{n-1}}\right)^2 \|T_n\|^2 = \frac{\pi}{2^{2n-1}}$$

(car $T_n \perp \mathbb{R}_{n-1}[T]$).

III.4.8 La forme demandée est :

$$(t^2e^{-2/t}y')' - \lambda e^{-2/t}y = 0.$$

Le poids est $w(t) := e^{-2/t}$.

III.4.9 c) On utilise les relations d'orthogonalité sur U en remarquant que les polynômes de Bessel restent solution de l'équation différentielle sur le plan complexe. f) On constate que le résultat de l'exercice III.4.2 reste valable, *mutatis mutandis*, pour les polynômes orthogonaux sur U. On obtient $a_n = 2n+1$ et $b_n = 1$. h) On obtient $\mu_n = (-2)^{n+1}/(n+1)!$, puis

$$D_n = (-1)^{\frac{n(n-1)}{2}}(-2)^{n^2}\prod_{s=1}^{n-1} s!/\prod_{s=n}^{2n-1} s!.$$

III.4.11 a) On montre que :

$$\big((1-t^2)P_n' + ntP_n | P_k\big) = 0$$

pour tout $k \in [\![0, n-2]\!]$.
b) On utilise l'équation différentielle de Legendre pour montrer la première relation, puis la relation de récurrence pour obtenir la deuxième.

III.4.15 a) Les opérateurs div et $\overrightarrow{\text{grad}}$ sont invariants par isométrie.
b) On note (ρ, θ) les coordonnées polaires.
On a :
$\rho\,\mathrm{d}\rho = x\,\mathrm{d}x + y\,\mathrm{d}y,\ \rho^2\,\mathrm{d}\theta = x\,\mathrm{d}y - y\,\mathrm{d}x$.
On obtient $\Delta f = \frac{\partial^2 f^*}{\partial \rho^2} + \frac{1}{\rho}\frac{\partial f^*}{\partial \rho} + \frac{1}{\rho^2}\frac{\partial^2 f^*}{\partial \theta^2}$.

III.4.16 a) On a : $\dim \mathcal{F}_k = k+1$.
d) Pour tout $k \in \mathbb{N}$, la dimension (sur $\mathbb{R}$ ou $\mathbb{C}$) est 2.

III.4.17 1. a) On
a $\dim \mathcal{F}_k = (k+1)(k+2)/2$.
c) $\Delta r^{2m} g = r^{2m}\Delta g$

$$+ 2m(2k-2m+1)r^{2m-2}g.$$

d) On raisonne par récurrence (finie) sur m.
e) On a : $\dim \mathcal{H}_k = 2k+1$.
2. d) On note :

$$Y_{k,m}(\theta, \varphi) = e^{im\theta} P_{k,m}(\cos\theta).$$

Les parties réelles et imaginaires des $r^m Y_{k,m}$:
$H_{n,m} = r^m \cos m\theta P_{k,m}(\cos\theta)$
et $H_{n,-m} = r^m \sin m\theta P_{k,m}(\cos\theta)$
$(m = 0, 1, \ldots, k)$ donnent, en revenant en coordonnées (x, y, z), les $2k+1$ éléments d'une base de $\mathcal{H}_k$.

III.4.19 b) On a, pour $|t| < 1$, $\frac{|L_n'(t)|}{n!} \leqslant 2^n$.
c) On a $L_n'(t) - nL_{n-1}'(t) = -nL_{n-1}(t)$.
On en déduit, en utilisant la relation de récurrence, $tL_n'(t) = nL_n(t) - n^2 L_{n-1}(t)$, et enfin
$tL_n''(t) + (1-t)L_n'(t) + nL_n(t) = 0$, à partir de la dérivée de cette dernière relation.

III.4.22 3. Utiliser l'équation différentielle des polynômes de Jacobi.
4. Utiliser 3. avec $\lambda := 0$ et $\lambda = 1$.

III.4.25 a) Utiliser l'assertion (i) de la proposition 50 de la page 594.
c) On retrouve que $(t, x) \mapsto \mathrm{e}^{2tx - x^2}$ est la fonction génératrice des polynômes d'Hermite.

III.4.26 a) Remarquer que, pour $k \in [\![0, n-1]\!]$,

$$\mu_k = \int_\alpha^\beta \lambda_{k+1}^k w(\lambda_{k+1})\,\mathrm{d}\lambda_{k+1},\ \ldots,$$

$$\mu_{k+n} = \int_\alpha^\beta \lambda_{k+1}^{k+n} w(\lambda_{k+1})\ \mathrm{d}\lambda_{k+1}.$$

c) On remarquera que la permutation $\sigma \in \mathfrak{S}_n$ agissant sur les n premières lignes d'un déterminant D le transforme en $\varepsilon(\sigma)D$.
d) Égaler les coefficients de t^n dans la représentation intégrale du a) et utiliser le corollaire 66 de la page 620.

III.4.28 c) Écrire :

$$\int_\alpha^\beta \arctan\frac{u-d}{y} w(u)\,\mathrm{d}u$$
$$= \int_\alpha^d \arctan\frac{u-d}{y} w(u)\,\mathrm{d}u$$
$$+ \int_d^\beta \arctan\frac{u-d}{y} w(u)\,\mathrm{d}u$$

et calculer séparément les limites des deux intégrales.

III.4.30 Utiliser une famille normale convenable.

III.4.31 c) Étudier $-n/x$ en fonction de n. On obtient $f(-10) = -0{,}09156$ à 3.10^{-5} près.

