UNIVERSITÉ DE LIÈGE
FACULTÉ DES SCIENCES APPLIQUÉES

APPENDICES COMMUNS AUX COURS

Théorie de l'information et du codage
Apprentissage inductif appliqué
Introduction aux processus stochastiques

Louis WEHENKEL

Octobre 1999. Version provisoire.

Copyright © Université de Liège, Belgique. 1998

Table des matières

A MOTIVATION

Le but de ces appendices est de collationner un certain nombre d'éléments du calcul des probabilités, de statistique et d'algèbre linéaire dont la bonne maîtrise est nécessaire pour la compréhension des méthodes stochastiques, couvertes dans les cours suivants

- Théorie de l'information et du codage
- Introduction aux processus stochastiques
- Apprentissage inductif appliqué
- Applications des méthodes stochastiques

L'expérience montre que ces matières, enseignées dans les premières années à l'université, sont souvent mal assimilées et/ou oubliées, sans doute faute d'avoir été mises en pratique suffisamment rapidement après avoir été apprises.

Alors que, traditionnellement, peu de place était laissée aux méthodes stochastiques dans l'enseignement des ingénieurs, ces méthodes sont aujourd'hui utilisées dans la très grande majorité des activités humaines. Plus particulièrement, les méthodes stochastiques sont nécessaires pour la modélisation, l'analyse et la conception des systèmes complexes tels que les réseaux informatiques et électriques. Elles jouent également un grand rôle en économie et en finance, ainsi que dans les sciences naturelles.

Organisation des appendices

Nous avons organisé ces appendices de rappels de façon modulaire, d'une part afin de mettre clairement en évidence les différentes disciplines, d'autre part, afin de faciliter la mise à jour progressive de certaines parties, susceptibles de devoir s'adapter dans le futur.

Les deux premières appendices traitent respectivement du calcul des probabilités et de notions élémentaires de statistique. Nous avons intentionnellement séparé les rappels de calcul de probabilités des rappels de statistique, de façon à mettre en évidence deux approches très différentes. En effet, le calcul des probabilités est une discipline purement déductive, telle que d'autres volets des mathématiques dont il fait partie. Même si l'interprétation de la notion de probabilité est sujette à débat, la théorie abstraite en est assez simple du point de vue conceptuel. Il en est tout autrement en ce qui concerne la statistique. Il s'agit d'une discipline extrêmement vaste, d'une

part, très vivante en ce qui concerne la recherche, d'autre part. De plus, et c'est sans doute ce qui importe le plus, elle ne peut se dissocier des problèmes réels auxquels elle s'applique. La statistique consiste en effet à recueillir des données sur des systèmes réels et à les interpréter, ce qui nécessite, certes, une bonne dose de savoir faire. Il s'en suit que l'apprentissage de la statistique passe certainement par la maîtrise du calcul des probabilités, mais surtout par les applications aux problèmes réels. Il faut également noter le rôle très important joué par l'informatique, dont les développements récents ont révolutionné la pratique de la statistique en permettant le traitement d'ensembles de données de très grande taille et la mise en oeuvre d'algorithmes de plus en plus sophistiqués.

Les deux dernières appendices fournissent respectivement des rappels de calcul vectoriel et matriciel, et une brève introduction aux espaces de Hilbert. Nous nous contentons de rappeler les notions et résultats fréquemment utilisés dans le cadre des méthodes stochastiques.

Il va de soi, que les notions traitées dans ces appendices seront (re)introduites au fur et à mesure des besoins dans le cadre des cours susmentionnés. Cependant, il nous a semblé utile de les présenter ici de façon indépendante et consistante.

Enfin, nous nous devons d'avertir le lecteur que nous avons volontairement restreint au stricte nécessaire les sujets abordés dans ces appendices. Nous n'avons nullement la prétention de nous substituer à la très vaste et souvent très bonne littérature qui couvre de manière plus complète et plus systématique les sujets abordés ici. Nous indiquerons en lieu voulu quelques références permettant au lecteur intéressé d'en savoir plus sur ce domaine passionnant.

Le stochastique en général

On fait appel aux méthodes stochastiques lorsqu'on est en présence de phénomènes qu'il n'est pas possible ou peu pratique d'étudier de façon détaillée et déterministe. C'est notamment le cas lorsque les systèmes étudiés présentent une très grande complexité, ou lorsqu'on ne dispose que d'une connaissance partielle de leurs caractéristiques. Les méthodes stochastiques permettent alors d'étudier les comportements en moyenne, en modélisant de façon probabiliste les parties d'un système qu'on ne souhaite pas ou qu'on ne peut pas décrire en détails. Les méthodes stochastiques fournissent les outils nécessaires pour déduire les distributions de probabilités des grandeurs de sortie importantes, en fonction de celles des entrées et du modèle (déterministe ou non) du système. Elles permettent ensuite d'utiliser au mieux ces informations pour prendre des décisions appropriées.

Pourquoi enseigner les méthodes stochastiques aux ingénieurs ?

Le besoin de méthodes stochastiques est reconnu depuis assez longtemps (enseignement du calcul des probabilités et des statistiques en candidature ingénieur, depuis quelques décennies). Cependant, cet enseignement a très peu évolué au cours du temps, et surtout le volume associé est resté constant. Il apparaît que la partie théorique relative au calcul des probabilités est mieux assimilée par les étudiants, alors que la statistique est assez vite oubliée, ce qui tient au fait qu'on ne l'utilise pas assez par après.

Si on fait un inventaire des travaux de fin d'études des ingénieurs électriciens (électronique, informatique, génie électrique) on se rend compte qu'une proportion importante fait appel de façon directe ou indirecte aux méthodes stochastiques.

Enfin, si on s'interroge sur le rôle des futurs ingénieurs, on se rend compte qu'on leur demande un esprit critique et une capacité d'innovation de plus en plus grande. Or, il est reconnu que l'enseignement actuel des ingénieurs, qui est surtout basé sur le raisonnement déductif (appliquer des "lois" générales aux cas particuliers), tend à étouffer l'imagination et la créativité. L'enseignement du stochastique a pour premier objectif d'engendrer une plus grande ouverture d'esprit et de renforcer la capacité de raisonnement inductif (c'est-à-dire à tirer des conclusions intéressantes à partir de cas particuliers). Il apparaît donc comme souhaitable, sinon nécessaire, pour renforcer la capacité d'innovation des ingénieurs.

Domaines de l'ingénieur électricien faisant appel au stochastique

Les télécommunications reposent sur les méthodes stochastiques en ce qui concerne l'optimisation des performances, le codage, et le filtrage du bruit. En particulier, les télécommunications font largement appel au traitement

du signal, aux techniques d'optimisation des performances de systèmes informatiques distribués, à la compression et au codage de données. Par exemple, dans un réseau ATM chaque connexion se présente sous la forme d'une suite virtuellement unique de cellules transmises, qui peut être représentée comme la réalisation d'un processus stochastique. Les méthodes stochastiques permettent alors d'étudier les performances du système lorsqu'il est soumis à différents types de trafic (communications téléphoniques, transferts de données numériques, trafic multimedia...). Elles permettent aussi l'optimisation du codage des données dans le but de minimiser les pertes d'informations suite au bruitage en cours de transmission.

Le traitement du signal (filtrage, traitement de la parole et de signaux physiologiques, traitement d'images) repose en grande partie sur des méthodes stochastiques. En imagerie spatiale, par exemple, ces techniques sont utilisées pour éliminer le bruit des informations brutes captées par les télescopes et ainsi identifier automatiquement les corps stellaires. En médecine, elles permettent l'interprétation automatique de signaux physiologiques (électrocardiogrammes, électroencéphalogrammes) et facilitent ainsi la surveillance des malades.

La théorie des systèmes est une discipline générale qui est utilisée pour l'étude et la conception d'une très grande diversité de systèmes, que ce soit en mécanique, en électricité, ou encore en informatique. Une partie de la théorie des système s'intéresse aux systèmes stochastiques qui sont notamment utilisés en estimation d'état et pour la conception de systèmes auto-adaptatifs. L'estimation d'état, permet de tirer le meilleur profit des informations fournies par divers capteurs, notamment en filtrant les erreurs de mesures et en permettant la détection de fonctionnements anormaux de certains capteurs. Les systèmes auto-adaptatifs sont capables d'adapter leur stratégie de commande en fonction de changements des caractéristiques de l'environnement, du système piloté, et/ou de ses objectifs de réglage.

En informatique de nombreuses questions font directement appel aux méthodes stochastiques : l'optimisation des performances des systèmes, la compression de données, l'intelligence artificielle, l'analyse de données, les réseaux de neurones... Par exemple, l'analyse de données est utilisée par certains constructeurs automobiles afin de détecter les raisons pour lesquelles certaines pannes répétitives sont observées; les méthodes stochastiques permettent dans ce contexte d'identifier les facteurs qui provoquent effectivement des pannes des nombreuses autres informations disponibles dans les bases de données. La compression des données est basée sur le codage de suites de symboles en fonction de leur probabilité d'apparition, les suites les plus fréquentes recevant les codes les plus courts; elle permet de réduire coûts de stockage et délais de transmission dans de nombreuses applications (stockage de masse, réseaux informatiques, disques compacts, multimedia, télévision digitale...).

La gestion des risques industriels et technologiques fait appel aux méthodes stochastiques pour l'analyse et la maîtrise de la fiabilité et de la sécurité des systèmes. Les méthodes stochastiques sont notamment utilisées pour la conception des centrales nucléaires, la planification des réseaux d'énergie électrique, l'évaluation des risques des moyens de transport en commun (aéronautique, trains à grande vitesse), la maîtrise de la fiabilité des lanceurs spatiaux... Ces techniques permettent notamment d'identifier les sources de pannes les plus probables et de déterminer des parades à la fois efficaces et aussi économiques que possibles. Notons que l'étude des performances des logiciels informatiques fait partie de ce domaine.

Disciplines de base

- **Théorie de l'information :** étude quantitative de la notion d'incertitude et d'information; optimisation des performances des systèmes de codage et de transmission de l'information.

- **Processus stochastiques :** méthodes statistiques de description des signaux temporels et techniques de traitement du signal.

- **Apprentissage automatique :** théorie de l'estimation statistique étendue à l'estimation de fonctions générales de plusieurs variables.

- **Sécurité et fiabilité des systèmes :** évaluation et amélioration de la capacité des systèmes à fonctionner de façon satisfaisante, malgré la possibilité de défaillances de certaines de leurs parties, et malgré les inévitables perturbations en provenance de leur environnement.

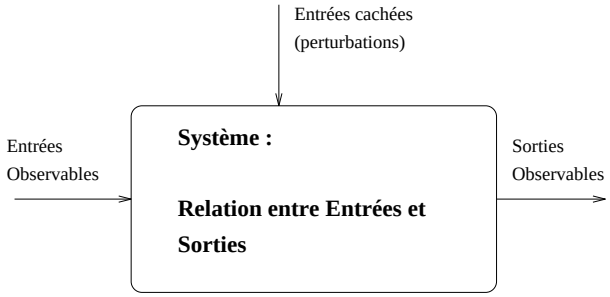

Figure A.1. Représentation graphique d'un système

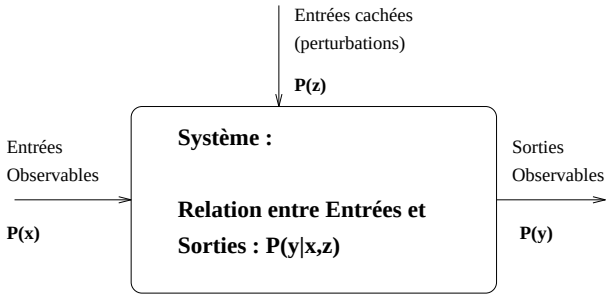

Figure A.2. Représentation probabiliste d'un système

Notion de système

La figure A.1 représente de façon graphique un système. Une telle représentation est en fait une abstraction (graphique) de la réalité physique; pour en faire l'étude on peut lui associer des objets mathématiques : espace d'entrée; espace de sortie; modèle entrée/sortie (équations différentielles, algébriques...). Si à des entrées fixées le modèle associe de manière unique les sorties, on dira que le modèle (et par extension le système) est déterministe. Un modèle non-déterministe est donc un modèle qui associe à une entrée donnée un ensemble de sorties possibles, c'est-à-dire compatibles avec le modèle. Notons d'emblée que *l'orientation* du modèle, c'est-à-dire le choix de ce qu'on convient d'appeller les entrées et les sorties est effectué en fonction des objectifs poursuivis. En particulier, un modèle déterministe peut devenir non-déterministe si on inverse le rôle joué par les entrées et les sorties. D'autre part, même dans le cas d'un modèle déterministe, il est souvent difficile ou impossible de déterminer toutes les entrées de manière précise (par exemple, certaines entrées sont qualifiées de *perturbations* inconnues); dans ce cas, on cherchera à déterminer les ensembles de sorties possibles, lorsque les entrées appartiennent à un sous-ensemble de l'espace d'entrée.

On conçoit que partant d'un système réel, on peut construire une hiérarchie de modèles comprenant des modèles très précis mais extrêmement complexes, voire impossibles à manipuler, ainsi que diverses approximations, plus simples à manipuler mais moins précises. Par exemple, en présence d'un simple circuit électrique (disons une "résistance" en série avec un "condensateur"[1] (Les notes sont regroupées à la fin de chaque chapitre)), l'ingénieur qui s'intéresse à la relation entre le courant dans ce circuit et la tension à ses bornes dispose de toutes une série de modèles : équations de Maxwell, équation différentielle des circuits en éléments condensés (linéaire ou non-linéaire), équation algébrique, relations qualitatives... Même le plus sophistiqué de ces modèles reste une abstraction de la réalité, et possède un domaine de validité restreint, sinon bien cerné. C'est tout l'art du métier d'ingénieur que de choisir le modèle approprié, à la fois suffisamment précis et adapté aux besoins pratiques, et ce choix dépend évidemment du contexte. Par exemple, dans le cas de notre mini-circuit il dépendra de l'espace d'entrées (contenu fréquentiel des signaux à l'entrée, amplitude), ainsi que de la nature de l'environnement (perturbations thermiques ou électromagnétiques) dans lequel le circuit est censé fonctionner. Mais le choix de modélisation dépend également de l'information disponible : par exemple la mise en oeuvre des équations de Maxwell nécessite des informations détaillées sur la géométrie des conducteurs et diélectriques des composants utilisés, informations qui ne sont pas nécessairement disponibles.

La figure A.2 schématise le point de vue adopté par les méthodes stochastiques pour l'étude des systèmes, anticipant sur la suite de ces appendices. Par rapport à la figure A.1, ce modèle est complété par des hypothèses sur les distributions de probabilités des entrées $P(x), P(z)$, la relation entrée/sortie est représentée par une distribution conditionnelle $P(y|x,z)$, et les sorties sont caractérisées par une distribution de probabilités $P(y)$. Ces distributions de probabilités représentent un certain niveau de connaissance du système physique modélisé, et englobent comme cas particuliers les systèmes déterministes[2]. L'avantage principal des cette vision des choses est de mettre en évidence explicitement le degré de non-déterminisme des différentes parties du modèle.

L'évolution de la technique engendre des systèmes de plus en plus complexes (systèmes de télécommunications, systèmes informatiques, réseaux d'énergie électrique…); l'amélioration des connaissances physiques fournit également des modèles de plus en plus sophistiqués. Cependant, la maîtrise de systèmes complexes réels passe le plus souvent par la mise en oeuvre de modèles "ad hoc", simplifiés pour en permettre la manipulation. Les méthodes stochastiques visent à quantifier les incertitudes résiduelles de ces modèles en mettant en oeuvre le calcul des probabilités et les outils statistiques. Elles permettent de construire des modèles probabilistes des systèmes à partir de données mesurées, et ensuite de manipuler ces modèles (analyse mathématique, simulations numériques) pour prendre des décisions.

Notes

1. Nous utilisons ici ces termes pour désigner les composants physiques et non les abstractions correspondantes de la théorie des circuits.

2. Notons que dans la représentation probabiliste de la figure A.2 nous avons fait implicitement l'hypothèse que les entrées observables et non-observables étaient indépendantes.

B RAPPELS DE PROBABILITES

"La théorie des probabilités n'est rien d'autre que le bon sens réduit sous forme de calcul."
- Pierre Simon Laplace, 1819

B.1 PROBABILITES VERSUS STATISTIQUE

La théorie des probabilités est une branche des mathématiques qui étudie les propriétés des structures (mathématiques) permettant de représenter les phénomènes où le "hasard" intervient.

Cette théorie permet donc de modéliser efficacement certains phénomènes aléatoires et d'en faire l'étude théorique. Elle fournit également une approche pour la formalisation du "raisonnement" en présence d'informations partielles et/ou contradictoires. Comme toute théorie mathématique, la théorie des probabilités est une science déductive, qui se base sur un certain nombre d'axiomes et utilise les techniques usuelles en mathématiques pour la démonstration de théorèmes. On y déduit donc des propriétés spécifiques, à partir d'hypothèses générales. Ses domaines d'application sont nombreux : la physique, l'intelligence artificielle, la théorie des systèmes, le traitement du signal, la statistique ... pour n'en citer que quelques uns.

La statistique, au sens le plus général, est une discipline qui consiste dans le recueil, le traitement et l'interprétation de données d'observations sur des systèmes physiques (réels ou simulés). En particulier, elle permet de construire des modèles (probabilistes ou non) qui représentent correctement la réalité mesurable du monde physique. Il s'agit d'une discipline faisant souvent appel au raisonnement inductif : à partir d'un certain nombre d'observations élémentaires on cherche à construire des lois générales qui "expliquent" ces observations. Etant donné ce caractère inductif, les résultats obtenus par la statistique peuvent être remis en question par de nouvelles observations, comme c'est le cas dans le domaine des sciences naturelles en général. Pour cette raison, une utilisation correcte de la statistique dans un domaine donné, va nécessairement de pair avec une bonne compréhension physique de ce domaine. Aussi, les résultats obtenus sont justifiés dans la mesure où ils sont opérationnels, et non pas parce qu'ils représenteraient la vérité absolue. Qu'on ne s'y trompe pas cependant, car l'utilisation des outils statistiques fait autant appel à la rigueur scientifique que les autres sciences expérimentales. Néanmoins, ces outils ne permettent de vérifier que la "plausibilité" (et non la "réalité") de la plupart des modèles utilisés par les ingénieurs et scientifiques de nombreuses disciplines. Etant donné la diversité des problèmes rencontrés en pratique, la statistique est un domaine extrêmement vaste dont l'appréhension d'ensemble nécessite du temps et de l'expérience pratique. Elle est surtout basée sur le calcul des probabilités, qui lui sert comme outil de raisonnement; elle fait cependant aussi appel à de nombreuses autres parties des mathématiques (analyse, algèbre,...).

Il y a donc une interdépendance forte entre les deux disciplines, mais également une différence fondamentale dans leur approche : déductive pour le calcul des probabilités; inductive pour la statistique.

Cet appendice s'intéresse d'abord au calcul des probabilités pour en rappeler les bases et les résultats les plus fondamentaux qui doivent être maîtrisés. Les rappels de statistique élémentaire font l'objet d'une appendice

séparée. Pour une très bonne introduction aux probabilités et à la statistique nous recommandons vivement [Sap90]. Pour en savoir plus sur les fondements mathématiques du calcul des probabilités nous suggérons la lecture de la référence [Bil79].

Pour conclure cette introduction, insistons sur le fait que la séparation probabilités/statistique que nous faisons volontairement n'est pas justifiée d'un point de vue fondamental, mais bien pour des raisons pédagogiques. Nous commençons en quelque sorte par analyser quelques arbres sans nous préoccuper de la forêt. Parmi les différentes possibilités qui s'offrent pour aborder un domaine comme celui-ci, le choix que nous avons fait est celui d'une rupture minimale (mais nécessaire) avec la façon habituelle de présenter probabilités et statistique dans un même cours, et sans séparation claire.

Nous souhaitons ainsi limiter la confusion entre les aspects déductifs et inductifs complémentaires du calcul des probabilités et de la statistique. Au fur et à mesure de sa familiarisation avec les méthodes stochastiques, l'étudiant prendra conscience comment ces deux disciplines couvrent les deux pans du *raisonnement en présence d'informations incomplètes*. Enfin, pour terminer ces commentaires introductifs, notons que le domaine des méthodes stochastiques est étroitement apparenté à la logique et à la philosophie des sciences, même si l'approche probabiliste ne constitue pas la seule réponse possible aux problèmes abordés dans ces domaines.

B.2 NOTION DE PROBABILITE - INTERPRETATIONS

Dans cette section nous allons introduire, tout d'abord intuitivement puis plus formellent la notion de probabilité. Ensuite nous discuterons très brièvement de ses différentes interprétations logiques et physiques.

B.2.1 Intuitivement

Comme mentionné plus haut, le calcul des probabilités est un outil mathématique qui permet de représenter et de manipuler des situations/expériences dont l'issue est aléatoire et/ou au sujet desquelles on dispose de connaissances incomplètes/incertaines.

Une connaissance (c'est-à-dire une affirmation logique) est dite *incertaine* dans un contexte donné, si dans ce contexte il est impossible aussi bien de réfuter sa véracité que de la prouver. La notion de probabilité permet d'ordonner de telles connaissances par ordre de *plausibilité* croissante, et de remettre à jour cet ordonnancement lorsque de nouvelles informations deviennent disponibles.

Expériences aléatoires. Une *expérience* est qualifiée d'aléatoire si on ne peut pas prévoir par avance son résultat, et donc si, répétée dans des conditions apparemment identiques, elle pourrait donner lieu à des *résultats* différents. Le calcul des probabilités permet de modéliser et de simuler de telles expériences.

Pour étudier une telle expérience on s'intéresse tout d'abord à l'univers de tous les résultats (ou objets) possibles : on note usuellement Ω cet ensemble fondamental, et ω un élément particulier de Ω, c'est-à-dire un résultat particulier parmi ceux possibles.

Exemple 1. Par exemple, si on s'intéresse au diagnostic médical, l'expérimentateur pourrait être un médecin particulier, et l'expérience le premier diagnostic de l'année (disons, le 2 janvier au matin, en l'an 2000). Nous pourrions alors définir l'ensemble Ω pour ce problème comme l'ensemble de tous les patients que ce médecin est susceptible de diagnostiquer (le médecin ne peut évidemment pas prévoir quel sera le patient particulier qui va se présenter devant lui).

Exemple 2. Un autre exemple intéressant concerne les réseaux informatiques. Plaçons nous en un noeud particulier du réseau Internet, et observons les messages par courrier électronique qui y transitent pendant une journée donnée. Un résultat particulier est alors la suite particulière de messages qui ont transité pendant la période d'observation. Avant d'avoir effectué l'expérience on ne peut évidemment pas prévoir quelle suite sera observée, et l'ensemble Ω est alors l'ensemble de toutes les suites de messages possibles pouvant transiter sur une journée, un ensemble certes très compliqué à caractériser mais néanmoins de taille finie.

Il faut noter que l'univers est défini en fonction de l'objectif particulier poursuivi. Ainsi, dans le premier exemple ci-dessus on aurait pu définir l'univers comme étant l'ensemble des maladies diagnostiquées par le médecin,

ou encore l'ensemble des médicaments préscrits. Dans le second exemple, on aurait pu s'intéresser uniquement à l'expéditeur des messages et définir le résultat de l'expérience comme étant l'ensemble des adresses email des expéditeurs ayant envoyé au moins un message email pendant la journée : l'univers serait alors l'ensemble de tous les sous-ensembles d'expéditeurs possibles de messages électroniques susceptibles de transiter par le noeud.

Evénements. Dans la terminologie usuelle, un **événement** désigne une assertion logique vérifiable relative au résultat d'une expérience. Ainsi, l'assertion logique suivante

> **le premier patient qui se présentera le 2 janvier 2000 au matin est un étudiant qui devrait passer un examen le 3 janvier**

définit un événement. Cette assertion logique est soit vraie soit fausse, et définit en fait un sous-ensemble des joyeux lurons parmi la "clientèle" encore en vie le 2 janvier 2000; nous allons noter cet ensemble $A \subset \Omega$.

Un autre événement correspondrait à l'assertion logique suivante

> **le premier patient . . . a trop festoyé le jour du réveillon et souhaite un certificat médical**

auquel on peut également associer un sous-ensemble $B \subset \Omega$, a priori différent de A.

A tout événement on peut donc faire correspondre un sous-ensemble de Ω. En particulier, à tout $\omega \in \Omega$ on peut associer un **événement élémentaire** correspondant au singleton $\{\omega\}$.

Probabilités. La notion de probabilité qui sera formalisée ci-dessous est une fonction qui mesure l'importance des événements : elle associe à un événement un nombre positif (entre 0 et 1) qui représente le degré de certitude qu'on peut associer à celui-ci a priori. Il traduit l'état de connaissance dans lequel on se trouve avant de réaliser une expérience.

Notons que si l'univers comprend un nombre fini d'éléments, les événements sont forcément des ensembles finis. Dans ce cas, un événement auquel on associe une probabilité égale à un est un événement dit certain (on est certain qu'il se réalisera); symétriquement, un événement auquel on associe une probabilité nulle est un événement impossible : on est certain qu'il ne se réalisera pas. Ces deux cas extrêmes sont les limites où le raisonnement probabiliste rejoint la logique classique : la partie intéressante concerne cependant tous les événements auquels on associe des probabilités intermédiaires. La mesure de probabilité permet de trier l'ensembles des événements par ordre croissant de leur probabilité a priori. Le calcul des probabilités permet de s'assurer que les raisonnements effectués à l'aide de ces nombres restent cohérents, d'une part, d'autre part, il permet de mettre à jour les probabilités en fonction des informations obtenues.

Remarque sur la notion d'expérience répétable. Les expériences que nous avons illustrées ci-dessus ne peuvent pas en principe être répétées : le 2 janvier 2000 au matin il n'y aura qu'un seul patient qui sera le premier. De même, au cours d'une journée donnée un seul ensemble de messages transitera en un noeud donné d'Internet.

Il est cependant souvent possible de supposer que les propriétés de certaines expériences ne changent pas au cours du temps : on peut alors répéter (éventuellement indéfiniment) cette expérience. Par exemple, on peut supposer qu'un dé ne s'use pas au cours du temps et que le résultat d'une expérience de lancer de dé ne dépend pas du temps, ou du nombre de fois qu'il a déjà été lancé. Similairement, lorsqu'on répète un certain nombre de fois une opération de mesure, on peut supposer que l'ensemble des résultats possibles ne change pas d'une fois à la suivante et que les probabilités des événements restent constantes.

Il est clair, néanmoins, que ce type de situation est une abstraction qui ne se réalise jamais parfaitement en pratique : il n'est pas possible d'observer un système physique sans le perturber. Nous verrons plus loin que cette abstraction est souvent vérifiée approximativement et est à la base d'une grande partie des statistiques. Nous allons pour le moment au moins admettre qu'un expérience peut être répétée. Nous discuterons à la section B.2.3 plus finement pourquoi il n'en est pas toujours ainsi, et pourquoi il est intéressant de se servir du calcul des probabilités lorsque ce n'est pas le cas.

B.2.2 Formellement

B.2.2.1 σ-Algèbre des événements.

Notations. Ci dessous nous utiliserons

- des lettres minuscules grecques ($\alpha, \beta, \ldots$) pour désigner les éléments de Ω
- des lettres majuscules latines (p.ex. $A, B, \ldots$) pour désigner des sous-ensembles de Ω.

Par ailleurs, nous désignons par

- 2^Ω l'ensemble de tous les sous-ensembles de Ω
- des lettres rondes (p.ex. $\mathcal{A}, \mathcal{B}, \ldots$) des sous-ensembles de 2^Ω, c'est-à-dire des ensembles de sous-ensembles de Ω.

Enfin, nous utiliserons également les notations

- $f(\cdot), g(\cdot), \ldots$ pour désigner une fonction définie sur (une partie de) Ω,
- $F(\cdot), G(\cdot), \ldots$ pour désigner des fonctions définies sur (une partie de) 2^Ω.

Enfin, nous désignerons par $\neg A$ le complémentaire dans Ω de A.

Définitions. Un σ-algèbre $\mathcal{E}$ d'événements[1] défini sur un univers Ω est une partie de 2^Ω (i.e. un ensemble de sous-sensembles de Ω) qui vérifie les propriétés suivantes :

- $\Omega \in \mathcal{E}$;
- $A \in \mathcal{E} \Rightarrow \neg A \in \mathcal{E}$;
- $\forall A_1, A_2, \ldots \in \mathcal{E}$ (en nombre fini ou dénombrable[2]) : $\bigcup_i A_i \in \mathcal{E}$.

Les éléments de $\mathcal{E}$ sont désignés par le terme d'événements. Il s'agit de parties de Ω auxquelles nous avons conféré un statut particulier, comme nous le verrons ci-dessous. Deux événements A et B sont dits *incompatibles* si $A \cap B = \emptyset$.

Système complet d'événements. $A_1, \ldots A_n$ forment un *système complet d'événements* si $\forall i \neq j \ : A_i \cap A_j = \emptyset$ (ils sont incompatibles deux à deux) et si $\bigcup_i^n A_i = \Omega$ (ils couvrent Ω). On dit aussi que les A_i forment une partition de Ω, et on supposera la plupart du temps que tous les A_i sont non-vides.

Remarques.

1. Les deux premières propriétés ci-dessus impliquent que l'ensemble vide (désigné par $\emptyset$) fait nécessairement partie de tout σ-algèbre d'événements.
2. La seconde et la troisième propriété impliquent également que $\bigcap_i A_i \in \mathcal{E}$.
3. $\{\emptyset, \Omega\}$ est un σ-algèbre d'événements : c'est le plus petit de tous.
4. 2^Ω est un σ-algèbre d'événements : c'est le plus grand de tous.
5. Si Ω est un ensemble fini, alors $\mathcal{E}$ l'est également.
6. Par contre, si Ω est infini (dénombable ou non), $\mathcal{E}$ peut être non-dénombrable, dénombrable, et même fini.

B.2.2.2 Probabilités.

Le statut particulier des événements est qu'il est possible de leur attribuer une probabilité, c'est-à-dire un nombre positif compris entre 0 et 1, qui doit répondre aux axiomes suivants.

Axiomes de Kolmogorov. On appelle probabilité sur $(\Omega, \mathcal{E})$ (ou loi de probabilité) une fonction $P(\cdot)$ définie sur $\mathcal{E}$ telle que :

- $P(A) \in [0, 1], \forall A \in \mathcal{E}$;
- $P(\Omega) = 1$;
- $\forall A_1, A_2, \ldots \in \mathcal{E}$ incompatibles : $P(\bigcup_i A_i) = \sum_i P(A_i)$.

Discussion.

On voit que l'utilisation du calcul des probabilités passe par trois étapes successives de modélisation : définition de l'univers Ω, choix d'un σ-algèbre d'événements $\mathcal{E}$, et enfin quantification par le choix de la mesure de probabilité. Les propriétés de base qui sont requises pour que l'ensemble soit cohérent sont le fait que $\mathcal{E}$ soit effectivement un σ-algèbre et que $P(\cdot)$ satisfasse les axiomes de Kolmogorov.

On constate également que le calcul des probabilités est compatible avec la logique classique (en tout cas, si Ω est fini). Il suffit de considérer le cas particulier où $P(\cdot)$ est définie sur $\{0, 1\}$, et associer à la valeur 1 la valeur de vérité "vrai" et à 0 la valeur "faux".

B.2.2.3 Propriétés remarquables.

Des axiomes de Kolmogorov on peut déduire immédiatement les propriétés suivantes (qu'on démontrera à titre d'exercice, en se restreignant au cas où Ω est fini).

1. $P(\emptyset) = 0$.
2. $P(\neg A) = 1 - P(A)$.
3. $A \subset B \Rightarrow P(A) \leq P(B)$.
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
5. $P(\bigcup_i A_i) \leq \sum_i P(A_i)$.
6. $A_i \downarrow \emptyset \Rightarrow \lim_i P(A_i) = 0$[3]

On a également :

1. $P(A) = 1 \Rightarrow P(A \cup B) = 1, \forall B \in \mathcal{E}$.
2. $P(A) = 1 \Rightarrow P(A \cap B) = P(B), \forall B \in \mathcal{E}$.
3. $P(A) = 0 \Rightarrow P(A \cap B) = 0, \forall B \in \mathcal{E}$.
4. $P(A) = 0 \Rightarrow P(A \cup B) = P(B), \forall B \in \mathcal{E}$.

B.2.2.4 Théorème des probabilités totales.

Soit B_i un système complet d'événements, alors $\forall A \in \mathcal{E}$: $P(A) = \sum_i P(A \cap B_i)$.

Remarque. Certains auteurs définissent une système complet d'événements de façon légèrement différente de celle que nous avons adpotée ci-dessus. Au lieu d'exiger que $\bigcup_i^n A_i = \Omega$ ils imposent la condition plus faible $P(\bigcup_i^n A_i) = 1$. Ce type de système, obéit également au théorème des probabilités totales.

D'ailleurs, on peut évidemment compléter un tel système avec $A_{n+1} = \neg \bigcup_i^n A_i$, avec $P(A_{n+1}) = 0$.

B.2.3 Différentes interprétations de la notion de probabilité

Il faut faire la distinction entre la formulation mathématique d'une théorie et l'utilisation que nous en faisons pour étudier des problèmes du monde réel qui nous entoure, c'est-à-dire son interprétation. Ceci est particulièrement vrai pour une théorie telle que le calcul des probabilités qui vise entre autres à modéliser une certaine forme du raisonnement humain, et qui s'adresse à des problèmes où l'incertitude joue un rôle fondamental, c'est-à-dire des problèmes où il pourrait être difficile de valider la théorie. En particulier, la théorie ne nous aide pas lorsqu'il s'agit de définir en pratique la loi de probabilité à associer aux événements choisis.

En réalité, depuis son origine, le calcul des probabilités a donné lieu à des débats intenses entre scientifiques, logiciens, physiciens, philosophes, en ce qui concerne la ou les interprétations physiques à donner à la notion même de probabilité. Ces débats sont encore d'actualité, et le resteront certainement encore longtemps; c'est la raison pour laquelle nous voulons mettre en évidence ici les différents points de vues qui s'opposent dans ce débat d'idées.

B.2.3.1 Le point de vue objectiviste.

La vision classique. La vision classique est héritée des jeux de hasard. Dans cette vision Ω est en général fini, et on considère alors comme σ-algèbre d'événements 2^Ω, fini lui aussi. Nous décrivons en détail la démarche adoptée pour alors définir la mesure de probabilité, car cette démarche est également utilisée dans la conception subjectiviste.