III.4.32 b) Pour montrer la deuxième relation, on utilisera une intégration par parties à partir de l'intégrale définissant J_{n-1}.

III.4.35 On trouve :

$$h_n(-10)/L_n(-10) = -191/2\,086$$
$$\approx -0{,}09156279 \text{ (à } 6.10^{-4} \text{ près)},$$

en accord avec $f(-10) \approx -0{,}09156$ trouvé dans l'exercice cité.

IV.1 : Notions fondamentales sur les probabilités

IV.1.1 On peut vérifier que les deux événements sont complémentaires.

IV.1.2 On pourra par exemple d'abord traiter le cas où les A_i sont deux à deux disjoints, puis s'y ramener en considérant, pour tout $I \subset [\![1, n]\!]$, la partie B_I formé des intersections des A_i pour $i \in I$ d'avec les A_j^c pour $j \notin I$.

IV.1.3 Il faut se rappeler que $\mathbb{P}(A) = 1$ n'implique pas que $A = \Omega$, et que $\mathbb{P}(A) = 0$ n'implique pas que $A = \varnothing$. Par exemple,
si $(\Omega, \mathcal{A}, \mathbb{P}) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$ et si A désigne l'ensemble des nombres rationnels de $[0, 1]$, alors $\mathbb{P}(A) = 0$!
Il faut utiliser les propriétés fondamentales des probabilités (proposition 1 de la page 669).

IV.1.4 Une réalisation est un couple $\omega = (\omega_1, \omega_2)$ où ω_1 est la couleur de la face exposée et ω_2 est la couleur de la face cachée. L'espace fondamental est donc $\Omega = \{(RR), (RB), (BR), (BB)\}$.
Attention : la probabilité sur cet espace n'est pas la probabilité uniforme !
Ensuite, il faut utiliser la définition d'une probabilité conditionnelle.

IV.1.5 Pour le 1., utiliser la formule des probabilités totales. Pour le 2., utiliser la formule de Bayes.
Réponses : 1. $2{,}8\%$ 2. $\sim 57{,}1\%$.

IV.1.6 En notant A_i l'événement « le parapluie est à l'étage i », on pourra utiliser le fait que les A_i sont des événements incompatibles.

IV.1.7 Utiliser la formule du conditionnement multiple.

IV.1.8 Utiliser la formule des probabilités totales.

IV.1.9 En utilisant la formule des probabilités totales, écrire une relation de récurrence entre p_n et p_{n-1}.
Avec la valeur de p_1, résoudre la relation de récurrence en exprimant p_n en fonction de n.

IV.1.10 On notera p_n la probabilité que le buveur ne boive pas le $n^{ème}$ jour. Pour le 1., utiliser la formule des probabilités totales pour écrire une relation de récurrence entre p_n et p_{n-1}, puis résoudre la récurrence.

IV.1.11 Pour le 1., appliquer la formule des probabilités totales.
Pour le 2., appliquer la formule de Bayes.
Pour le 3., appliquer aussi la formule des probabilités totales.

IV.1.12 Pour le calcul de $\mathbb{P}(B_2)$, $\mathbb{P}(A_3)$, on peut appliquer la formule du conditionnement multiple. On peut aussi appliquer la formule du conditionnement multiple pour établir une relation de récurrence sur $\mathbb{P}(A_{2n-1})$ d'une part et sur $\mathbb{P}(B_{2n})$ d'autre part.
Pour le 2., décomposer A et B comme une union dénombrable d'événements incompatibles de la forme A_{2n-1} et B_{2n}, puis utiliser la σ-additivité.

IV.1.13 Pour le 1., on peut commencer par calculer la probabilité qu'un étudiant obtienne k bonnes réponses sachant qu'il a préparé l'examen, ainsi que la probabilité qu'un étudiant obtienne k bonnes réponses sachant qu'il n'a pas préparé l'examen. Avec la formule des probabilités totales, on trouve alors la probabilité qu'un étudiant obtienne k bonnes réponses, puis la probabilité qu'un étudiant réussisse l'examen.
Pour le 2., utiliser les résultats du 1. et la formule de Bayes.

IV.2 : Variables aléatoires discrètes

IV.2.1 On peut noter V_k l'événement « la porte n'est toujours pas ouverte après k essais ». En remarquant qu'il s'agit d'une suite décroissante d'événements, exprimer la probabilité p_k qu'il faille exactement k essais pour ouvrir la porte en fonction de la probabilité de V_k. Puis calculer la probabilité de V_k dans chacun des deux cas. Appliquer ensuite la définition de l'espérance d'une v.a. discrète.

IV.2.2 Développer l'expression de $\mathbb{E}[(X-c)^2]$, qui, en tant que fonction de c, est un polynôme du second degré avec un coefficient dominant strictement positif. En déduire l'existence, l'unicité, la position et la valeur du minimum de cette fonction.

IV.2.3 Décomposer les événements E et F comme des unions disjointes d'événements élémentaires de la forme $\{X=k\}$. Écrire les probabilités de E et F comme des sommes, et reconnaître les développements en séries entières de fonctions hyperboliques.

IV.2.4 Pour le 1., utiliser la formule des probabilités totales pour calculer $\mathbb{P}(X=i)$. Identifier la loi de X comme une loi de Poisson dont on identifiera le paramètre.
Pour le 2., calculer $\mathbb{P}(X=i, Y=j)$ et comparer avec $\mathbb{P}(X=i)\mathbb{P}(Y=j)$.
Pour le 3., décomposer l'événement en question comme une union dénombrable d'événements de la forme $\{X=i, Y=i\}$. On pourra reconnaître le développement en série entière d'une fonction tabulée (fonction de Bessel I_0).