Dans ce cas, il suffit d'attribuer des probabilités aux événements élémentaires $P(\omega); \forall w \in \Omega$; les probabilités des autres événements s'en déduisent par application du troisième axiome de Kolmogorov. En particulier , on aura $P(\Omega) = \sum_\omega P(\omega)$ ce qui en vertu du second axiome de Kolmogorov impose évidemment que $\sum_\omega P(\omega) = 1$.

Sous cette contrainte, la vision classique impose des arguments de symétrie, posant que les événements élémentaires sont équiprobables. Par conséquent, $P(\omega) = |\Omega|^{-1}$ (où $|\Omega|$ désigne la taille finie de l'univers).

C'est cette démarche qui conduit à associer aux 6 faces d'un dé "parfait", une probabilité de $\frac{1}{6}$.

La principale faiblesse de cette approche est qu'elle repose sur un postulat de symétrie idéal (donc irréalisable en pratique) et ne permet pas la remise en question des probabilités en fonction d'informations supplémentaires (obtenues par exemple en effectuant des expériences de lancer de dé). Une autre faiblesse est que cette approche ne s'étend pas au cas ou Ω est non-dénombrable (voir à ce sujet la discussion au §B.2.4).

La vision fréquentiste. Elle repose sur la loi des grands nombres et sur une autre idéalisation, à savoir la notion d'expérience indéfiniment répétable dans les mêmes conditions. La loi des grand nombres assure en effet que dans une telle expérience la fréquence relative observée d'un événement converge vers la probabilité de celui-ci.

Notons que cette vision est aussi appelée la vision "orthodoxe" : toute comme dans la vision classique, la notion de probabilité est supposée définie de façon unique (c'est-à-dire indépendamment de l'observateur), et tous les observateurs doivent se soumettre à une expérience similaire pour en déterminer la valeur.

Il est clair que la procédure expérimentale n'est pas pratiquement réalisable. D'autre part, elle n'autorise pas l'utilisation du calcul des probabilités pour raisonner sur des événements incertains mais non répétables (et en pratique, aucun événement n'est parfaitement répétable). Enfin, elle est basée sur un cercle vicieux logique : cette définition repose sur la loi des grands nombres qui elle-même suppose déjà défini le concept de probabilité.

B.2.3.2 Le point de vue subjectiviste.

Les faiblesses des deux approches précédentes et le fait que d'un point de vue logique il soit souhaitable de permettre la remise en question de la probabilité d'un événement suite à l'obtention de nouvelles informations (par exemple, si nous apprenons que le dé est imparfait) conduisent à nier l'existence de la notion de probabilité "objective".

Ainsi, dans la conception subjectiviste on modélise l'état de connaissance d'un observateur. On peut alors argumenter que pour être cohérent avec lui-même, un observateur doit assigner des probabilités aux événements qui respectent les axiomes de Kolmogorov, mais, différents observateurs, ayant éventuellement des connaissances différentes, peuvent aboutir à des assignations différentes. De plus, un même observateur peut remettre à jour ces probabilités lorsque de nouvelles informations se présentent.

Mesure d'incertitude. La probabilité objective n'existe pas et n'est donc pas une grandeur mesurable; la probabilité subjective est simplement une mesure d'incertitude, qui peut varier avec les circonstances et avec l'observateur.

Puisque la notion d'expérience répétable n'est plus nécessaire, on peut étendre le domaine d'application du calcul des probabilités aux événements non répétables, c'est-à-dire au raisonnement en présence d'incertitudes (par exemple en intelligence artificielle, pour modéliser le raisonnement humain).

La vision bayesienne. Nous reviendrons plus loin sur cette approche, après avoir développé le calcul des probabilités. Pour le moment, contentons nous d'indiquer que cette approche consiste à attribuer des probabilités à tout ce qui est incertain. En particulier, cette approche consiste à attribuer des lois de probabilités aux probabilités des événements, si les informations disponibles ne sont pas suffisantes pour déterminer leurs valeurs exactes.

Ainsi, en présence d'un problème de jeu de "pile ou face", un bayesien va commencer par admettre qu'il ne connait pas suffisamment bien la pièce pour fixer a priori la probabilité de "pile". En d'autre mots, il admet avoir une incertitude sur la valeur de cette probabilité, qu'il va modéliser par une (meta)loi de probabilités. Ensuite, il

va utiliser cette loi de probabilités pour faire des prédictions, et si des expériences sont effectuées (par exemple des lancers de pièce) il va utiliser le calcul des probabilités (la formule de Bayes) pour remettre à jour la valeur des méta-probabilités en fonction de l'issue de l'expérience.

Il faut remarquer que cette approche n'est pas non plus entièrement satisfaisante (au grand damne de ses défenseurs) puisqu'il reste une phase arbitraire qui consiste à choisir les méta-probabilités. Signalons simplement plus haut que certains arguments de symétrie et d'"esthétique" sont utilisés par les bayesiens pour fixer de façon "objective" les méta-probabilités...

Cependant, sans prendre parti disons qu'il nous semble que l'approche bayesiennne présente une certaine souplesse qui va de pair avec la démarche scientifique.

B.2.4 Ensembles infinis, voire non-dénombrables

Pour terminer ces commentaires philosophiques, et avant de nous attaquer au coeur du calcul des probabilités, nous voulons faire ici quelques remarques générales sur la nécessité de pouvoir manipuler des lois de probabilités définies sur des univers de taille infinie ou dénombrables.

Lorsque l'ensemble Ω est fini, l'algèbre des événements l'est également. Par contre, lorsque Ω est infini, il est possible d'y définir des algèbres d'événements finis, dénombrables ou non-dénombrables. Par exemple, si Ω est infini mais reste dénombrable (c'est-à-dire peut être mis en bijection avec l'ensemble $\mathbb{N}$ des entiers naturels), l'algèbre complet est non-dénombrable (il peut être mis en bijection avec l'ensemble $\mathbb{R}$).

Dans la pratique, l'application du calcul des probabilités aux ensembles infinis peut conduire à un certain nombre de difficultés conceptuelles, dues aux passages à la limite. En particulier, un événement de probabilité nulle n'est pas nécessairement impossible, et corrolairement un événement de probabilité égale à un n'est pas nécessairement certain. Par exemple, si nous prenons comme univers l'intervalle $]0, 1]$ de la droite réelle, muni de l'algèbre induit par les semi-intervalles $]a, b]$ ($a < b \in [0, 1]$) [4], et muni de la loi de probabilité uniforme qui associe à un intervalle $]a, b]$ la probabilité $b - a$, les événements élementaires de $]0, 1]$ sont des singletons de probabilité nulle et non moins possibles. Dans de telles situations il faut utiliser quelques précautions oratoires et parler d'événements presque impossibles ou presque certains.

D'un point de vue conceptuel, il est cependant important de se souvenir que ces ensembles infinis sont des constructions mathématiques obtenues par passage à la limite sur des ensembles finis. Le langage mathématique associé à ces ensembles permet d'écrire de façon synthétique des propriétés qui sont vraies pour les ensembles finis et qui le restent lors du passage à la limite. Mais, ce language ne doit pas changer la signification de propriétés, ni nous induire en erreur.

D'un point de vue pratique, on peut donc adopter le point de vue que le monde tel qu'il est accessible à l'expérimentation physique est essentiellement fini (c'est d'ailleurs évident en ce qui concerne le monde de l'informatique digitale). On pourrait donc parfaitement justifier une approche qui consisterait à développer les théories sur base de modélisations par ensembles finis, et qui expliciterait les passages à la limite sur les résultats plutôt que sur les concepts de départ. On pourrait alors se débarasser des difficultés engendrées par l'analyse moderne (calcul infinitésimal, théorie de la mesure, des distributions...) au prix d'une lourdeur d'écriture accrue (et souvent excessive) d'un certain nombre de propriétés et de raisonnements.

Nous pensons que l'utilisation de l'analyse classique est un outil mathématique qui est non seulement intéressant du point de vue conceptuel, mais également justifié par son caractère opérationnel. Cependant, la compréhension des principes de base importants dans le domaine du calcul de probabilités peut par contre très bien se faire sans y faire appel à tours de bras. En clair, nous suggérons aux étudiants d'effectuer leurs raisonnements dans le cadre d'univers finis, afin de bien assimiler la signification mathématique et physique des principales notions. Une fois bien maîtrisé le cas fini, ils pourront ensuite se poser la question de savoir ce qui se passe lors du passage à la limite.

B.3 ELEMENTS DE BASE DU CALCUL DE PROBABILITES

B.3.1 Loi de probabilité conditionnelle

Partons d'un espace probabilisé $(\Omega, \mathcal{E}, P(\cdot))$, et supposons que l'on sache qu'un événement B est réalisé. Cherchons à savoir ce que devient alors la probabilité qu'un événement A quelconque soit réalisé, que nous allons noter $P(A|B)$.

Si A et B sont incompatibles il est clair que A ne se réalisera pas et $P(A|B) = 0$. Par contre, si $A \cap B \neq \emptyset$, alors la réalisation de A est possible, mais seule la partie de A qui est dans B nous intéresse. Si $A \cap B = B$ alors nous sommes certains que A se réalisera : $P(A|B) = 1$ dans ce cas. Tout se passe donc comme si nous avions restreint notre univers à l'événement B et que nous nous intéressions uniquement aux probabilités relatives des parties des événements situées dans B.

Définition. Nous supposerons que B est de probabilité non-nulle (dans le cas fini il est ridicule d'envisager qu'un événement de probabilité nulle se soit réalisé), et nous *définissons* la probabilité conditionnelle de A sachant que B est réalisé par

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}. \tag{B.1}$$

Notons que $A \supset B \Rightarrow P(A|B) = 1$, mais (attention) la réciproque est fausse!

Il s'agit bien d'une loi de probabilité. En effet, elle vérifie les axiomes de Kolmogorov puisque :

- $A, B \in \mathcal{E} \Rightarrow A \cap B \in \mathcal{E}$ et donc $P(A|B)$ est bien définie sur $\mathcal{E}$.
- $P(A|B) \geq 0$.
- $A \cap B \subset B \Rightarrow P(A \cap B) \leq P(B) \Rightarrow P(A|B) \leq 1$.
- $P(\Omega|B) = 1$ (trivial).
- $P\left(\bigcup_i A_i|B\right) = \frac{P\left(\left(\bigcup_i A_i\right)\cap B\right)}{P(B)} = \frac{P\left(\bigcup_i A_i \cap B\right)}{P(B)} = \sum_i \frac{P(A_i \cap B)}{P(B)} = \sum_i P(A_i|B)$, car si les A_i sont incompatibles les $A_i \cap B$ le sont également.

Evénements indépendants. On dit que A est indépendant de B si $P(A|B) = P(A)$, c'est-à-dire si le fait de savoir que B est réalisé ne change en rien la probabilité de A. On utilise souvent la notation

$$A \perp B \tag{B.2}$$

pour indiquer que A est indépendant de B.

Si $P(B) \in]0, 1[$, on a A indépendant de B si, et seulement si, $P(A|B) = P(A|\neg B)$.
Suggestion : calculer alors $P(A)$ par le théorème des probabilités totales.

Indépendance conditionnelle. Soient A, B, C trois événements, et $P(C) \neq 0$. Alors, on dit que A *est indépendant de B conditionnellement à C*, que l'on note par

$$A \perp B \mid C \tag{B.3}$$

si $P(A|B \cap C) = P(A|C)$.

Notons que

$$P(A|B \cap C) \triangleq \frac{P(A \cap (B \cap C))}{P(B \cap C)} = \frac{P((A \cap B) \cap C)}{P(C)} \frac{P(C)}{P(B \cap C)} = \frac{P(A \cap B|C)}{P(B|C)}.$$

Discussion. Il est important de remarquer que la loi de probabilité conditionnelle est bien définie sur l'algèbre de départ $\mathcal{E}$, et ceci bien que ses valeurs ne dépendent en fait que de probabilités de sous-ensembles de B.

Pour deux événement donnés A et B non indépendants on peut avoir soit $P(A|B) < P(A)$ ou $P(A|B) > P(A)$. Un événement peut donc devenir plus ou moins certain lorsque on dispose d'informations nouvelles.[5]

Dans le cas d'un univers fini tous les événements possibles sont de probabilité strictement positive, et définissent par conséquent une loi de probabilité conditionnelle. Nous avons déjà illustré un cas d'univers infini, où le fait qu'un événement se réalise n'implique pas nécessairement que sa probabilité a priori était non-nulle; il est alors nécessaire de recourir à un artifice pour définir la notion de probabilité conditionnelle vis-à-vis de tels événements (passage à la limite).

B.3.2 Sur la notion d'indépendance

La notion d'indépendance est une notion centrale en théorie des probabilités. Aussi allons nous détailler les diverses propriétés immédiates qui découlent de sa définition. Nous supposerons ci-dessous que $P(A) \in]0,1[$ (resp. $P(B) \in]0,1[$), et dans le cas contraire nous dirons que A (resp. B) est un événement trivial. Nous laissons au lecteur le soin de vérifier dans quels cas (et comment) ces conditions peuvent être relaxées à des événements triviaux.

Propriétés positives. Nous demandons au lecteur de démontrer, à titre d'exercice immédiat, celles parmi les propriétés suivantes dont nous ne donnons pas la preuve.

- $\emptyset$ est indépendant de tout autre événement.
- Un événement de probabilité nulle est indépendant de tout autre événement : $P(A) = 0 \Rightarrow P(A \cap B) = 0 \Rightarrow P(A|B) = 0$.
- Tout événement est indépendant de Ω.
- Tout événement est indépendant de tout événement certain : $P(A) = 1 \Rightarrow P(A \cap B) = P(B)$ et donc $P(B|A) = P(B)$.
- "A indépendant de B" $\Leftrightarrow P(A \cap B) = P(A)P(B)$ *(conséquence directe de la définition).*
- "A indépendant de B" $\Rightarrow$ "B indépendant de A".
- "A indépendant de B" $\Rightarrow$ "$\neg A$ indépendant de B".
- "A indépendant de B" $\Rightarrow$ "A indépendant de $\neg B$".
- "A indépendant de B" $\Rightarrow$ "$\neg A$ indépendant de $\neg B$".

On peut donc utiliser en lieu et place de la définition de l'indépendance la définition suivante.

Définition alternative de l'indépendance. Deux événements A et B sont indépendants si $P(A \cap B) = P(A)P(B)$.

Il est à noter que cette définition peut être étendue au cas où $P(A)$ et/ou $P(B)$ sont nulles, la propriété étant trivialement vérifiée dans ce cas (car $P(A \cap B) \leq min\{P(A), P(B)\}$).

Propriétés négatives. L'assimilation de celles-ci sont aussi importantes pour la bonne compréhension de la notion d'indépendance.

Remarquons tout d'abord que des événements indépendants pour une loi de probabilité donnée peuvent très bien être non indépendants pour une autre loi de probabilité. En d'autres mots, la propriété d'indépendance dépend bien du choix de la loi de probabilité et pas seulement des propriétés ensemblistes.

Nous suggérons au lecteur de chercher des contre-exemples pour démontrer les propriétés négatives suivantes.

- Un événement quelconque non trivial n'est jamais indépendant de lui-même!
- A indépendant de B et B indépendant de $C \not\Rightarrow A$ indépendant de C
  (suggestion : prendre A et B non-triviaux indépendants et $C = A$).
- A dépendant de B et B dépendant de $C \not\Rightarrow A$ dépendant de C
  (suggestion : prendre A et C indépendants, et $B = A \cap C$ non trivial.)
- A indépendant de $B \not\Rightarrow A$ indépendant de B conditionnellement à C
  (suggestion : prendre comme exemple le double "pile ou face" avec une pièce équilibrée, comme événements A "face au premier lancer", B "face au second lancer", C "même issue aux deux lancers").

Indépendance mutuelle. On peut étendre la deuxième définition de l'indépendance au cas de n événements. On dira que les événements $A_1, A_2, \ldots, A_n$ sont mutuellement indépendants si pour toute partie I de l'ensembles des indices allant de 1 à n on a :

$$P\left(\bigcap_I A_i\right) = \prod_I P(A_i). \tag{B.4}$$

Il est important de noter que l'indépendance mutuelle est une condition plus forte que l'indépendance deux à deux.

Pour s'en convaincre il suffit de reconsidérer notre double lancer de pile ou face ci-dessus. Dans cet exemple on a en effet, A indépendant de B, B indépendant de C, et C indépendant de A, alors que C n'est pas indépendant $A \cap B$.

Autres formules utiles.

$$P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C) \tag{B.5}$$

$$P(A \cap B|C) = P(A|C)P(B|C \cap A) \tag{B.6}$$

Notations. Dans la suite nous utiliserons de façon interchangeable les notations suivantes pour désigner l'occurence simultanée de plusieurs événements :

- $A_1 \cap A_2 \cap \ldots \cap A_n$: la notation ensembliste (on insiste sur le fait que les A_i sont vus comme des ensembles).
- $A_1 \wedge A_2 \wedge \ldots \wedge A_n$: la notation logique (on insiste sur le fait que les A_i sont vus comme des formules logiques).
- $A_1, A_2, \ldots A_n$: notation indifférente (plus légère).

B.3.3 Formules de Bayes

Le formules de Bayes permettent d'exprimer $P(A|B)$ en fonction de $P(B|A)$.

Première formule de Bayes.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{B.7}$$

Théorème des probabilités totales. Si $B_1, B_2, \ldots, B_n$ est un système complet d'événements non triviaux alors le théorème des probabilités totales peut s'écrire sous la forme suivante

$$P(A) = \sum_i P(A|B_i)P(B_i) \tag{B.8}$$

Deuxième formule de Bayes. Dès lors la formule de Bayes peut s'écrire

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_k P(A|B_k)P(B_k)}, \tag{B.9}$$

qui s'appelle aussi théorème sur la "probabilité des causes", car il permet de calculer les probabilités des causes possibles d'un événement sachant qu'une conséquence s'est réalisée, connaissant la probabilité de cette dernière sous l'hypothèse de chaque cause et connaissant la probabilité des causes a priori.

Il ne faut pas confondre la (ou les) **formules** *Bayes avec la notion de* **règle** *de Bayes utilisée en théorie de la décision. La règle de Bayes est une régle de décision qui minimise une probabilité d'erreurs. Nous en verrons des exemples en Théorie de l'information et en Apprentissage inductif.*

Discussion. Le théorème de Bayes (on donne souvent le nom de théorème de Bayes aux deux formules de Bayes) joue un rôle très important dans le cadre du calcul des probabilités. Il sert de fondement au raisonnement incertain probabiliste et est à la base de toute une branche de la statistique appelée *statistique bayesienne*.

Il permet de remettre à jour les probabilités d'un certain nombre d'alternatives B_i en fonction d'informations nouvelles (le fait que A soit réalisé). On utilise souvent le terme de *probabilités a priori* pour désigner les $P(B_i)$ et le terme de *probabilités a posteriori* pour désigner les $P(B_i|A)$.

Par exemple, dans le cadre du diagnostic médical cette formule permet à un médecin de remettre à jour la plausibilité de certaines maladies (désignées par les B_i) à partir des symptômes observés (désignés conjointement par A), partant d'une connaissance de la probabilité a priori d'observer les différentes maladies (obtenues par exemple en effectuant des statistiques) et une connaissance des probabilités d'observer les symptômes A pour chacune de ces maladies (obtenues également par application de méthodes statistiques). Il s'agit d'une extension de la logique classique au raisonnement plausible.

B.4 ESPACE PRODUIT

Nous introduisons ci-dessous quelques notions et terminologies qui seront utilisées et illustrées plus loin, notamment dans le cadre des variables aléatoires.

B.4.1 Définition

Etant donnés un nombre fini d'espaces probabilisés $(\Omega_i, \mathcal{E}_i, P_i(\cdot))$ $(i = 1, \ldots, n)$ on peut définir un espace probabilisé produit $(\Omega, \mathcal{E}, P(\cdot))$ obtenu de la manière suivante :

- $\Omega = \Omega_1 \times \ldots \times \Omega_n$, (produit cartésien).
- $\mathcal{E} = \{A = \bigcup_j A_j \subset \Omega | \forall j, \forall i = 1, 2, \ldots, n, \exists A_{i,j} \in \mathcal{E}_i \ : A_j = A_{1,j} \times \ldots \times A_{n,j}\}$,
  (extension des événements par produit cartesien et union dénombrable, où on peut exiger sans perte de généralité que les A_j soient disjoints)
- $P(A) = \sum_j P(A_j)$ et $P(A_j) = \prod_i P_i(A_{i,j})$ (factorisation de la loi de probabilités).

On peut se convaincre que cette définition conduit bien à un espace probabilisé. On dira que les Ω_i sont les axes "orthogonaux" de l'espace produit et on parlera de la projection d'un événement A sur les axes pour désigner les A_i, et d'événements parallèles à un axe i si $A_i = \Omega_i$. On peut alors montrer que si deux événements de l'espace produit sont parallèles à des ensembles d'axes complémentaires, ils sont indépendants. En d'autres termes, si un évenement ne spécifie rien selon un certain nombres d'axes, alors le fait de savoir qu'un événement soit réalisé qui ne spécifie que de l'information relative à ces axes ne fournit aucune information sur cet événement.

B.4.2 Séries d'épreuves identiques et indépendantes

Un cas particulièrement intéressant en pratique d'espace produit est celui où tous les $(\Omega_i, \mathcal{E}_i, P_i(\cdot))$ sont identiques. Un tel type d'espace produit permet de modéliser les séries d'épreuves identiques et indépendantes, rencontrées en théorie de l'échantillonnage et à la base des statistiques.

B.4.3 Factorisation

Dans certains cas il est possible de factoriser un espace de départ en effectuant l'opération inverse, c'est-à-dire de l'écrire sous la forme du produit cartésien d'espaces indépendants (non nécessairement identiques).

(Suggestion : partir d'un espace fini dont on suppose que l'algèbre est engendré par deux événements indépendants A et B, et montrer qu'il peut se factoriser en deux axes correspondant à ces événements.)

B.4.4 Marginalisation

Partant d'un espace produit, il est possible de reconstituer les espaces produits correspondant à certains de ces axes par une opération de projection. En calcul de probabilité on dit qu'on "marginalise" les autres axes. Notons que cette opération peut s'effectuer sur un espace produit dont la loi de probabilité n'est pas factorisable.

B.5 VARIABLES ALEATOIRES

B.5.1 Définition générale

Soient un espace probabilisé $(\Omega, \mathcal{E}, P(\cdot))$ de départ et un univers Ω' de destination muni d'un algèbre d'événements $\mathcal{E}'$. Une fonction $f(\cdot)$ de Ω dans Ω' est une variable aléatoire si elle possède la propriété suivante :

$$\forall A' \in \mathcal{E}' \ : f^{-1}(A') \in \mathcal{E}, \tag{B.10}$$

où $f^{-1}(A')$ désigne $\{w \in \Omega | f(w) \in A'\}$. On dit alors que la fonction $f(\cdot)$ est $(\mathcal{E}, \mathcal{E}')$-mesurable, ou simplement mesurable si aucune confusion sur les σ-algèbres n'est possible.

Si cette propriété est vérifiée alors la variable aléatoire induit une loi de probabilité sur $(\Omega', \mathcal{E}')$ définie de la manière suivante :

$$P_f(A') = P\left(f^{-1}(A')\right). \tag{B.11}$$

Une variable aléatoire est donc une fonction définie sur un espace probabilisé qui est compatible avec les algèbres d'événements définis dans son espace d'origine et de destination, et qui induit donc une loi de probabilité sur l'espace de destination à partir de la loi de probabilité définie sur l'espace de départ. *(Suggestion : montrer que cette loi vérifie bien les axiomes de Kolmogorov.)*

Notons d'emblée que si les deux univers sont finis et munis des algèbres maximaux (et même si seulement l'espace de départ vérifie cette propriété), alors toute fonction de Ω dans Ω' est une variable aléatoire.

Par conséquent, si nous avons pris la précaution de formuler les restrictions ci-dessus, c'est que nous voulons appliquer le concept de variable aléatoire dans des situations où l'espace de départ est infini (p.ex. $\Omega = \mathbb{R}^p$).

Interprétation. Il est à noter que toute fonction de Ω dans Ω' induit à partir de $\mathcal{E}'$ un nouvel algèbre d'événements sur l'espace de départ Ω. *(Nous suggérons au lecteur de s'en convaincre en montrant que l'ensemble des parties de Ω qui peuvent s'écrire sous la forme $f^{-1}(A')$ avec $A' \in \mathcal{E}'$ est un σ-algèbre si $\mathcal{E}'$ est un σ-algèbre.)*

Le sens profond de la condition (B.10) est donc que l'algèbre des événements induit par la fonction $f(\cdot)$ à partir de $\mathcal{E}'$ sur Ω doit être inclus dans l'algèbre sur lequel la loi de probabilité $P(\cdot)$ est connue. Donc, une variable aléatoire a pour effet de remplacer l'algèbre $\mathcal{E}$ par l'algèbre $f^{-1}(\mathcal{E}')$ qui contient en général moins de sous-ensembles que $\mathcal{E}$. Elle opère donc sur un espace probabilisé en condensant l'information dont on disposait au départ en une information plus grossière.

Il s'agit bien là du sens physique profond de la notion de variable aléatoire : l'observation d'une variable aléatoire fournit une information généralement partielle sur les événements de l'univers de départ.

En résumé, une variable aléatoire transpose une loi de probabilité d'un espace de départ vers un espace de destination et elle transpose une structure d'algèbre de l'espace de destination vers l'espace de départ.

B.5.2 Fonction d'une variable aléatoire

Soit une variable aléatoire X à valeurs dans Ω', et une certaine fonction $\phi(\cdot)$ définie sur Ω' et à valeurs dans Ω''. Alors, si $\phi(\cdot)$ a le statut de variable aléatoire sur Ω' (compatibilité de $\mathcal{E}'$ et $\mathcal{E}''$), la fonction composée $\phi \circ X(\cdot) = \phi(X(\cdot))$ définit également une v.a. sur Ω.

Nous verrons ci-dessous le cas particulièrement intéressant en pratique où les deux fonctions sont des v.a. réelles.

B.5.3 Variable aléatoire discrète

Une variable aléatoire est discrète si l'ensemble de ses valeurs possibles est fini. Une telle variable aléatoire définit un système complet d'événements mutuellement exclusifs; en fait elle est équivalente à la donnée d'un système complet d'événements discrets. Aussi nous ne distinguerons plus dans la suite ces deux notions. Notons que si l'espace Ω est fini, alors toute variable aléatoire est nécessairement discrète.

On utilisera communément la notation $\mathcal{X} = \{X_1, \ldots, X_k\}, \mathcal{Y} = \{Y_1, \ldots, Y_l\}, \ldots$ pour désigner l'ensemble des valeurs possibles de telles variables aléatoires, et par extension la variable aléatoire elle-même sera désignée par $\mathcal{X}(\cdot)$ ou simplement $\mathcal{X}$.

D'autre part, nous assimilerons dans nos notations les valeurs prises par la variable aléatoire avec les événements (sous-ensembles de Ω) qui leurs correspondent. Par exemple la notation $P(X_i)$ désignera la probabilité de $\{\omega \in \Omega \, : \mathcal{X}(\omega) = X_i\}$.

Remarque. Dans la littérature on désigne souvent par variable aléatoire discrète une v.a. pouvant prendre un nombre dénombrable (éventuellement infini) de valeurs. Comme nous ne nous intéresserons que très marginalement au cas particulier infini, nous sous-entendrons (sauf mention explicite du contraire) que la variable discrète est aussi finie.

B.5.4 Variables aléatoires réelles

Une v.a. est dite réelle si $\Omega' = \mathbb{R}$. Elle peut être discrète ou non. Evidemment si Ω est fini, elle sera nécessairement discrète.

B.5.4.1 Fonction de répartition. Par définition, la fonction de répartition $F_{\mathcal{X}}$ d'une v.a. réelle $\mathcal{X}$ est une fonction de $\mathbb{R}$ dans $[0, 1]$ définie par

$$F(x) = P(]-\infty, x[), \tag{B.12}$$

Elle peut donc en principe avoir des discontinuités (à droite), si certaines des valeurs sont de probabilité non-nulle. Elle est monotonément croissante, et $F(-\infty) = 0$ et $F(+\infty) = 1$.

Cette fonction *caractérise* la variable aléatoire et permet de calculer la probabilité de tout intervalle de $\mathbb{R}$ par

$$P(a \leq \mathcal{X} < b) = F(b) - F(a). \tag{B.13}$$

Remarque. On peut tout aussi bien définir la fonction de répartition de façon à ce qu'elle soit continue à droite, on a alors

$$F'(x) = P(]-\infty, x]). \tag{B.14}$$

C'est la convention qu'on trouve généralement dans la littérature anglo-saxonne.

Lorsque la v.a. réelle est aussi discrète, il n'y a qu'un nombre fini de points de $\mathbb{R}$ de probabilité non-nulle. La fonction de répartition prend alors l'allure indiquée à la figure B.1.

B.5.4.2 Variable aléatoire (réelle) continue. Une variable aléatoire est dite continue si elle admet une densité, c'est-à-dire s'il existe une fonction $f(\cdot)$ définie sur $\mathbb{R}$ telle que $\forall a \leq b$ on a

$$P(]a, b[) = P([a, b]) = P([a, b[) = P(]a, b]) = \int_a^b f(x)dx. \tag{B.15}$$

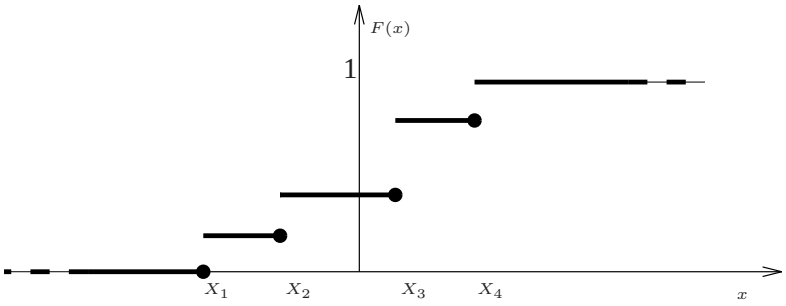

Figure B.1. Fonction de répartition d'un v.a. réelle discrète

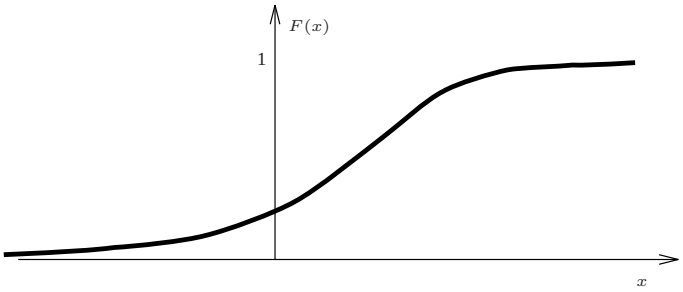

Figure B.2. Fonction de répartition d'un v.a. réelle continue

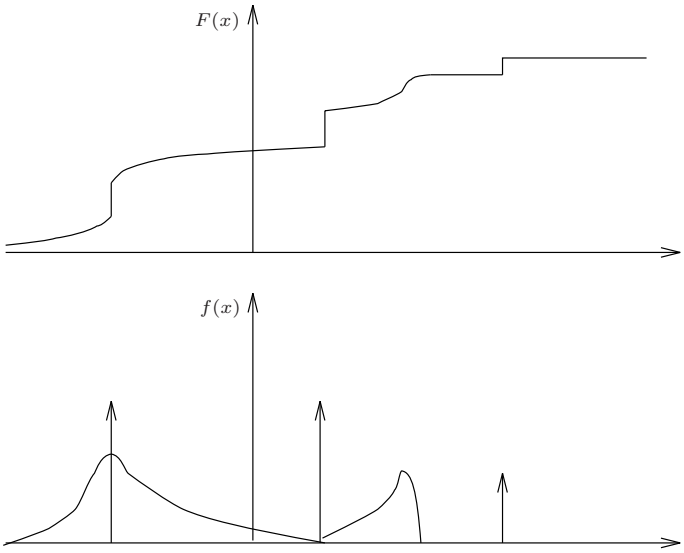

Figure B.3. Fonction de réparition et distribution de probabilité

Dans ce cas, $F(\cdot)$ est dérivable (et donc continue) et admet $f(\cdot)$ comme dérivée. Par conséquent, $f(\cdot)$ est positive et d'intégrale sur $\mathbb{R}$ égale à 1. La figure B.2 représente graphiquement la fonction de répartition de ce type de variable aléatoire.

Il est évident qu'une variable aléatoire peut être ni discrète, ni continue. Elle ne peut cependant qu'avoir au plus un nombre dénombrable de points de discontinuité.

Dans la suite nous utiliserons la notation $\mathcal{X} \sim f(x)$ pour indiquer qu'une v.a. suit une certaine distribution.

B.5.4.3 Cas général. Dans le cas général on peut séparer la v.a. réelle en la somme d'une composante continue et d'une composante discrète[6].

Notons qu'on peut dans le cas général aussi faire appel à la théorie des distributions pour définir cette fois la densité comme une *distribution* qui s'écrit sous la forme d'une combinaison linéraire d'un fonction et d'un série d'impulsions de Dirac. Cette situation est schématisée graphiquement à la figure B.3.

B.5.5 Indépendance de deux variables aléatoires

Deux variables aléatoires $\mathcal{X}$ et $\mathcal{Y}$ définies sur un même espace probabilisé sont indépendantes si et seulement si, $\forall A \in \mathcal{E}_X, \forall B \in \mathcal{E}_Y$ on a

$$P(\mathcal{X}^{-1}(A) \cap \mathcal{Y}^{-1}(B)) = P(\mathcal{X}^{-1}(A))P(\mathcal{Y}^{-1}(B)). \tag{B.16}$$

En d'autres termes, la loi induite par la v.a. $\mathcal{XY}(\cdot) = (\mathcal{X}(\cdot), \mathcal{Y}(\cdot))$ sur l'espace produit $\Omega_{\mathcal{X}} \times \Omega_{\mathcal{Y}}$, est factorisable, et on a

$$P_{\mathcal{XY}} = P_{\mathcal{X}} P_{\mathcal{Y}}. \tag{B.17}$$

Dans le cas où les variables aléatoires sont réelles cette condition se traduit par

$$H(x, y) \stackrel{\triangle}{=} P(X < x \wedge Y < y) = F(x)G(y), \tag{B.18}$$

où $F(\cdot)$ et $G(\cdot)$ sont les fonctions de répartition respectivement de $\mathcal{X}$ et $\mathcal{Y}$. Si de plus X et Y admettent les densités $f(\cdot)$ et $g(\cdot)$, alors il en est de même pour le couple, dont la densité est le produit :

$$h(x, y) = f(x)g(y).$$

Notons que nous pouvons étendre ces notions et propriétés par induction au cas d'un nombre fini quelconque de v.a. On parle alors de vecteurs aléatoires et le cas particulier intéressant est celui où celui-ci appartient à $\mathbb{R}^p$.