IV.2.5 Utiliser le théorème de dérivation terme à terme d'une série entière.

IV.2.6 Pour le 1., il faut vérifier que :

$$\sum_{k=0}^{+\infty} \mathbb{P}(X=k) = 1.$$

Pour le 2., on peut se servir de l'exercice précédent.
Pour le 3., il faut calculer $\mathbb{P}(X>s)$.

IV.2.7 En notant A_k l'événement « il y a rencontre au tirage d'ordre k », utiliser la formule d'inclusion-exclusion.

IV.2.8 On peut chercher le nombre de tirages qui ont été nécessaires pour qu'il reste exactement r allumettes dans l'autre boîte lorsque le fumeur s'aperçoit qu'une boîte est vide. Puis on peut examiner comment

ont pu se répartir ces tirages (entre poche gauche et poche droite) pour que cet événement se produise.

La valeur de la somme $\sum_{r=0}^{N} u_r = 1$ peut s'obtenir sans calcul.

IV.2.9 Pour le 1(a), il faut remarquer que, pour $v \in \mathbb{N}^*$ et $w \in \mathbb{Z}$, l'événement $\{V = v, W = w\}$ coïncide avec $\{Y = v, X = v + w\}$ si $w \geqslant 0$ et avec $\{X = v, Y = v - w\}$ si $w \leqslant 0$. On calcule alors la loi jointe du couple (V, W) en utilisant l'indépendance de X et Y.

Pour le 1(b), utiliser la formule des marginales.

Pour le 1(c), vérifier que :

$$\mathbb{P}(V = v, W = w) = \mathbb{P}(V = v)\mathbb{P}(W = w)$$

pour tous v, w.

Pour le 2., calculer de deux façons différentes :

$$\frac{\mathbb{P}\left(\{X = n+1\} \cap \{Y = n\}\right)}{\mathbb{P}\left(\{X = n\} \cap \{Y = n\}\right)}.$$

IV.2.10 Pour le 1., on peut calculer $\mathbb{P}(X_m > n)$, en remarquant que $\{X_m > n\} = \{X_1 > n, X_2 > n\}$.

Montrer que X_m suit une loi géométrique, dont on identifiera le paramètre.

IV.2.11 Pour le 1., il faut déterminer λ de telle manière que $\sum_{i,j \in \mathbb{N}} p_{i,j} = 1$.

Pour le 2., utiliser la formule des marginales.

IV.2.12 Utiliser la formule des probabilités totales pour calculer $\mathbb{P}(S = s)$ en décomposant sur la partition $\{\nu = k\}$, $k \in \mathbb{N}^*$.

IV.2.13 Utiliser les relations entre les dérivées en 1 de la fonction génératrice et les moments d'une v.a. discrète.

Lorsque ν est déterministe, on doit retrouver les formules usuelles sur les sommes de v.a. indépendantes : $\mathbb{E}[S] = \nu\mathbb{E}[X]$ et $\mathrm{Var}(S) = \nu^2\mathrm{Var}(X)$.

IV.3 : Variables aléatoires à densité

IV.3.1 Déterminer d'abord la fonction de répartition de Y.

IV.3.2 Il faut déterminer a de telle manière que $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$.

Pour déterminer la loi de Y, on peut chercher sa fonction de répartition.

Pour calculer l'espérance de Y, on peut calculer l'intégrale $\int_{\mathbb{R}} yg(y)dy$ où g est la densité de Y, mais il vaut mieux se servir simplement du fait que $\mathbb{E}[Y] = \mathbb{E}[X^2]$ qui a déjà été calculé au 2.

IV.3.3 Il faut déterminer C de telle manière que $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$.

Pour le 3., calculer puis identifier la fonction de répartition de $G(U)$.

IV.3.4 Pour le 2., calculer l'intégrale par morceaux.

IV.3.5 Pour le 2.(a), on peut raisonner par l'absurde : si l'on suppose qu'il existe t tel que $\mathbb{P}(T > t) = 0$, alors on montre par récurrence sur n que $\mathbb{P}(T > t2^{-n}) = 0$, et les événements $\{T > t2^{-n}\}$ forment une suite décroissante.

Pour le 2.(b), on peut se servir de la monotonie de la fonction de répartition pour le passage de $\alpha \in \mathbb{Q}_+$ à $\alpha \in \mathbb{R}_+$.

IV.3.6 Pour le 2. on pourra introduire une v.a. discrète de loi uniforme sur $\{a_1, \ldots, a_n\}$.

IV.3.7 Pour le 2., utiliser la formule des marginales.
Pour montrer que deux v.a. réelles ne sont pas indépendantes, il suffit de trouver deux réels x et y tels que $\mathbb{P}(X > x)\mathbb{P}(Y > y) \neq \mathbb{P}(X > x, Y > y)$.
Pour le 4, il faut calculer les moments $\mathbb{E}[XY]$, $\mathbb{E}[X]$ et $\mathbb{E}[Y]$.

IV.3.8 On mesure les angles par rapport à la position de la cerise. Le découpage se fait selon deux angles θ_1 et θ_2 tirés uniformément sur $]0, 2\pi[$. En vous aidant d'un dessin, vérifier que la part contenant la cerise mesure $2\pi - (\theta_2 - \theta_1)$ si $\theta_2 > \theta_1$ et $2\pi - (\theta_1 - \theta_2)$ si $\theta_2 \leqslant \theta_1$. Donc cette part mesure $2\pi - |\theta_2 - \theta_1|$. Il ne reste plus qu'à faire les calculs.
La part contenant la cerise a une taille moyenne de $4\pi/3$.
La probabilité que la part contenant la cerise soit la plus grande est $3/4$.