B.5.6 Espérance mathématique

Pour une variable réelle aléatoire discrète on définit l'espérance mathématique (on dit aussi sa moyenne) par

$$E\{\mathcal{X}\} = \sum_{i=1}^{k} X_k P(X_k). \tag{B.19}$$

Notons que dans le cas fini cette grandeur existe toujours. Dans le cas infini (dénombrable) il est facile de construire des exemples tels que cette série ne converge pas.

Pour une variable continue on a

$$E\{\mathcal{X}\} = \int_{\mathbb{R}} x f(x) dx. \tag{B.20}$$

Dans le cas plus général on peut combiner les deux formules. L'écriture générale est alors la suivante

$$E_P\{\mathcal{X}\} = \int_{\Omega} X(\omega) dP(\omega), \tag{B.21}$$

où le dP indique que l'intégrale est prise par rapport à la mesure P définie sur l'espace de départ Ω, ce qui est équivalent à

$$E\{\mathcal{X}\} = \int_{\mathbb{R}} x dP_{\mathcal{X}}(x), \tag{B.22}$$

où le $dP_{\mathcal{X}}$ indique que l'intégrale est prise par rapport à la mesure induite sur l'espace d'arrivée.

Il faut souligner, même si c'est évident, que l'intégrale n'est pas toujours définie. Il existe des distributions de probabilités continues pour lesquelles l'espérance mathématique n'est pas définie.

Par exemple, la distribution de *Cauchy*

$$f(x) = \frac{1}{\pi(1 + x^2)}, \tag{B.23}$$

n'admet pas d'espérance.

Cependant, toute fonction continue et à support compact étant intégrable, toute variable aléatoire réelle continue et bornée admet une espérance mathématique.

L'espérance mathématique d'une fonction ($\phi(\cdot) : \mathbb{R} \to \mathbb{R}$) est définie par

$$E_P\{\phi(\mathcal{X})\} = \int_\Omega \phi(X(\omega))dP(\omega), \tag{B.24}$$

et dans le cas continu on a

$$E_P\{\phi(\mathcal{X})\} = \int_{\mathbb{R}} \phi(x)f(x)dx. \tag{B.25}$$

Cas particuliers :

1. Fonction constante ($\phi(x) = a$) : $E\{\phi\} = a$.
2. Fonction linéaire ($\phi(x) = ax + b$) : $E\{\phi\} = aE\{\mathcal{X}\} + b$.
3. Somme de deux v.a. ($\phi(x, y) = x + y$) : $E\{\phi\} = E\{\mathcal{X}\} + E\{\mathcal{Y}\}$.
4. Produit de deux v.a. ($\phi(x, y) = xy$) :

$$E\{\mathcal{X}\mathcal{Y}\} = \int_{\mathbb{R}^2} xydP_{\mathcal{X}\mathcal{Y}}(x, y).$$

Lorsque $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes, la mesure $dP_{\mathcal{X}\mathcal{Y}}(x, y)$ se factorise et l'intégrale double peut se décomposer en produit des deux intégrales simples :

$$E\{\mathcal{X}\mathcal{Y}\} = \int_{\mathbb{R}} xdP_{\mathcal{X}}(x) \int_{\mathbb{R}} ydP_{\mathcal{Y}}(y) = E\{\mathcal{X}\}E\{\mathcal{Y}\}.$$

La réciproque n'est pas vraie.

B.5.7 Variance et écart-type

Lorsque l'espérance existe, la **variance** est définie par

$$V\{\mathcal{X}\} = \sigma^2 = E\{(\mathcal{X} - m)^2\}, \tag{B.26}$$

lorsque cette grandeur existe, où $m = E\{\mathcal{X}\}$.

L'écart-type est la racine carrée σ de la variance.

Propriétés de la variance. On a

$$E\{(\mathcal{X} - a)^2\} = V\{\mathcal{X}\} + (E\{\mathcal{X}\} - a)^2, \tag{B.27}$$

et par conséquent, la variance est la valeur minimale de $E\{(\mathcal{X} - a)^2\}$, et $a = E\{\mathcal{X}\}$ minimise $E\{(\mathcal{X} - a)^2\}$. Cette propriété est exploitée très largement en statistiques, dans le domaine de l'estimation au sens des moindres carrés.

On en déduit, en prenant $a = 0$ que

$$V\{\mathcal{X}\} = E\{X^2\} - (E\{\mathcal{X}\})^2. \tag{B.28}$$

L'espérance et l'écart-type sont reliés par *l'inégalité de Bienaymé-Tchebyshev* :

$$P(|X - E\{\mathcal{X}\}| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}, \tag{B.29}$$

dont on déduit que si $\sigma = 0$ la v.a. est presque sûrement égale à sa moyenne, c'est-à-dire constante. La variance mesure donc bien le caractère aléatoire d'une v.a.

Par ailleurs, on a

- $V\{\mathcal{X}+a\} = V\{\mathcal{X}\}$.
- $V\{a\mathcal{X}\} = a^2 V\{\mathcal{X}\}$.
- $V\{\mathcal{X}+\mathcal{Y}\} = V\{\mathcal{X}\} + V\{\mathcal{Y}\} + 2\text{cov}\{\mathcal{X},\mathcal{Y}\}$, où

$$\text{cov}\{\mathcal{X},\mathcal{Y}\} \stackrel{\triangle}{=} E\left\{(\mathcal{X}-E\{\mathcal{X}\})(\mathcal{Y}-E\{\mathcal{Y}\})\right\} = E\{\mathcal{X}\mathcal{Y}\} - E\{\mathcal{X}\}E\{\mathcal{Y}\}$$

désigne la covariance.
Si les v.a. sont indépendantes, on a $\text{cov}\{\mathcal{X},\mathcal{Y}\} = 0$ et donc

$$V\{\mathcal{X}+\mathcal{Y}\} = V\{\mathcal{X}\} + V\{\mathcal{Y}\}.$$

La réciproque n'est pas vraie.

B.5.7.1 Inégalité de Jensen. *Si $\phi(\cdot)$ est une fonction convexe (voir appendice D, pour la définition de la convexité), et si $\mathcal{X}$ est une v.a. réelle, alors*

$$E_P\{\phi(\mathcal{X})\} \geq \phi(E_P\{\mathcal{X}\}), \tag{B.30}$$

et si la fonction est strictement convexe, alors l'égalité implique $X = cnste$, sauf éventuellement sur un ensemble de probabilité nulle.

Cette inégalité est très pratique dans le contexte d'un certain nombre de démonstrations en théorie de l'information et dans le domaine des processus aléatoires.

B.6 VARIABLES ALEATOIRES COMPLEXES

Tout ce qui vient d'être dit concernant les variables aléatoires réelles peut, à peu de choses près, être appliqué aux variables aléatoires à valeurs complexes. D'ailleurs, on peut séparer toute fonction complexe en ses parties réelle et immaginaire qui sont des fonctions réelles. Une variable aléatoire complexe est donc de ce point de vue équivalente à un couple de variables aléatoires réelles.

B.7 COUPLES DE V.A. ET CONDITIONNEMENT

B.7.1 Cas discret

Nous étudions ici les couples de v.a. $(\mathcal{X},\mathcal{Y})$ tels que $\mathcal{X}$ et $\mathcal{Y}$ prennent leurs valeurs dans un ensemble fini désigné respectivement par $\mathcal{X} = \{X_1,\ldots,X_k\}$ et $\mathcal{Y} = \{Y_1,\ldots,Y_l\}$ munis de leur $\sigma-$algèbre maximal. Dans ce cas, le couple prend ses valeurs dans $\mathcal{X}\times\mathcal{Y}$ muni du $\sigma-$algèbre produit, qui est également maximal. On suppose que la fonction ainsi induite de $\Omega \to \mathcal{X}\times\mathcal{Y}$ est bien $\mathcal{E}$-mesurable.

B.7.1.1 Lois associées.

Loi (con)jointe. La loi de probabilité du couple $P_{\mathcal{X},\mathcal{Y}}$ est déterminée complètement par la conaissance des kl nombres

$$p_{i,j} \stackrel{\triangle}{=} P(\mathcal{X} = X_i \cap \mathcal{Y} = Y_j), \forall i = 1,\ldots,k, \forall j = 1,\ldots,l. \tag{B.31}$$

On a bien sur $\sum_{i=1}^{k}\sum_{j=1}^{l} p_{i,j} = 1$.

Lois marginales. La loi marginale de $\mathcal{X}$ est évidemment

$$p_{i,\cdot} \stackrel{\triangle}{=} P(\mathcal{X} = X_i) = \sum_{j=1}^{l} p_{i,j}. \tag{B.32}$$

	Y_1	$\cdots$	Y_j	$\cdots$	Y_l	
X_1			$\vdots$			
$\vdots$			$\vdots$			
X_i	$\cdots$	$\cdots$	$p_{i,j}$	$\cdots$	$\cdots$	$p_{i,\cdot}$
$\vdots$			$\vdots$			
X_k			$\vdots$			
			$p_{\cdot,j}$			

Figure B.4. Table de contingence

De même, la loi marginale de $\mathcal{Y}$ est

$$p_{\cdot,j} \triangleq P(\mathcal{Y} = X_j) = \sum_{i=1}^{k} p_{i,j}. \tag{B.33}$$

On représente souvent un couple de v.a. à l'aide d'une *table de contingences*, telle qu'illustrée à la Figure B.4

Lois conditionnelles. La loi conditionnelle de $\mathcal{X}$ connaissant $\mathcal{Y}$ est définie par

$$p_{X_i|Y_j} \triangleq P(\mathcal{X} = X_i|\mathcal{Y} = Y_j) = \frac{p_{i,j}}{p_{\cdot,j}}, \tag{B.34}$$

et celle de $\mathcal{Y}$ connaissant $\mathcal{X}$ par

$$p_{Y_j|X_i} \triangleq P(\mathcal{Y} = Y_j|\mathcal{X} = X_i) = \frac{p_{i,j}}{p_{i,\cdot}}. \tag{B.35}$$

B.7.1.2 Moments conditionnels.

Supposons que $\mathcal{Y}$ soit une v.a. réelle (où éventuellement complexe).

Espérance conditionelle. Alors on définit l'espérance conditionnelle de $\mathcal{Y}$ par

$$E\{\mathcal{Y}|X\} \triangleq \sum_{j=1}^{l} Y_j p_{Y_j|X}. \tag{B.36}$$

$E\{\mathcal{Y}|X\}$ est donc une fonction (réelle où complexe) de X. Cette fonction s'appelle **fonction de régression** de $\mathcal{Y}$ en X. Comme $\mathcal{X}$ est une v.a. cette fonction définit une v.a. reélle ou complexe. Cette variable aléatoire présente un certain nombre de propriétés remarquables que nous allons énumerer.

En premier lieu elle est linéaire (il s'agit d'une espérance). Donc,

$$E\{\mathcal{Y}_1 + \mathcal{Y}_2|\mathcal{X}\} = E\{\mathcal{Y}_1|\mathcal{X}\} + E\{\mathcal{Y}_2|\mathcal{X}\}. \tag{B.37}$$

Mais surtout, elle satisfait au **théorème de l'espérance totale** :

$$E\{E\{\mathcal{Y}|\mathcal{X}\}\} = E\{\mathcal{Y}\}. \tag{B.38}$$

En effet, on peut calculer son espérance mathématique ce qui donne

$$E\{E\{\mathcal{Y}|\mathcal{X}\}\} = \sum_{i=1}^{k} p_{i,\cdot} \sum_{j=1}^{l} Y_j p_{Y_j|X_i} \tag{B.39}$$

$$= \sum_{j=1}^{l} Y_j \sum_{i=1}^{k} p_{i,\cdot} p_{Y_j|X_j} \tag{B.40}$$

$$= \sum_{j=1}^{l} Y_j p_{\cdot,j}. \tag{B.41}$$

Variance conditionnelle. On définit, similairement la variance conditionnelle comme une v.a. qui prend la valeur

$$V\{\mathcal{Y}|\mathcal{X}\} \stackrel{\triangle}{=} E\left\{\left(\mathcal{Y} - E\{\mathcal{Y}|\mathcal{X}\}\right)^2 |\mathcal{X}\right\}. \tag{B.42}$$

On a le **théorème de la variance totale** qui s'écrit comme suit :

$$V\{\mathcal{Y}\} = E\left\{V\{\mathcal{Y}|\mathcal{X}\}\right\} + V\left\{E\{\mathcal{Y}|\mathcal{X}\}\right\}. \tag{B.43}$$

B.7.2 Variables continues

B.7.2.1 Une des deux variables est continue. On peut directement étendre ce qui précède au cas où $\mathcal{Y}$ est une variable continue en remplaçant les probabilités par les fonctions de répartition où des densités. On note

$$G(y|x) \stackrel{\triangle}{=} P(\mathcal{Y} < y|\mathcal{X} = x), \tag{B.44}$$

puis si elle existe, la densité marginale est la dérivée de $G(y)$. La fonction de répartition marginale s'écrit

$$G(y) \stackrel{\triangle}{=} \sum_{i=1}^{k} p_{i,\cdot} G(y|X_i) \tag{B.45}$$

qui dérivée terme à terme donne la densité marginale

$$g(y) \stackrel{\triangle}{=} \sum_{i=1}^{k} p_{i,\cdot} g(y|X_i). \tag{B.46}$$

Les théorèmes de l'espérance et de la variance totales restent également d'application.

On peut également écrire

$$P(\mathcal{X} = x|\mathcal{Y} < y) = \frac{G(y|x)P(x)}{G(y)}, \tag{B.47}$$

mais nous ne pouvons pas pour le moment écrire

$$P(\mathcal{X} = x|\mathcal{Y} = y) = \frac{g(y|x)P(x)}{g(y)}, \tag{B.48}$$

car $Y = y$ est un événement de probabilité nulle par rapport auquel on ne peut pas en principe conditionner. Nous allons indiquer ci-dessous sous quelles conditions un conditionnement de ce type est permis.

Illustration. Un exemple pratique important où on considère les dépendances entre variables continues et discrète est fourni par la théorie de la décision, qui intervient dans les problèmes de classification en apprentissage automatique, et également dans les problèmes de transmission de données numériques à l'aide de signaux analogiques.

Prenons par exemple, le problème de l'allocation de crédit bancaire qui se ramène à celui de l'étude des relations entre variables numériques (montant du crédit souhaîté, niveau de salaire, endettement, âge ...) et discrètes décrivant la situation financière et sociale d'un demandeur de crédit (état civil, propriétaire, statut professionnel...), et la décision optimale d'une banque (Accord ou non du crédit).

Du point de vue du banquier non altruiste, la décision optimale est en l'occurence celle qui maximise l'espérance mathématique du bénéfice de la banque. Si le crédit est accordé, ce bénéfice dépendra du fait que le demandeur sera capable de rembourser les mensualités ou non. Si le crédit n'est pas accordé, le bénéfice est nul. Le banquier fera donc appel à un logiciel qui déterminera, sur base des informations fournies par le demandeur, la probabilité de remboursement complet du crédit (disons $P(\mathcal{R} = V|\mathcal{I})$, où $\mathcal{R}$ désigne une variable qui vaut V s'il y a remboursement et F sinon, et $\mathcal{I}$ symbolise les informations propres au demandeur), à partir de laquelle on pourra déterminer l'espérance mathématique du bénéfice par la formule de l'espérance totale (conditionnée par l'information fournie par le demandeur)

$$E\{\mathcal{B}|\mathcal{I}\} = E\{\mathcal{B}|\mathcal{R} = V, \mathcal{I}\}P(\mathcal{R} = V|\mathcal{I}) + E\{\mathcal{B}|\mathcal{R} = F, \mathcal{I}\}P(\mathcal{R} = F|\mathcal{I}). \tag{B.49}$$

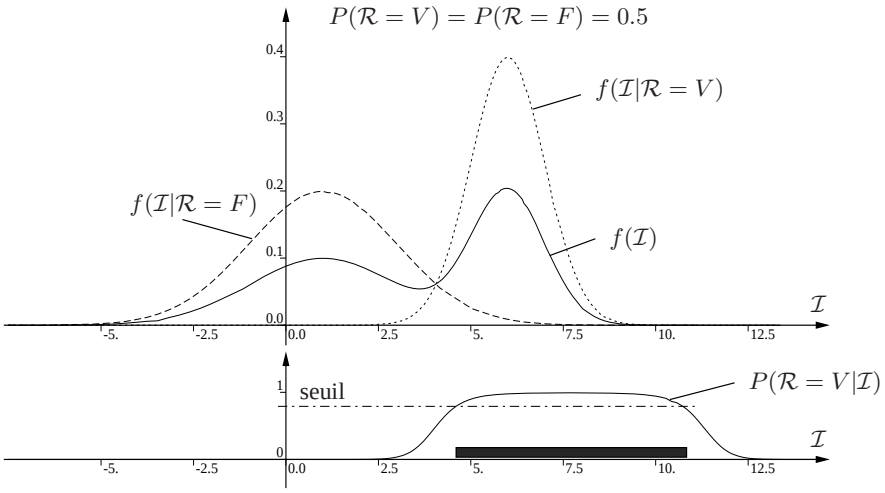

Figure B.5. Illustration des densités conditionnelles

Dans cette formule, on a évidemment

$$P(\mathcal{R}=F|\mathcal{I})=1-P(\mathcal{R}=V|\mathcal{I}),$$

et le chiffre $E\{\mathcal{B}|\mathcal{R}=V,\mathcal{I}\}$ correspond au gain de la banque calculé au moyen de formules d'actualisation tenant compte des conditions du crédit (intérêt, type de remboursement, ...), du coût de l'argent immobilisé que la banque doit assumer, et est évidemment proportionnel au montant du crédit. D'autre part, le terme $E\{\mathcal{B}|\mathcal{R}=F,\mathcal{I}\}$ est quant à lui un "bénéfice" negatif.

Par conséquent, le crédit sera alloué si

$$E\{\mathcal{B}|\mathcal{I}\}>0 \Leftrightarrow P(\mathcal{R}=V|\mathcal{I})>\frac{-E\{\mathcal{B}|\mathcal{R}=F,\mathcal{I}\}}{E\{\mathcal{B}|\mathcal{R}=V,\mathcal{I}\}-E\{\mathcal{B}|\mathcal{R}=F,\mathcal{I}\}}, \tag{B.50}$$

et le problème se ramène donc essentiellement au calcul de $P(\mathcal{R}=V|\mathcal{I})$ et à la comparaison de celle-ci à un certain seuil, $\mathcal{R}$ étant une variable discrète et $\mathcal{I}$ une ensembles de variables généralement mixtes discrètes/continues. Nous verrons au cours d'apprentissage automatique que les méthodes utilisées par les banquiers se fondent essentiellement sur une approximation de $P(\mathcal{R}=V|\mathcal{I})$ obtenue à partir de bases de données des clients antérieurs de la banque et grâce aux méthodes d'apprentissage.

Notons que nous avons utilisé la notation explicite $\mathcal{R}=V$ ou $\mathcal{R}=F$ pour bien mettre en évidence le conditionnement sur des valeurs prises par la v.a. discrète $\mathcal{R}$. Selon notre convention, la notation $f(\mathcal{I}|\mathcal{R})$ désigne en effet une fonction à deux arguments définie par

$$f(\mathcal{I}|\mathcal{R}) \quad = \quad \begin{cases} f(\mathcal{I}|\mathcal{R}=V) \text{ si } \mathcal{R}=V \\ f(\mathcal{I}|\mathcal{R}=F) \text{ si } \mathcal{R}=F \end{cases}, \tag{B.51}$$

et $f(\mathcal{I},\mathcal{R})$ est définie par

$$f(\mathcal{I}|\mathcal{R})P(\mathcal{R}) \tag{B.52}$$

où $\mathcal{R}$ peut désigner soit la valeur V soit la valeur F.

Cette remarque étant faite, illustrons ces idées graphiquement pour un cas simple où $\mathcal{I}$ se réduit à une seule variable numérique (disons un chiffre magique obtenu en combinant les différentes informations selon une formule pré-établie) et faisons l'hypothèse que cette variable est continue. La figure B.5 représente graphiquement la situation, en terme des densités de probabilité $f(\mathcal{I})$, $f(\mathcal{I}|\mathcal{R}=V)$, $f(\mathcal{I}|\mathcal{R}=F)$ et la probabilité conditionnelle $P(\mathcal{R}=V|\mathcal{I})$.

Notons qu'à la figure B.5 les distributions conditionnelles $f(\mathcal{I}|\mathcal{R}=V)$ et $f(\mathcal{I}|\mathcal{R}=F)$ sont gaussiennes, de moyennes et de variances différentes. On suppose donc que les bons et les mauvais clients présentent des valeurs assez différentes de notre variable "magique". On suppose également qu'a priori dans la population qui s'adresse

aux banques pour obtenir des crédit on a la même proportion de bons et de mauvais clients, ce qui se traduit par l'égalité des probabilités a priori $P(\mathcal{R} = V)$ et $P(\mathcal{R} = F)$. On a,

$$f(\mathcal{I}) = f(\mathcal{I}|\mathcal{R} = V)P(\mathcal{R} = V) + f(\mathcal{I}|\mathcal{R} = F)P(\mathcal{R} = F), \tag{B.53}$$

et la probabilité a posteriori $P(\mathcal{R} = V|\mathcal{I})$ représentée sur la partie inférieure de la figure B.5 est obtenue par la formule de Bayes

$$P(\mathcal{R} = V|\mathcal{I}) = \frac{f(\mathcal{I}|\mathcal{R} = V)P(\mathcal{R} = V)}{f(\mathcal{I})}, \tag{B.54}$$

et voit qu'elle vaut 0.5 au point de croisement des trois courbes du haut, c'est-à-dire au point où,

$$f(\mathcal{I}) = f(\mathcal{I}|\mathcal{R} = V) = f(\mathcal{I}|\mathcal{R} = F), \tag{B.55}$$

parce que les classes sont a priori équiprobables.

Sur la partie inférieure de la figure on a illustré la règle de décision du banquier par un seuil supposé indépendant de $\mathcal{I}$ (ce qui n'est pas nécessairement vrai en pratique comme il ressort des formules générales indiqués ci-dessus). L'ensemble des valeurs de $\mathcal{I}$ pour lesquelles le crédit est alloué est l'intervalle indiqué sur la figure.

(Suggestion : trouver l'expression générale en fonction de $P(\mathcal{R} = V)$, de la relation entre $f(\mathcal{I}|\mathcal{R} = V)$ et $f(\mathcal{I}|\mathcal{R} = F)$ au point où $P(\mathcal{R} = V|\mathcal{I}) = seuil$.)

B.7.2.2 Cas le plus général. Nous reférons le lecteur intéressé par les conditions d'existence de mesures de probabilités conditionnelles vis-à-vis d'événements de probabilité nulle à [Bil79] et à [Sap90] pour une discussion des implications en terme de conditionnement vis-à-vis de v.a. quelconques.

On peut résumer la situation de la manière suivante : si $\mathcal{Y}$ est une variable aléatoire réelle, et si $\mathcal{X}$ est une variable aléatoire soit discrète, soit à valeurs dans $\mathbb{R}^p$, alors il est permis de conditionner Ω et donc $\mathcal{Y}$ par rapport à $\mathcal{X}$ localement. De plus, si $E\{\mathcal{Y}\}$ existe alors il existe une v.a. aléatoire "espérance conditionnelle" qui satisfait au théorème de l'espérance totale. Enfin, si $V\{\mathcal{Y}\}$ existe aussi alors cette v.a. satisfait aussi au théorème de la variance totale.

Enfin, les formules de conditionnement des densités s'obtiennent par analogie au cas discret.

En particulier on a

$$g(y|x) = \frac{h(x,y)}{f(x)} \tag{B.56}$$

$$E\{\mathcal{Y}|x\} = \int yg(y|x)dy \tag{B.57}$$

et la formule de Bayes

$$g(y|x) = \frac{f(x|y)g(y)}{f(x)}. \tag{B.58}$$

B.8 LOIS DE PROBABILITE D'USAGE COURANT

A toutes fins utiles, nous rapelons ici quelques lois de probabilités usuellement rencontrées et nous énonçons, sans les démontrer, leurs propriétés principales.

B.8.1 Lois discrètes

B.8.1.1 Uniforme. Elle est définie sur $\{1, 2, \ldots, n\}$ et associe une probabilité de $\frac{1}{n}$ à chacune de ces n valeurs possibles.

On a $E\{\mathcal{X}\} = \frac{n+1}{2}$ et $V\{\mathcal{X}\} = \frac{n^2-1}{12}$.

B.8.1.2 Bernouilli. C'est une loi d'une v.a. $\mathcal{X}$ ne pouvant prendre que deux valeurs possibles 1 ou 0, avec les probabilités p et $1-p$. En d'autres termes, $\mathcal{X}$ est la fonction indicatrice d'un événement A de probabilité p.

On a $E\{\mathcal{X}\} = p$ et $V\{\mathcal{X}\} = p(1-p)$.

B.8.1.3 Binomiale. Supposons qu'on répète n expériences de Bernouilli, et qu'on compte le nombre de fois sur n que l'événement A est réalisé. Désignons par $\mathcal{X}$ la variable aléatoire qui désigne le compte. $\mathcal{X}$ est la somme de n v.a. indépendantes et identiquement distribuées (i.i.d.)

$$\mathcal{X} = \sum_{i=1}^{n} \mathcal{X}_i.$$

La loi de cette v.a. est par définition la loi binomiale $\mathcal{B}(n,p)$. Les valeurs possibles de $\mathcal{X}$ sont $\{0,1,\ldots,n\}$

On a $E\{\mathcal{X}\} = np$, et $V\{\mathcal{X}\} = np(1-p)$. D'autre part, on a

$$P(\mathcal{X}=k) = C_n^k p^k (1-p)^{(n-k)}.$$

On a la propriété importante (et évidente) suivante : soient $\mathcal{X} \sim \mathcal{B}(n_1,p)$ et $\mathcal{Y} \sim \mathcal{B}(n_2,p)$, alors

$$\mathcal{Z} = \mathcal{X} + \mathcal{Y} \sim \mathcal{B}(n_1+n_2,p).$$

La loi binomiale permet la modélisation des tirages sans remise.

B.8.1.4 Poisson. La loi de Poisson $\mathcal{P}(\lambda)$ et la loi d'une v.a. entière positive ou nulle qui satisfait à

$$P(\mathcal{X}=x) = \exp(-\lambda)\frac{\lambda^x}{x!}.$$

On a $E\{\mathcal{X}\} = \lambda$, et $V\{\mathcal{X}\} = \lambda$. On montre que si $\mathcal{X}_n \sim \mathcal{B}(n,p)$ est une suite de v.a. binomiales telle que

$$\lim_{n\to\infty} np = \lambda,$$

alors $\mathcal{X}_n$ converge en loi (voir ci-dessous) vers $\mathcal{P}(\lambda)$.

B.8.2 Lois continues

B.8.2.1 Uniforme. La loi uniforme sur $[0,a]$, notée $\mathcal{U}_{[0,a]}$ est définie par la densité uniforme $u_{[0,a]}(x) = \frac{1}{a}$ sur $[0,a]$, et 0 ailleurs.

On a $E\{\mathcal{X}\} = \frac{a}{2}$, et $V\{\mathcal{X}\} = \frac{a^2}{12}$.

La somme de deux v.a. uniformes identiques et indépendantes est une loi triangulaire sur $[0,2a]$.

B.8.2.2 Exponentielle. La densité de la loi exponentielle de paramètre λ est

$$f(x) = \lambda \exp(-\lambda x)$$

si $x > 0$, 0 ailleurs.

On a $E\{\mathcal{X}\} = \frac{1}{\lambda}$, et $V\{\mathcal{X}\} = \frac{1}{\lambda^2}$.

B.8.2.3 Gaussienne (ou normale). $\mathcal{X}$ suit une loi Gaussienne de moyenne μ et de variance σ^2, notée $\mathcal{G}(\mu,\sigma^2)$ ou $\mathcal{N}(\mu,\sigma^2)$, si sa densité est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

On a $E\{\mathcal{X}\} = \mu$, et $V\{\mathcal{X}\} = \sigma^2$. Si $\mu = 0$ on dit que la loi est centrée. Si $\sigma = 1$ on dit qu'elle est réduite.

Additivité. Si $\mathcal{X} \sim \mathcal{N}(\mu_1,\sigma_1^2)$ et $\mathcal{Y} \sim \mathcal{N}(\mu_2,\sigma_2^2)$ sont deux variables aléatoires **indépendantes**, alors leur somme suit encore une loi normale et on a

$$\mathcal{Z} = \mathcal{X} + \mathcal{Y} \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2).$$

La loi Gaussienne joue un rôle très important, notamment à cause du théorème central-limite qui permet d'affirmer que la loi est d'application dans de nombreuses situations pratiques. Nous allons en voir la généralisation aux vecteurs aléatoires de dimension p.

B.9 VECTEURS ALEATOIRES

Nous nous intéressons ici aux v.a. à valeurs dans l'espace euclidien $\mathbb{R}^p$. Nous rappelons d'abord quelques notations et propriétés générales de telles variables aléatoires, puis nous nous focaliserons sur les lois Gaussiennes multi-dimensionnelles.

Ci-dessous nous indiquerons en gras les vecteurs (colonnes) et matrices. Etant donné un vecteur ou une matrice $\boldsymbol{V}$ nous noterons par $\boldsymbol{V}^T$ le vecteur ou la matrice transposée. Etant donné une matrice $\boldsymbol{M}$ nous noterons par $|\boldsymbol{M}|$ son déterminant.

B.9.1 Généralités sur les v.a. vectorielles

Une v.a. vectorielle ou vecteur (colonne) aléatoire $\boldsymbol{\mathcal{X}}$ est une application mesurable de $(\Omega, \mathcal{E}, P)$ dans $\mathbb{R}^p$ muni de son σ-algèbre borélien (produit cartésien de p σ-algèbres boréliens sur $\mathbb{R}$).

La fonction de répartition d'un vecteur aléatoire est une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$ définie par

$$F(x_1, x_2, \ldots, x_p) \stackrel{\triangle}{=} P(\mathcal{X}_1 < x_1, \ldots, \mathcal{X}_p < x_p), \tag{B.59}$$

où $\mathcal{X}_i$ désigne la i-ème composante de $\boldsymbol{\mathcal{X}}$.

Si la densité existe, elle est définie par

$$f(x_1, x_2, \ldots, x_p) \stackrel{\triangle}{=} \frac{\partial^p F}{\partial x_1 \ldots \partial x_p}. \tag{B.60}$$

On note par $\boldsymbol{\mu}$ (ou $\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}$, si nécessaire; certains auteurs utlisent la notation $\overline{\boldsymbol{x}}$) le vecteur colonne

$$\boldsymbol{\mu} \stackrel{\triangle}{=} E\{\boldsymbol{\mathcal{X}}\} = \begin{bmatrix} E\{\mathcal{X}_1\} \\ \vdots \\ E\{\mathcal{X}_p\} \end{bmatrix}, \tag{B.61}$$

dont les composantes sont les espérances mathématiques des p composantes de $\boldsymbol{\mathcal{X}}$.

On note par $\boldsymbol{\Sigma}$ (ou $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{X}}}$, si nécessaire) la matrice $p \times p$ de variance-covariance définie par

$$\boldsymbol{\Sigma} \stackrel{\triangle}{=} E\{(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})^T\} = E\{\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{X}}^T\} - \boldsymbol{\mu}\boldsymbol{\mu}^T, \tag{B.62}$$

dont l'élément i, j est

$$\boldsymbol{\Sigma}_{i,j} = \text{cov}(\mathcal{X}_i, \mathcal{X}_j). \tag{B.63}$$

En particulier, on a $\boldsymbol{\Sigma}_{i,i} = V\{\mathcal{X}_i\}$. Notons que $\boldsymbol{\Sigma}$ est symétrique et semi-définie positive. Elle peut donc être diagonalisée au moyen d'une transformation orthogonale. Le résultat de cette transformation donne un vecteur aléatoire dont les composantes sont *décorrélées*, mais pas nécessairement indépendantes.

Transformations linéaires. Soit $\boldsymbol{A}$ une matrice $r \times p$ et $\boldsymbol{\mathcal{X}}$ un v.a. de $\mathbb{R}^p$. Alors $\boldsymbol{\mathcal{Y}} = \boldsymbol{A}\boldsymbol{\mathcal{X}}$ est un vecteur aléatoire de $\mathbb{R}^r$.

On a $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}} = \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}$ et $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Y}}} = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{X}}}\boldsymbol{A}^T$.

Théorème de Cramer-Wold. On montre que la loi de $\boldsymbol{\mathcal{X}}$ est entièrement déterminée par celles de toutes les combinaisons linéaires de ses composantes $\boldsymbol{a}^T\boldsymbol{\mathcal{X}}$, $\forall \boldsymbol{a} \in \mathbb{R}^p$.

B.9.2 Vecteurs aléatoires Gaussiens

Définition. $\boldsymbol{\mathcal{X}}$ est (par définition) un vecteur aléatoire Gaussien à p dimensions (on note $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, où $\boldsymbol{\Sigma}$ est la matrice de variance-covariance de $\boldsymbol{\mathcal{X}}$, et $\boldsymbol{\mu}$ sa moyenne), si toute combinaison linéaire de ses composantes $\boldsymbol{a}^T\boldsymbol{\mathcal{X}}$, $\forall \boldsymbol{a} \in \mathbb{R}^p$, suit une loi de Gauss à une dimension.

La propriété d'être Gaussien est donc invariante vis-à-vis de toute transformation linéaire (rotation, dilatation, translation,...) de l'espace $\mathbb{R}^p$. Cette propriété implique en particulier que toutes les composantes suivent des lois Gaussiennes (mais la réciproque est fausse).