IV.3.9 Pour calculer la loi de U, on peut calculer $\mathbb{P}(U > u)$ en utilisant le fait que $\{U > u\} = \{X > u, Y > u\}$.
Pour calculer la loi du couple (U, V), on peut se servir du fait que :

$$\begin{aligned}\{U > u, V > v\} &= \{X > u, Y > X + v\} \\ &\quad \cup \{X > Y + v, Y > u\}.\end{aligned}$$

Pour le 2., exprimer l'événement en question en termes de V et Z.
Pour le 3., exprimer T en termes de U et Z.

IV.3.10 Pour le 2., calculer le discriminant de l'équation du second degré, et calculer la probabilité qu'il soit strictement positif ou strictement négatif.

IV.3.11 Pour le 1., calculer la fonction de répartition de Y.

IV.3.12 Considérer la fonction définie sur $]0, +\infty[$ à valeurs dans $[0, 1]$:

$$\sigma \mapsto f(\sigma) := \mathbb{P}(X \in [a-1, a+1])$$

pour $X \sim \mathcal{N}(0, \sigma^2)$. Vérifier que cette fonction est de classe $\mathcal{C}^1$ et établir son tableau de variation. Montrer que f admet un unique maximum et identifier-le.
Le résultat est : $\sigma^2 = \dfrac{1}{2a} \ln\left(\dfrac{a+1}{a-1}\right)$.

IV.3.13 Utiliser le fait que :

$$\{X > x\} = \bigcap_{i=1}^{n} \{U_i > x\}$$

et :

$$\{Y \leqslant y\} = \bigcap_{i=1}^{n} \{U_i \leqslant y\}$$

pour calculer les fonctions de répartition de X et Y.

IV.4 : Théorèmes limites et estimation

IV.4.1 La moyenne et la variance empiriques sont $\bar{X}_{10} = 5{,}23$ et $V_{10} = 0{,}78$. Un intervalle de confiance au niveau 90% de la moyenne est :

$$\left[\bar{X}_{10} - \frac{\sqrt{V_{10}}t_{0,9}}{\sqrt{10}}, \bar{X}_{10} + \frac{\sqrt{V_{10}}t_{0,9}}{\sqrt{10}}\right],$$

où $\mathbb{P}(-t_{0,9} < Z < t_{0,9}) = 0{,}9$ avec $Z \sim T_9$.
Un intervalle de confiance au niveau 90% de la variance est :

$$\left[\frac{9V_{10}}{t_{0,9,1}}, \frac{9V_{10}}{t_{0,9,0}}\right],$$

où $\mathbb{P}(Z < t_{0,9,0}) = 0{,}05$ et $\mathbb{P}(Z < t_{0,9,1}) = 0{,}95$ avec $Z \sim \chi^2_9$.
Pour le cas avec 100 mesures, il faut utiliser les approximations gaussiennes des lois T_{99} et χ^2_{99}.

IV.4.2 Chaque erreur a une chance sur 1000 d'être à la page considérée. Les erreurs sont indépendantes. On peut en déduire la loi du nombre d'erreurs sur la page considérée.

IV.4.3 Pour le 1., il faut majorer $\mathbb{E}[(S_n/n - x)^2]$ uniformément en $x \in [0,1]$.
Pour le 2., il faut se servir du fait qu'une fonction sur un intervalle compact est bornée d'une part, uniformément continue d'autre part.
Pour le 3., en notant

$$f_n(x) = \sum_{k=0}^{+\infty} e^{-nx} \frac{(nx)^k}{k!} \cos\left(\frac{k}{n}\right),$$

il faut montrer que $f_n(x_n)$ ne tend pas vers 0 quand n tend vers $+\infty$.

IV.4.4 Relire la section « Sondages » du module. Le résultat est que la fourchette sur les intentions de vote sera de $\pm 1\%$ avec le niveau de confiance 95% si $n \geqslant 10^4$.

IV.4.5 Pour le 1. et le 2., relire la section « Sondages » du module.
Pour le 2., chercher d'abord une fonction h telle que $\mathbb{P}_p(p \geqslant h(M_n)) \geqslant 0{,}95$. La réponse est $m_0 = 527$.
Pour le 3., on cherche à évaluer $p = (p_1 + p_2)/2$. Calculer la variance de l'estimateur usuel $\hat{p}_n = M_n/n$ d'une part et la variance du nouvel estimateur :

$$\check{p}_n = \frac{\check{p}_{f,n/2} + \check{p}_{h,n/2}}{2},$$

où $\check{p}_{f,n/2} := M_{f,n/2}/(n/2)$ et $M_{f,n/2}$ est le nombre de femmes parmi les $n/2$ interrogées qui déclarent voter pour B. De même, $\check{p}_{h,n/2} := M_{h,n/2}/(n/2)$ et $M_{h,n/2}$ est le nombre d'hommes parmi les $n/2$ interrogés qui déclarent voter pour B.

IV.4.6 Dans le B3., commencer par montrer par récurrence que $a_n \leqslant a_c$.
Dans le C3.(b), montrer et utiliser le fait que A_∞ est la limite de la suite croissante d'événements A_n.
Dans le C6.(b), évaluer le signe de la dérivée seconde de G pour montrer la convexité. Pour montrer que G a au plus deux points fixes, on peut raisonner par l'absurde en supposant que G a au moins trois points fixes sur $[0,1]$, et utiliser le théorème de Rolle pour aboutir à une contradiction.
Pour le C6.(d), si $m \leqslant 1$, montrer que $G(z) > z$ pour tout $z < 1$. Si $a < 1$, montrer que $m > 1$.
Pour le C7., utiliser l'équation (18).

Index

A

aire 220
aléatoire
    générateur pseudo — 16
    somme — 748
    vecteur — 730, 770
algébrique (clôture —) 508
algorithme
    d'Euclide 5, 53
    de Box-Muller 789
    de Cornacchia 34
    du pivot sur un anneau euclidien 6
alternée (matrice —) 77
anisotrope (forme quadratique —) 91
anneau
    élément irréductible d'un — 7
    élément premier d'un — 7
    euclidien 4
    factoriel 7
    polynôme sur un — factoriel 10, 12
    principal 4, 5, 7
application d'Artin-Schreier 97
approximants de Padé 621
Arf (invariant de —) 97, 119, 121
Artin-Schreier (application d'—) 97
Artin-Whapples (théorème d' —) 530
atome d'hydrogène 600
automorphisme de Frobenius 21, 71
autosimilaire 518

B

babylonienne (tablette —) 510
Balmer (série de —) 600
base
    Hilbertienne 340
    symplectique 81
Bayes (formule de —) 684
Bernoulli
    loi de — 710
    schéma de — 739
Bessel (polynômes de —) 556, 568, 580, 632, 652
biais d'un estimateur 811
Bienaymé-Tchebychev (inégalité de —) 723, 769
binaire
    forme quadratique — 89, 95
binomiale
    loi — 711
    négative (loi —) 741
    variable aléatoire — 712
Borel (tribu de —) 322, 668
boréliens (sous-ensembles —) 322
Box-Muller (algorithme de —) 789

C

Cantor (ensemble de —) 270
caractéristique 2
    formes quadratiques en — 89, 120
Carmichael (nombres de —) 14, 29, 43
carré sommable (fonction de —) 336
Cauchy
    conditions de — 376
    critère de — 265
    transformée de — 641

Cauchy-Schwarz (inégalité de —) 338
centrale (théorème de la limite —) 803
certificat
de composition 46
de primalité 46
changement de variable 215, 269
Chasles
relation de — 264, 275, 343
χ_n^2 (loi du) 782
chiffres 431, 439
Christoffel (nombres de —) 550
classe(s)
d'un discriminant 89
nombre de — 89
NP 46
P 46
clôture algébrique 508
coût du calcul
de $a^m \pmod{n}$ 41
du pgcd 41
du symbole de Jacobi 42
cohérente (suite —) 441
composition
certificat de — 46
test de — 44
comptage (mesure de —) 371
conditionnelle (probabilité —) 680
conditionnement multiple 685
cône
de révolution 134
elliptique 158
confiance (intervalle de —) 816
congruentes (matrices —) 92
conique (nappe —) 151
conoïde (nappe —) 152
consistance d'un estimateur 811
contenu d'un polynôme 12
continue (fraction —) 56, 641
continuité d'une fonction définie par une intégrale 305
convergence
théorème de — dominée 301, 320
théorème de — encadrée 300, 305, 319
théorème de — monotone 296, 318
Cornacchia (algorithme de —) 34
corps
extension finie d'un — 508
formellement réel 492
résiduel 448
de $\mathbb{Z}[\mathrm{i}]$ 71
corrélation 733
courbe
paramétrée 130
tracée sur une nappe 130
Cousin (lemme de —) 342
covariance 733, 776
création-annihilation (opérateurs de —) 593
crible d'Ératosthène 42
critère
de Cauchy 265
pour les intégrales 343
de Darboux 186
critique (point —) 161
cryptographie à clé publique 18
cubable (partie —) 198
cyclique (groupe —) 14
cylindre
elliptique 162
hyperbolique 162
parabolique 140, 164
cylindrique (nappe —) 151

D

Darboux
critère de — 186
sommes de — 185
théorème de — 186
Darboux-Christoffel (formule de —) 543
De Morgan (formules de —) 667
décomposé (nombre premier —) 30
décomposition de Witt 103
découpage 247
définie (forme quadratique —) 91
degré d'une extension 508
densité 205
de probabilité 758

dérivabilité d'une fonction définie par une intégrale 306
descente infinie 33
déterminant
de Hankel 617
et volume 353
difféomorphisme
direct 132
indirect 132
diophantienne (équation —) 66
Dirac (masse de —) 554
Dirichlet
conditions de — 376
fonction de — 260
discrète (valuation —) 494
discriminant
classes d'un — 89
d'une forme quadratique 94
dominée
théorème de convergence — 301, 320
droite-gauche (procédé —) 247

E

écart-type 722
écriture
décimale 431
impropre 432
propre 432
égalité de Parseval 340
Eisenstein (critère d'—) 69
ellipsoïde 158
elliptique
paraboloïde — 140, 162
point — 139
empirique
moyenne — 812
variance — 813
non biaisée 815
encadrée (théorème de convergence —) 300, 305, 319
engendrée
tribu — 322
ensemble
de représentants des irréductibles d'un anneau factoriel 9
fondamental 665
négligeable 187
au sens de Riemann 187
entiers de Gauß 29
équation
aux dérivées partielles de Laplace 585
aux différences 701
de Schrödinger 597
stationnaire 597
des ondes 564
différentielle
d'Hermite 571
de Jacobi 569
de Laguerre généralisée 570
des fonctions d'Hermite 592
diophantienne 66
réduite 159, 163, 164
équivalentes
formes quadratiques — 90
formes symplectiques — 77
erf (fonction —) 762
escalier du diable 270
espace
de probabilité 669
hyperbolique 102
quadratique 90
symplectique 77
tangent 136, 146
ultramétrique 452
espace quadratique
non dégénéré 91
espérance 717, 764
estimateur 810
de la moyenne 812
de la variance 813
étagée (fonction —) 317
états stationnaires 597
Euclide
algorithme d'— 53
Euclide (algorithme d'—) 5