Propriétés fondamentales. On a les propriétés fondamentales suivantes :

- En général, si $\boldsymbol{A}$ est une matrice $r \times p$ et $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ un v.a. Gaussien de $\mathbb{R}^p$, alors $\boldsymbol{\mathcal{Y}} = \boldsymbol{A}\boldsymbol{\mathcal{X}}$ est un v.a. Gaussien de $\mathbb{R}^r$, et on a $\boldsymbol{\mathcal{Y}} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.
(Suggestion : montrer que $\boldsymbol{\mathcal{Y}}$ est bien un v.a. Gaussien, en prouvant que toutes ses projections sont de v.a. réelles Gaussiennes.)
- Donc, si $\mathcal{Y} = \boldsymbol{a}^T\boldsymbol{\mathcal{X}}$ alors $E\{\mathcal{Y}\} = \boldsymbol{a}^T\boldsymbol{\mu}$, et $V\{\mathcal{Y}\} = \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}$, donc $\mathcal{Y} \sim \mathcal{N}_r(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})$.
- On déduit de la propriété précédente que les distributions marginales (des composantes de $\boldsymbol{\mathcal{X}}$) sont les suivantes : $\mathcal{X}_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$ conformément à l'intuition.
(Suggestion : appliquer la propriété précédente au vecteur $\boldsymbol{a}$ de composantes $a_j = \delta_{i,j}$.)
- Les composantes de $\boldsymbol{\mathcal{X}}$ sont mutuellement indépendantes si, et seulement si, $\boldsymbol{\Sigma}$ est une matrice diagonale, c'est-à-dire si les composantes sont décorrélées deux à deux.
(Suggestion : montrer que la condition est suffisante)
- Lorsque $\boldsymbol{\Sigma}$ est régulière, et seulement dans ce cas, la densité existe et vaut

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{\Sigma}|}} exp\left(\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right). \tag{B.64}$$

Distributions conditionnelles. Si on partitionne $\boldsymbol{\mathcal{X}}$ en deux sous-vecteurs $\boldsymbol{\mathcal{X}}_1$ et $\boldsymbol{\mathcal{X}}_2$ à k et $p-k$ composantes, respectivement de moyennes $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$:

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} \boldsymbol{\mathcal{X}}_1 \\ \boldsymbol{\mathcal{X}}_2 \end{bmatrix}, \tag{B.65}$$

la moyenne se partitionne selon

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \tag{B.66}$$

et la matrice de variance-covariance se partitionne

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} & \boldsymbol{\Sigma}_{1,2} \\ \boldsymbol{\Sigma}_{2,1} & \boldsymbol{\Sigma}_{2,2} \end{bmatrix} \tag{B.67}$$

La loi conditionnelle de $\boldsymbol{\mathcal{X}}_1$ lorsque $\boldsymbol{\mathcal{X}}_2$ est connu est alors une Gaussienne à k dimensions

- d'espérance

$$E\{\boldsymbol{\mathcal{X}}_1|\boldsymbol{\mathcal{X}}_2\} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{1,2}\boldsymbol{\Sigma}_{2,2}^{-1}(\boldsymbol{\mathcal{X}}_2 - \boldsymbol{\mu}_2). \tag{B.68}$$

- de matrice de variance-covariance

$$\boldsymbol{\Sigma}_{1,1|2} = \boldsymbol{\Sigma}_{1,1} - \boldsymbol{\Sigma}_{1,2}\boldsymbol{\Sigma}_{2,2}^{-1}\boldsymbol{\Sigma}_{2,1}. \tag{B.69}$$

On constate donc que la matrice de variance-covariance ne dépend pas de la valeur de $\boldsymbol{\mathcal{X}}_2$.

Cas particulier : $p = 2$. Dans le cas particulier où $p = 2$ on a

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \tag{B.70}$$

où

$$\rho \triangleq \frac{\text{cov}\{\mathcal{X}_1, \mathcal{X}_2\}}{\sigma_1\sigma_2}$$

est le coefficient de corrélation linéaire.

La distribution conditionnelle de $\mathcal{X}_1$ étant donné $\mathcal{X}_2$ est alors

$$f(\mathcal{X}_1|\mathcal{X}_2) \sim \mathcal{N}(\mu_1 + \rho\sigma_1\frac{\mathcal{X}_2 - \mu_2}{\sigma_2}, \sigma_1\sqrt{1-\rho^2}). \tag{B.71}$$

La densité n'exsiste que si $|\rho| < 1$ dans ce cas particulier.

Remarques et interprétations. On voit que la distribution Gaussienne est fortement liée à la notion de linéarité. Une distribution Gaussienne est en effet une distribution qui garde sa structure Gaussienne lorsqu'on effectue des transformations linéaires. D'autre part, pour des variables conjointement Gaussiennes, l'espérance conditionnelle est une fonction linéaire, et la matrice de variance-covariance conditionnelle est indépendante de la valeur de la variable qui conditionne. Enfin, il est possible de diagonaliser la matrice de variance-covariance au moyen d'une transformation linéaire (orthogonale). Une fois diagonalisée, les composantes sont indépendantes, ce qui veut dire que dans le cas de distributions Gaussiennes la notion de dépendance probabiliste et celle de dépendance linéaire coïncident essentiellement.

Enfin, on voit que pour un couple de v.a. conjointement Gaussiennes, le coefficient de corrélation linéaire ρ mesure la dépendance entre celles-ci. Il est nul si, et seulement si, les v.a. sont statistiquement indépendantes; il vaut 1 si, et seulement si, l'une des variables est une fonction linéaire de l'autre. Enfin, il prend une valeur non triviale si, et seulement si, les deux variables peuvent s'exprimer sous la forme de deux combinaisons linéaires **linéairement indépendantes** de deux v.a. gaussiennes indépendantes.

Nous verrons dans l'annexe sur les statistiques que les distributions Gaussiennes jouent un rôle très important en estimation statistique, notamment à cause des fortes propriétés mathématiques qui les caractérisent. Pour terminer, signalons que le théorème central-limite formulé ci-dessous s'applique également au cas des v.a. Gaussiennes de $\mathbb{R}^p$.

B.10 SUITES DE V.A. ET NOTIONS DE CONVERGENCE

Il existe différentes façons de définir la notion de convergence de suites de v.a.. Nous les rappelons brièvement ci-dessous en indiquant les relations qui existent entre ces notions, s'il y a lieu.

B.10.1 Convergence en probabilité

Notation: $\xrightarrow{P}$

La suite $(\mathcal{X}_n)$ de v.a. réelles converge en probabilité vers la constante a; si $\forall\epsilon$ et η (arbitrairement petits), $\exists n_0$ tel que $n > n_0$ entraîne

$$P(|\mathcal{X}_n - a| > \epsilon) < \eta. \tag{B.72}$$

On note alors $(\mathcal{X}_n) \xrightarrow{P} a$.

On définit la convergence en probabilité d'une suite de v.a. $(\mathcal{X}_n)$ vers une v.a. $\mathcal{X}$ comme la convergence vers 0 de la suite $(\mathcal{X}_n - \mathcal{X})$.

B.10.2 Convergence presque sûre ou convergence forte

Notation: $\xrightarrow{p.s.}$

La suite $(\mathcal{X}_n)$ de v.a. réelles converge presque sûrement vers $\mathcal{X}$) si :

$$P(\{\omega| \lim_{n\to\infty} \mathcal{X}_n(\omega) \neq \mathcal{X}(\omega)\}) = 0. \tag{B.73}$$

On note alors $(\mathcal{X}_n) \xrightarrow{p.s.} \mathcal{X}$.

La convergence presque sûre implique la convergence en probabilité, c'est pourquoi on l'appelle aussi convergence forte.

B.10.3 Convergence en moyenne d'ordre p

Si $E\{(\mathcal{X}_n - \mathcal{X})^p\}$ existe $\forall n$, alors on a

$(\mathcal{X}_n) \to \mathcal{X}$ en moyenne d'ordre p si $E\{(\mathcal{X}_n - \mathcal{X})^p\} \to 0$.

Le cas pratique usuel est la moyenne quadratique ($p = 2$).

La convergence en moyenne d'ordre p implique la convergence en probabilité.

B.10.4 Convergence en en loi

Notation: $\xrightarrow{\mathcal{L}}$

La suite $(\mathcal{X}_n)$ de v.a. réelles converge en loi vers $\mathcal{X}$ de fonction de répartition $F(\cdot)$ si en tout point de continuité x de $F(\cdot)$, la suite $(F_n(x))$ converge ponctuellement vers $F(x)$. On note

$$(\mathcal{X}_n) \xrightarrow{\mathcal{L}} \mathcal{X}. \tag{B.74}$$

Il s'agit de la convergence la plus faible. En particulier, la convergence en probabilité implique la convergence en loi. Cette dernière est très utilisée en pratique car elle permet d'approximer la fonction de répartition de $(\mathcal{X}_n)$ par celle de $\mathcal{X}$, et réciproquement.

On montre que si $F(\cdot)$ est continue alors la convergence est uniforme (plus que ponctuelle). De plus, si les $F_n(\cdot)$ admettent des densités alors la convergence en loi implique la convergence ponctuelle des densités.

B.11 THEOREMES DE CONVERGENCE

B.11.1 Moivre-Laplace

Ce théorème utile en statistiques, permet d'approximer une loi binomiale par une loi Gaussienne. Il dit que, si $(\mathcal{X}_n)$ forme une suite de v.a. binomiales $\mathcal{B}(n,p)$, alors

$$\frac{\mathcal{X}_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{B.75}$$

B.11.2 Théorème central-limite

Ce théorème établit la convergence en loi vers la loi normale d'une somme de v.a. i.i.d. sous des hypothèses très peu contraignantes. Il dit que, si $(\mathcal{X}_n)$ forme une suite de v.a. i.i.d. de moyenne μ et d'écart-type σ (ces deux moments sont donc supposés exister), alors

$$\left(\frac{\sum_{i=1}^n \mathcal{X}_i - n\mu}{\sigma\sqrt{n}}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1). \tag{B.76}$$

On retrouve comme cas particulier le théorème de Moivre-Laplace, en prenant des variables de Bernouilli.

Contre-exemple : loi de Cauchy.

B.11.3 Lois des grands nombres

B.11.3.1 Loi faible des grands nombres. Soient $\mathcal{X}_i, \forall i = 1, \dots n$ indépendantes d'espérance μ_i finies et de variances σ_i finies, alors

Si $\frac{1}{n}\sum_{i=1}^n \mu_i \to \mu$ *et* $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 \to 0$, *alors* $\overline{\mathcal{X}} \stackrel{\triangle}{=} \frac{1}{n}\sum_{i=1}^n \mathcal{X}_i$ *est telle que*

$$\overline{\mathcal{X}} \xrightarrow{P} \mu. \tag{B.77}$$

Cas particulier : les v.a. $\mathcal{X}_i$ sont i.i.d. μ, σ. On a alors $\frac{1}{n}\sum_{i=1}^n \mu_i = \mu$ et $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 = \frac{\sigma^2}{n}$.

B.11.3.2 Loi forte des grands nombres. *Si Si* $\frac{1}{n}\sum_{i=1}^n \mu_i \to \mu$ *et* $\sum_{i=1}^n \frac{\sigma_i^2}{i^2} \to a$, *alors* $\overline{\mathcal{X}} \stackrel{\triangle}{=} \frac{1}{n}\sum_{i=1}^n \mathcal{X}_i$ *est telle que*

$$\overline{\mathcal{X}} \xrightarrow{p.s.} \mu. \tag{B.78}$$

Cas particulier : les v.a. $\mathcal{X}_i$ sont i.i.d. μ, σ. On a alors $\frac{1}{n}\sum_{i=1}^n \mu_i = \mu$ et $\sum_{i=1}^n \frac{{\sigma_i}^2}{i^2} = \sigma^2 \sum_{i=1}^n \frac{1}{i^2}$, qui converge.

Notes

1. Le terme consacré est en réalité "σ-algèbre de Boole" ou "tribu". Le terme "algèbre" est normalement réservé au cas où la troisième propriété est relaxée à l'union finie. Cependant, dans la suite nous utiliserons la plupart du temps simplement le terme "algèbre" étant entendu que dans le cas infini il faut comprendre σ-algèbre.

2. Dorénavant nous utiliserons la notation $A_1, A_2, \ldots$ pour désigner une suite dénombrable (éventuellement finie) d'ensembles.

3. La notation $A_i \downarrow A$ désigne une suite d'ensembles, telle que $A_{i+1} \subset A_i$ et $\bigcap_i A_i = A$.

4. C'est-à-dire le σ-algèbre de tous les ensembles qui peuvent s'exprimer sous la forme d'une union ou d'une intersection finie ou dénombrable de semi-intervalles. On appelle cet algèbre la tribu Borelienne.

5. Cependant, dans le cours de théorie de l'information on montrera qu'en moyenne l'incertitude concernant une expérience aléatoire diminue, lorsqu'on utilise de l'information complémentaire.

6. En toute généralité, on peut montrer que toute fonction de répartition peut se décomposer en une somme de trois termes ($F(x) = F_c(x) + F_d(x) + F_s(x)$) tels que F_c soit absolument continue (continue et dérivable), F_d est discrète, et F_s (composante singulière) est continue mais ne possède pas de dérivée. Nous supposons que $F_s = 0$.

C RAPPELS DE STATISTIQUE

"No amount of experiments can ever prove me right; a single experiment may at any time prove me wrong."
- Albert Einstein

C.1 INTRODUCTION

La statistique comporte deux volets principaux complémentaires.

Le premier volet est appelé statistique exploratoire ou descriptive et a pour but de synthétiser, résumer, structurer l'information contenue dans des bases de données, dans le but de permettre à un expert humain de l'analyser. Elle utilise pour cela des représentations des données sous diverses formes numériques et graphiques adaptées aux facultés d'analyse humaines et aux types d'informations qu'on cherche à présenter. On utilise aussi le terme plus neutre d'analyse de données. L'outil informatique joue évidemment un rôle prépondérant en analyse de données, et il existe à l'heure actuelle un nombre croissant de logiciels interactifs qui permettent de synthétiser et de visualiser l'information contenue dans des bases de données de grande taille.

Le deuxième volet est la statistique inférentielle, qui vise à inférer à partir des échantillons observés des caractéristiques relatives à une population. La statistique inférentielle fait partie d'un domaine plus général, appelé apprentissage automatique et faisant appel à l'inférence inductive. L'inférence inductive consiste en effet à formuler (ou à rejeter) des hypothèses générales à partir d'observations particulières. A l'intérieur de ce domaine, les méthodes d'inférence statistique se caractérisent essentiellement par le fait qu'elles formulent des hypothèses de nature probabiliste. Une fois formulées, ces hypothèses permettent ensuite de faire des prédictions quant au comportement de futures observations. Le calcul des probabilités joue évidemment un rôle privilégié dans ce contexte.

Il faut noter que les activités d'analyse de données et d'inférence inductive constituent un pan important des méthodes utilisées dans la plupart (sinon toutes) les disciplines scientifiques. Par ailleurs, dans la vie de tous les jours leur utilisation (informelle) nous permet d'apprendre à partir de nos expériences. Mais, comme le souligne la citation d'Albert Einstein en tête de chapitre, dans le domaine du raisonnement inductif on ne peut tirer des conclusions définitives, car on n'est jamais à l'abri de nouvelles observations qui viennent contredire ce qui paraissait presque certain au vu des observations antérieures. C'est la raison pour laquelle le calcul des probabilités joue un rôle aussi fondamental dans ce domaine, car il fournit un modèle mathématique non seulement pour décrire les phénomènes physiques, mais aussi (et sans doute de manière plus fondamentale) pour quantifier le degré de confiance que nous pouvons avoir en une telle description.

La démarche habituelle, lorsqu'on est en présence d'une base de données, consiste d'abord à utiliser des méthodes d'analyse de données de façon à se familiariser avec ces données et à formuler des hypothèses générales (indépendance, linéarité, normalité…). Ensuite, on utilise des méthodes inférentielles afin de construire des modèles plus précis et de quantifier la confiance qu'on peut avoir dans ces modèles.

Dans cet appendice nous ne parlerons pas des méthodes d'analyse de données (les méthodes élémentaires ont été vues au cours de statistique, et d'autres seront vues au cours d'apprentissage inductif appliqué). Nous nous contenterons de rappeler un certain nombre de concepts et de résultats importants en statistique inférentielle, en nous limitant pour l'essentiel aux problèmes unidimensionnels.

C.2 NOTION D'ECHANTILLON STATISTIQUE

On dispose d'une suite a priori ordonnée $x_1, x_2, \ldots, x_n$ d'observations d'une certaine grandeur physique. Les x_i peuvent représenter des données simples (par exemple, les tailles d'individus successifs rencontrés par une personne particulière) ou complexes par exemple les messages électroniques successifs reçus par une certaine personne. Nous utiliserons généralement le terme "observation" pour faire référence à l'un des x_i et le terme "échantillon" pour faire référence à l'ensemble d'observations $x_1, x_2, \ldots, x_n$.

La notion d'échantillon *statistique* apparaît lorsqu'on formule une ou plusieurs hypothèses de nature probabiliste sur le mécanisme qui produit les échantillons. Par exemple, une bonne partie (mais certainement pas toutes) des méthodes d'inférence statistique font l'hypothèse d'observations "indépendantes et identiquement distribuées" (i.i.d.).