euclidien(ne)
algorithme du pivot sur un anneau — 6
anneau — 4
division — 4
stathme — 4
Euler
indicatrice d'— 13
série d'— 645
théorème d'— 14
événements 666
deux à deux indépendants 688
incompatibles 667
indépendants 686
mutuellement indépendants 689
exponentielle
loi — 761, 768
série génératrice — 604
exposant d'un groupe 70
extension
algébrique d'un corps 508
degré d'une — 508
finie d'un corps 508

F

factoriel (anneau —) 7
factorisation (difficulté de —) 20
Fatou (lemme de —) 299, 319
Fermat
nombres premiers de — 50
petit théorème de — 14
témoin de — 44
fonction
absolument intégrable 286
arithmétique multiplicative 14, 70
continue à support compact 336
d'Hermite 591
de Cantor-Lebesgue 270
de carré sommable 336
de Dirichlet 260
de Heaviside 555
de répartition 554, 757
en escalier 178
erf 762
étagée 317
Gamma 240
génératrice 604, 741
des polynômes de Laguerre 606
des polynômes de Legendre 606
harmonique 585
négligeable 313
poids 533
sphérique 585
sphérique zonale 587
théorème des —s implicites 149
Zeta de Riemann 38
fondamental (ensemble —) 665
forçage de points marqués 250, 252
forme de Sturm-Liouville 557, 560
forme(s) quadratique(s)
anisotrope 91
binaire 89
binaires 95
définie 91
discriminant d'une — 94
en caractéristique 2 89, 120
équivalentes 90
non dégénérée 91
noyau d'une — 91
rang d'une — 91
formellement réel
corps — 492
formes symplectiques 77
formes symplectiques
équivalentes 77
forme bilinéaire
alternée
noyau d'une — 77
rang d'une — 77
non-dégénérée 77
formule(s)
d'inclusion-exclusion 677
d'intégration par parties
généralisée 574
d'Olinde Rodrigues 574
de Bayes 684
de changement de variable 269

de Darboux-Christoffel 543
de Green-Riemann 214
de Hadamard 463
de Huyghens 767
de Legendre 523
de quadrature de Gauß 549
de Taylor 138
des marginales 731, 771
des probabilités totales 682
fractales 518
fraction continue 56, 57, 641
convergent d'un développement en — 62
développement en — d'un réel 61
développement en — d'une série formelle 64
développement en — dans $\mathbb{Q}$ et dans $K(X)$ 56
générale 59
quotient complet d'un développement en — 59, 61
quotient incomplet d'un développement en — 56, 61
réduite d'une — 58
simple 57, 58
simple propre 58
Frobenius (automorphisme de —) 21, 71
Fubini (théorème de —) 208, 345

G

Gamma (fonction —) 240
Gauß
entiers de — 29
formules de quadrature de — 549
lemme de — 10
loi de — 761, 768
méthode de — 549
gaussien(ne)
loi — normale 768
vecteur — 780
Gegenbauer (polynômes de —) 581, 657
générateurs
d'un groupe cyclique 14
linéaires congruentiels 17
pseudo-aléatoires 16
génératrice
fonction — 604, 741
série — 604
exponentielle 604
géométrie des nombres 26
géométrique
loi — 715
variable aléatoire — 715
gradient 146
grands nombres (loi des —) 798
Green-Riemann (formule de —) 214
groupe
cyclique 14
des unités 448

H

Hadamard (formule de —) 463
Hankel (déterminants de —) 617
harmonique
fonction — 585
sphérique 587
Hasse (invariant de —) 115
Hasse-Minkowski
théorème de — 112
Heaviside (fonction de —) 555
Heisenberg
principe d'incertitude de — 596
relation d' — 594
Hensel (série de —) 442
Henstock (lemme de —) 279, 280
Henstock-Kurzweil (intégrale de —) 252
Hermite
fonctions d' — 591
polynômes d' — 535
hessienne (matrice —) 139
Hilbert (symbole de —) 113
histogrammes 709
horizontalement simple
ouvert — 213, 214
Huygens (formule de —) 767

hyperbolique
espace — 102
paraboloïde — 140, 153, 162
plan — 99
point — 139
hyperboloïde
à deux nappes 159
à une nappe 155, 158
hypergéométrique (loi —) 717

I

implicites (surfaces —) 144
impropre (écriture —) 432
inclusion-exclusion (formule d' —) 677
indépendant(e)s 686, 689
événements — 686, 689
variables aléatoires — 735, 773
indicatrice d'Euler 13
indice de Witt 110
inégalité(s)
de Bienaymé-Chebychev 723, 769
de Cauchy-Schwarz 338, 733, 778
de Jensen 769
de Markov 769
inerte (nombre premier —) 30
intégrable
fonction — au sens de Riemann 180
partie — 311
intégrale 182
de Henstock-Kurzweil 252
de Laguerre 645
inférieure 180
multiple 175, 341
supérieure 180
intégration par parties 269
intervalle de confiance 816
invariant
de Arf 97, 119, 121
de Hasse 115
inversion (théorème d'— locale) 138
involutif (sous-espace —) 79
irréductible (élément — d'un anneau) 7
isotrope
sous-espace — 77, 91
vecteur — 91

J

Jacobi
cas de — 569
matrice de — 539
symbole de — 28
jacobien 132, 215
jauge 249
Jensen (inégalité de —) 769

L

Lagrangien (sous-espace —) 79
Laguerre
intégrale de — 645
polynômes de — 535
Laplace
équation aux dérivées partielles de —) 585
laplacien 654
sphérique 586
Lebesgue
mesure de — 311, 352, 676
tribu de — 322
Legendre
formule de — 523
polynômes de — 535
symbole de — 22
lemme
de Cousin 342
de Fatou 299, 319
de Gauß 10
de Henstock 280
de Morse 142, 149, 161
de recouvrement de Vitali 283
limite
inférieure 299
supérieure 299
théorème de la — centrale 803
logarithme p-adique 465, 525

loi(s)
binomiale 711
binomiale négative 741
d'une variable aléatoire discrète 707
d'une variable aléatoire réelle 757
de Bernoulli 710
de Gauß 761, 768
de Poisson 729
de réciprocité quadratique 25
de Student 785
des grands nombres 798
diffuse 759
du χ_n^2 782
exponentielle 761, 768
faible des grands nombres 798
forte des grands nombres 798
gaussienne normale 768
gaussienne ou normale 761
géométrique 714
hypergéométrique 717
multinomiale 731, 732
normale 768
uniforme 759, 767

M

marginales (formule des —) 731, 771
Markov (inégalité de —) 769
masse 205
de Dirac 554
matrice(s)
alternée 77
congruentes 92
de Jacobi 539
meilleure approximation quadratique (polynôme de —) 547
Mersenne (nombres premiers de —) 52
mesurable (partie —) 321
mesure 553
de comptage 371
de Lebesgue 311, 352, 676
extérieure 325
modèle statistique 810
moment(s)
d'inertie 205
problème des — 618
série génératrice des — 633, 637
monotone
théorème de convergence — 318
Morse (lemme de —) 142, 149, 161
moyenne empirique 812
multinomiale (loi —) 731
multiple (point —) 129
multiplicative
fonction arithmétique — 14, 70

N

nappe
cartésienne 128
conique 151
conoïde 152
cylindrique 151
de révolution 153
géométrique 131
paramétrée 128
réglée 151
simple 129
négligeable
ensemble — 187
au sens de Riemann 187
fonction — 313
Neumann (conditions de —) 376
nombre(s) premier(s) 3
de Fermat 50
de Mersenne 52
décomposé 30
il y a une infinité de — 37
inerte 30
répartition des — 38
ramifié 30
représentable sous la forme $x^2 + 2y^2$ 67
représentable sous la forme $x^2 + dy^2$ 71
représentable sous la forme $x^2 + xy + y^2$ 71
représentable sous la forme $x^2 + y^2$ 31
théorème des — 38

nombres
de Carmichael 14, 29, 43
de Christoffel 550
décadiques 441
entiers décadiques 441
parfaits 52
quantiques 601
non dégénéré(e)
espace quadratique — 91
forme bilinéaire — 77
forme quadratique — 91
normale
à une surface 147
loi — 768
normalisation 534
norme algébrique 30
noyau
d'une forme bilinéaire alternée 77
d'une forme quadratique 91
NP (classe de complexité) 46

O

Olinde Rodrigues (formule d' —) 574
opérateur
de création-annihilation 593
différentiel 558
nombre de particules 594
nombre de quanta 594
ordre
d'une écriture 442
d'un nombre b-adique 450
orientable (nappe —) 132
orthogonal(e)
réflexion — 105
renversement 105
symétrie — 105
orthogonaux 77, 91
polynômes — 534
oscillateur harmonique 596
oscillation d'une fonction 191
ouvert
horizontalement simple 213, 214
verticalement simple 213, 214

P

P (classe de complexité) 46
P = NP ? 46
p-adique
logarithme — 465, 525
valuation — 449, 451, 494
Padé (approximants de —) 621
paraboloïde
elliptique 140, 162
hyperbolique 140, 153, 162
paradoxe de Banach-Tarski 327
paramétrage 129
parfait (nombre —) 52
Parseval (égalité de —) 340
partie
cubable 198
mesurable 321
pavable 196
partition 667
pas d'une subdivision 176
pavable (partie —) 196
pavé 176
Pépin (test de primalité de —) 51
Pfaffien 82
formule explicite pour le — 83
universel 83
plan
hyperbolique 99
quadratique 90
symplectique 80
tangent 134, 146
plus fine (subdivision —) 178
poids (fonction —) 533
point(s)
critique 161
elliptique 139
hyperbolique 139
marqué 246
multiple 129
parabolique 139
régulier 133, 144
simple 129
singulier 133, 144

Poisson (loi de —) 729
Poisson-Charlier (polynôme de —) 653
polynôme(s)
contenu d'un — 12
continuant 73
d'Hermite 535, 571
de Bessel 556, 568, 580, 632, 652
de Gegenbauer 581, 657
de Jacobi 569
de Laguerre 66, 535, 570
de Laguerre généralisés 570
de Legendre 535
de meilleure approximation quadratique 547
de Poisson-Charlier 653
de Tchebychev 563, 657
orthogonaux 534
classiques 555
de seconde espèce 540
primitif 12
sur un anneau factoriel 10, 12
premier
élément — d'un anneau 7
nombre — 3
presque partout 188
primalité
certificat de — 46
test de — 42, 47
test de — de Solovay-Strassen 29
PRIMES
est dans NP 46
est dans P 48
primitif (polynôme) 12
primitives (racines — modulo p) 15
principal (anneau —) 4, 5, 7
Prix du millénaire de la fondation Clay
Fonction Zeta 38
P = NP ? 46
probabilité(s) 668
conditionnelle 680, 681
totales (formule des —) 682
uniforme 672, 675
problème des moments 618
procédé droite-gauche 247
propre (écriture —) 432

Q

quadratique
espace — 90
loi de réciprocité — 25
plan — 90
résidu — 22
quadrique 156
à centre 156
quantiques (nombres —) 601
quaternions 36, 838

R

racines primitives modulo p 15
ramifié (nombre premier —) 30
rang
d'une forme bilinéaire alternée 77
d'une forme quadratique 91
réalisation 665
réciprocité (loi de — quadratique) 25
recouvrement de Vitali 283
récurrence (relation de — des polynômes orthogonaux) 537
réduites d'une fraction continue 58, 641
réflexion orthogonale 105
règle de correspondance 596
réglée (nappe —) 151
régulier
point — 133, 144
sous-espace — 77, 92
relation
canonique de commutation 596
d'Heisenberg 594
de Chasles 264, 275, 343
de récurrence des polynômes orthogonaux 537
renversement orthogonal 105
répartition (fonction de —) 554, 757
représentants (ensemble de — des irréductibles d'un anneau factoriel) 9
représente 90
résidu quadratique 22

résiduel (corps —) 448
révolution (nappe de —) 153
Riemann (sommes de —) 186
risque quadratique moyen d'un estimateur 812
RSA 18
fiabilité 20

S

schéma de Bernoulli 739
Schrödinger (équation de —) 597
seconde espèce
polynômes orthogonaux de — 540
séparation des variables 586
méthode de — 564
série
d'Euler 65, 645
de Hensel 442
de Laurent
(partie entière d'une —) 64
formelle (développement en fraction continue d'une —) 64
génératrice 604
des moments 633
des moments décalée 637
exponentielle 604
simple
nappe — 129
point — 129
simplification (théorème de —) 110
singulier
point — 133, 144
sous-espace — 77, 92
Solovay 327
Solovay-Strassen
test de primalité de — 29, 45
somme
aléatoire 748
de Darboux 185
de deux carrés 33
de deux parties 483
de quatre carrés 35
de Riemann 186
de trois carrés 35
directe orthogonale 90, 92
sondage 824
sous-espace
involutif 79
isotrope 77, 91
Lagrangien 79
régulier 77, 92
singulier 77, 92
totalement isotrope 91
maximal 91
sous-variété 145
sphérique
fonction — 585
harmonique — 587
laplacien — 586
squeeze theorem 266
stathme euclidien 4
stationnaire
équation de Schrödinger — 597
Stieltjes (transformation de —) 635
Student (loi de —) 785
Sturm-Liouville (forme de —) 557, 560
subdivision 176, 246
adaptée 178
partielle 280
plus fine 178
pointée 176, 246
produit 178
suite
cohérente 441
croissante d'événements 671
surfaces
implicites 144
tangentes 148
transverses 148
symbole
de Hilbert 113
de Jacobi 28
de Legendre 22
symétrie 105
orthogonale 105
symplectique
base — 81
espace — 77
forme — 77
plan — 80

T

table de la loi gaussienne 806
taille d'un entier 41
tangent(e)(s)
droite — 134
plan — 134, 146
surfaces — 148
Taylor (formules de —) 138
Tchebychev (polynôme de —) 563, 657
témoin
de Fermat 44
de non-primalité 45
test de composition 44
test de primalité 42, 47
de Pépin 51
de Solovay-Strassen 29, 45
théorème
d'Artin-Whapples 530
d'Egorov 332
d'Euler 14
d'inversion locale 138
de convergence dominée 301, 320
de convergence encadrée 300, 305, 319
de convergence monotone 296, 318
de Darboux 186
de Fischer-Riesz 335
de Fubini 208, 345
de Hake 277
de Hasse-Minkowski 112
de la limite centrale 803
de Lebesgue-Denjoy 285
de Lebesgue-Vitali 193
de simplification de Witt 110
de Witt 109
des fonctions implicites 149
des nombres premiers 38
fondamental de l'analyse 267
petit — de Fermat 14
tore 154
totalement isotrope
sous-espace — 91
totales (formule des probabilités —) 682
transformation de Stieltjes 635
transformée de Cauchy 641
transverses (surfaces —) 148
tribu 322, 667, 668
de Borel 322, 668
de Lebesgue 322
engendrée 322
troncature 432
à gauche 441

U

ultramétrique (espace —) 452
uniforme
loi — 759, 767
probabilité — 672, 675
unités (groupe des —) 448
univers 665

V

valuation 10, 494
p-adique 449, 451, 494
discrète 494
variable(s) aléatoire(s)
binomiale 712
géométrique 715
indépendantes 735, 773
intégrable 717
réelle 755
à densité 758
variance 722, 767
empirique 813
non biaisée 815
vecteur
aléatoire 730, 770
gaussien 780
isotrope 91
verticalement simple (ouvert —) 213, 214
vibration
d'un cercle 564
zonale 564
Vitali (recouvrement de —) 283
volume 196, 220, 349, 353
d'un ouvert 349
Von Neumann, John 16

W

Witt
- décomposition de — 103
- indice de — 110
- théorème de — 109

Z

zéros des polynômes orthogonaux 544
zonale
- fonction sphérique — 587
- vibration — 564