Il est important de remarquer à ce stade que l'hypothèse faite par l'analyste peut ne pas coller avec la réalité. Par exemple, si les observations sont effectuées sur un système dont les caractéristiques évoluent au cours du temps ou qui présente une certaine mémoire, l'hypothèse "i.i.d" ne sera pas valable.

Par ailleurs, dans de très nombreux cas, les échantillons sont sélectionnés d'une manière ou d'une autre par l'expérimentateur, ce qui peut introduire également des biais involontaires. Dans d'autres cas (par exemple dans les sondages) on introduit volontairement un biais lors de l'échantillonnage dans le but de maximiser l'information utile dans l'echantillon.

Nous reviendrons sur ces aspects ultérieurement. Pour le moment nous supposerons que nous avons effectivement à faire à des échantillons i.i.d. Notons qu'on peut voir, du point de vue probabiliste, un tel échantillon de deux façons :

- les $x_1, x_2, \ldots, x_n$ corespondent à n réalisations d'une variable aléatoire $\mathcal{X}(\omega)$: $x_i = \mathcal{X}(\omega_i)$, avec $\omega \sim (\Omega, \mathcal{E}, P(\cdot))$.
- $(x_1, x_2, \ldots, x_n) = \mathcal{X}^n(\omega^n)$ correspond à une unique réalisation d'une variable vectorielle définie sur l'espace produit $\omega^n \sim (\Omega^n, \mathcal{E}^n, P^n(\cdot))$.

C.3 THEORIE DE L'ECHANTILLONNAGE

La théorie de l'échantillonnage étudie les propriétés du n-tuple $(x_1, x_2, \ldots, x_n)$ et principalement des caractéristiques le résumant, à partir de la distribution supposée connue de la variable $\mathcal{X}$. On appelle de façon générique *statistique* toute grandeur qui peut s'écrire comme une fonction (au sens tout à fait général du terme)

$$f(x_1, x_2, \ldots, x_n). \tag{C.1}$$

Comme nous pouvons voir notre échantillon comme une réalisation d'une v.a. vectorielle sur l'espace produit, il s'en suit qu'une statistique définit une variable aléatoire. Notons qu'une statistique peut être à valeurs discrètes, réelles (le cas le plus fréquent), vectorielles ou même fonctionnelles. Par exemple, en apprentissage automatique on déduit d'un échantillon des fonctions entrée/sortie (p.ex. des règles de décision); ces fonctions sont donc des fonctions aléatoires, ce qui veut dire qu'elles définissent en réalité une famille de variables aléatoires indicée par les variables d'entrée. D'ailleurs, le premier exemple de statistique que nous allons discuter ci-dessous est également une fonction aléatoire.

Notons qu'un certain nombre de résultats en statistique sont de nature asymptotique, ce qui veut dire qu'ils sont valables lorsque $n \to +\infty$.

C.3.1 Fonction de répartition empirique d'un échantillon

Nous supposons que $x_i \in \mathbb{R}$. Soit alors un échantillon $x_1, \ldots, x_n$ i.i.d. et désignons par

$$F^*_{x_1,\ldots,x_n}(x) \tag{C.2}$$

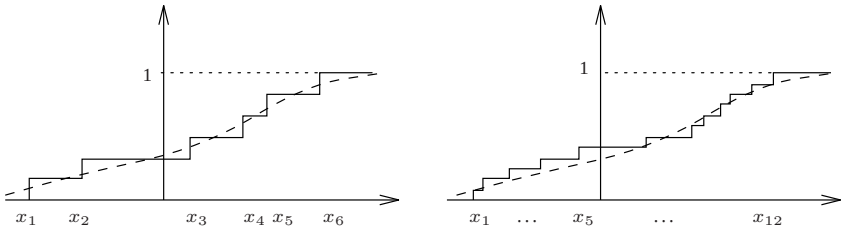

Figure C.1. Fonctions de répartition empiriques

la proportion des n valeurs de x_i inférieures à x. Il est clair que quel que soit l'échantillon cette fonction est croissante, et que

$$\lim_{x\to-\infty} F^*_{x_1,\ldots,x_n}(x) = 0 \text{ et } \lim_{x\to+\infty} F^*_{x_1,\ldots,x_n}(x) = 1. \tag{C.3}$$

On voit également que la fonction est constante par morceaux et présente des discontinuités (sauts d'amplitude $\frac{1}{n}$) aux points $x = x_i$ qui correspondent aux observations.

La figure C.1 résume graphiquement deux exemples de fonctions $F^*(x)$ obtenus pour différents échantillons d'une même variable aléatoire dont la fonction de répartition $F(x)$ est également illustrée (en traits interrompus). Ici nous avons supposé que les échantillons ont été numérotés par ordre croissant de leurs valeurs.

Si les x_i sont numérotés par ordre croissant de leurs valeurs on a

$$\begin{array}{lll} F^*_{x_1,\ldots,x_n}(x) = 0 & \text{si} & x < x_1, \\ F^*_{x_1,\ldots,x_n}(x) = \frac{i-1}{n} & \text{si} & x_{i-1} \le x < x_i, \\ F^*_{x_1,\ldots,x_n}(x) = 1 & \text{si} & x \ge x_n. \end{array} \tag{C.4}$$

On voit que $F^*_{x_1,\ldots,x_n}(x)$ prend l'allure d'une fonction de répartition d'une variable discrète. Nous allons voir que lorsque la taille de l'échantillon tend vers l'infini, cette fonction converge vers la fonction de répartition $F(x)$ de la variable aléatoire $\mathcal{X}$, et que la façon dont cette convergence se manifeste est indépendante de la forme de $F(x)$. La valeur $F^*_{x_1,\ldots,x_n}(x)$, pour n donné et $\forall x \in \mathbb{R}$ constitue une variable aléatoire réelle comprise dans l'intervalle $[0, 1]$; nous la noterons $F^*_n(x)$.

On a alors les trois théorèmes suivants :

Théorème 1.

Pour tout x et pour $n \to \infty$, on a $F^*_n(x) \overset{p.s.}{\to} F(x)$.

Théorème 2 (Glivenko-Cantelli).

La convergence de $F^*_n(\cdot)$ vers $F(\cdot)$ est presque sûrement uniforme, c'est-à-dire que :

$$D_n = \sup_{x\in\mathbb{R}} |F^*_n(x) - F(x)| \overset{p.s.}{\to} 0.$$

La grandeur D_n est aussi appelée distance de Kolmogorov-Smirnov entre la fonction de répartition empirique $F^*_n(x)$ et la fonction de répartition $F(x)$.

Théorème 3 (Kolmogorov).

Pour tout n, la distribution de la variable D_n est indépendante de l'allure de la fonction $F(x)$ qui caractérise la v.a. étudiée.

Asymptotiquement ($n > 35$), elle obéit à l'équation suivante, $\forall y > 0$:

$$\lim_{n\to+\infty} P(\sqrt{n}D_n < y) = K(y) = \sum_{k=-\infty}^{k=+\infty} (-1)^k e^{-2k^2y^2}.$$

Ce dernier théorème permet de formuler de nombreux tests d'hypothèse en statistique inférentielle. Les quantiles de D_n correspondant à différentes valeurs de n sont tabulées dans la plupart des recueils de statistique et tables numériques.

Ces trois théorèmes constituent sans doute le résultat le plus fondamental en statistique, puisqu'il est à la base de la justification de l'usage des échantillons.

Notons que ces résultats se généralisent de diverses manières, en particulier on montre que si on dispose de deux échantillons $x_1, \ldots, x_n$ et $x'_1, \ldots, x'_{n'}$ issus d'une même population (même $F(x)$) alors la distance de Kolmogorov-Smirnov entre les deux fonctions de répartition empiriques $F^*_n(x)$ et $F^*_{n'}(x)$ a une distribution indépendante de la forme de $F(x)$. Ce résultat permet en particulier de formuler un test d'hypothèse "non-paramétrique" pour comparer deux échantillons.

C.3.2 Distributions d'échantillonnage de certains moments

Les notions d'estimateurs, de biais et de variance sont définies de manière plus précise à la section suivante.

Nous décrivons ci-dessous les propriétés élémentaires des statistiques les plus utilisées, et en anticipant sur la section suivante nous indiquons les qualités des estimateurs qu'on peut en déduire.

C.3.2.1 Moyenne d'échantillon.

Nous rappelons que la statistique $\overline{\mathcal{X}}$ est la moyenne d'échantillon définie par

$$\overline{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i. \tag{C.5}$$

On a

$$E\{\overline{\mathcal{X}}\} = E\{\mathcal{X}\}, \tag{C.6}$$

et

$$V\{\overline{\mathcal{X}}\} = \frac{1}{n} V\{\mathcal{X}\}. \tag{C.7}$$

pour autant que $V\{\mathcal{X}\}$ existe.

Lorsque $n \to \infty$, $V\{\overline{\mathcal{X}}\} \to 0$. Il s'ensuit que $E\{\overline{\mathcal{X}}\}$ converge en moyenne quadratique vers $E\{\mathcal{X}\}$. Les lois des grands nombres permettent d'affirmer que pour autant que $V\{\mathcal{X}\}$ est finie, $E\{\overline{\mathcal{X}}\} \stackrel{P}{\to} E\{\mathcal{X}\}$ et $E\{\overline{\mathcal{X}}\} \stackrel{p.s.}{\to} E\{\mathcal{X}\}$.

Le théorème central limite nous dit que sous les mêmes conditions $E\{\overline{\mathcal{X}}\}$ converge en loi vers une loi Gaussienne, indépendamment de la forme de $F(x)$. Cependant, nous attirons l'attention du lecteur sur le fait que la forme de $F(x)$ influence (notablement) la vitesse à laquelle cette convergence en loi est assurée.

En particulier, nous savons déjà que si $\mathcal{X} \sim \mathcal{N}(\mu, \sigma^2)$ alors $\overline{\mathcal{X}}$ qui est combinaison linéaire de variables Gaussiennes indépendantes est distribuée de façon Gaussienne, $\forall n$ fini.

C.3.2.2 Probabilité d'un événement.

Soit une variable aléatoire discrète $\mathcal{X}$ prenant la valeur 1 avec une certaine probabilité p et la valeur 0 avec la probabilité $q = 1 - p$. On dit qu'une telle variable est une variable indicatrice d'un événement. En effet, on peut toujours définir l'ensemble $A \subset \Omega$ défini par

$$A = \{\omega \in \Omega | X(\omega) = 1\}. \tag{C.8}$$

Cet ensemble doit faire partie de $\mathcal{E}$, car sinon la fonction $X(\cdot)$ ne définirait pas une variable aléatoire (elle ne serait pas mesurable). Nous utiliserons fréquemment de telles variables indicatrices, et nous utiliserons la notation

$$X(\cdot) = 1_A(\cdot). \tag{C.9}$$

Notons que $E\{1_A\} = P(A)$ et $V\{1_A\} = P(A)(1 - P(A))$. Par conséquent, la moyenne d'échantillon de 1_A fournit une estimée de $P(A)$, et asymptotiquement cette estimée sera distribuée en loi Gaussienne de variance $\frac{1}{n}P(A)(1 - P(A))$. Pour n fini, la loi est une loi Binomiale $\mathcal{B}(n, P(A))$.

C.3.2.3 Variance empirique.

La variance empirique, désignée par $\mathcal{S}^2$ (nous utiliserons également la notation $\hat{\sigma}^2$) est la grandeur suivante

$$\mathcal{S}^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}(\mathcal{X}_i - \overline{\mathcal{X}})^2. \tag{C.10}$$

Il est important de remarquer que dans cette formule $\overline{\mathcal{X}}$ est la moyenne d'échantillon et non la valeur de $E\{\mathcal{X}\}$.

On montre que

$$E\{\mathcal{S}^2\} = \frac{n-1}{n} V\{\mathcal{X}\}, \tag{C.11}$$

ce qui veut dire que la variance empirique sous-estime la variance d'autant plus que l'échantillon est petit. Elle est donc biaisée.

On montre que $\mathcal{S}^2$ et $\overline{\mathcal{X}}$ sont asymptotiquement non corrélées quelle que soit la fonction de répartition $F(x)$. De plus, si la loi est symétrique (i.e. $F(x+\mu) = 1 - F(x-\mu)$ où $\mu = E\{\mathcal{X}\}$), alors $\mathcal{S}^2$ et $\overline{\mathcal{X}}$ sont non corrélées pour des tailles finies de l'échantillon. Si $\mathcal{X}$ est Gaussienne, alors

$$\frac{n\mathcal{S}^2}{V\{\mathcal{X}\}}$$

suit une loi en χ^2_{n-1}. De plus, comme cette distribution est symétrique, $\mathcal{S}^2$ et $\overline{\mathcal{X}}$ sont non corrélés, et on montre qu'elles sont également indépendantes. Sous ces conditions, toute l'information contenue dans l'échantillon relative à $F(x)$ est fournie par ces deux statistiques.

On montre que la statistique suivante (obtenue en remplaçant dans (C.10) $\overline{\mathcal{X}}$ par $E\{\mathcal{X}\}$)

$$\mathcal{S}'^2 \triangleq \frac{1}{n} \sum_{i=1}^{n} (\mathcal{X}_i - E\{\mathcal{X}\})^2. \tag{C.12}$$

est un estimateur non-biaisé de $V\{\mathcal{X}\}$:

$$E\{\mathcal{S}'^2\} = V\{\mathcal{X}\}. \tag{C.13}$$

Cet estimateur n'est cependant utilisable que si on dispose d'une connaissance a priori de $E\{\mathcal{X}\}$.

Evidemment l'estimateur

$$\mathcal{S}^2_{n-1} \triangleq \frac{n}{n-1} \mathcal{S}^2, \tag{C.14}$$

est non-biaisé. Cependant, sa variance est plus élevée que celle de $E\{\mathcal{S}^2\}$.

C.3.2.4 Vecteur aléatoire Gaussien.

Soit $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ un échantillon d'un vecteur aléatoire $\boldsymbol{\mathcal{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Notre échantillon peut donc être vu comme un tableau de n lignes et p colonnes.

On désigne par $\overline{\boldsymbol{\mathcal{X}}}$ la moyenne (vecteur) de l'échantillon définie par

$$\overline{\boldsymbol{\mathcal{X}}} \triangleq \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\mathcal{X}}_i, \tag{C.15}$$

et par $\boldsymbol{V}$ la matrice d'ordre p de variance de l'échantillon définie par

$$\boldsymbol{V} \triangleq \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\mathcal{X}}_i - \overline{\boldsymbol{\mathcal{X}}})(\boldsymbol{\mathcal{X}}_i - \overline{\boldsymbol{\mathcal{X}}})^T. \tag{C.16}$$

On montre, à partir des propriétés des lois Gaussiennes (cf. appendice B), que $\overline{\boldsymbol{\mathcal{X}}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$. La matrice $\boldsymbol{V}$ suit une loi de Wishart à $n-1$ degrés de libertés (voir [Sap90]).

C.4 ESTIMATION

L'estimation consiste à donner des valeurs approchées aux paramètres d'une population $(\mu, \sigma, F(x), \ldots)$. Idéalement, le but est évidemment de choisir des estimateurs à la fois précis et robustes. Malheureusement, ces deux qualités sont en général antagonistes.

Dans ce qui suit nous allons désigner de façon générique par θ le paramètre qui est estimé. Nous supposerons que θ peut prendre ses valeurs dans un certain ensemble Θ défini a priori. Cet ensemble peut être une partie de $\mathbb{R}$, de $\mathbb{R}^p$ ou d'un espace de fonctions.

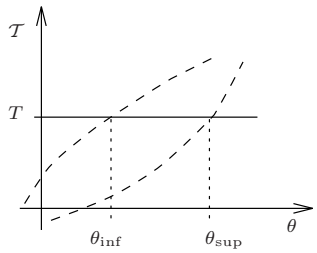

Figure C.2. Intervalles de confiance

Une partie très importante des statistiques est dediée au développement d'estimateurs ayant de bonnes performances dans diverses conditions. Les conditions pour lesquelles un estimateur est développé peuvent être spécifiées selon différents critères

- taille des échantillons (statistiques asymptotiques vs petits échantillons),
- caractéristiques des distributions $F(x)$ (statistiques paramétriques vs non-paramétriques)
- robustesse (vis-à-vis d'écarts aux hypothèses)
- échantillonnage (indépendant, structuré, corrélé,...)
- avec connaissances a priori (de certains moments, de la forme de $F(x)$, ...)

Dans ce qui suit nous allons surtout considérer le cas des échantillons i.i.d. et définir la notion d'estimateur et les principales caractéristiques utilisées pour le décrire.

Notons qu'il existe essentiellement trois stratégies pour estimer une grandeur

Estimation ponctuelle. Dans ce cas on utilise une fonction de l'échantillon (une statistique $\mathcal{T}$) pour estimer la valeur du paramètre. Nous venons de donner quelques exemples d'estimateurs ponctuels dans la section précédente.

L'estimation ponctuelle consiste donc à choisir une valeur $\theta_0 \in \Theta$ comme étant la valeur la plus plausible du paramètre.

Estimation par région de confiance. L'idée est de construire un sous-ensemble de valeurs du paramètre θ qui sont compatibles avec la valeur observée de la statistique $\mathcal{T}$. Pour une valeur donnée de θ on peut construire un intervalle de valeurs pour $\mathcal{T}$ de probabilité donnée $(1-\alpha)$. On accepte alors une valeur de θ comme plausible si cet intervalle contient la valeur de $\mathcal{T}$ déduite de l'échantillon. Enfin, on définit la région de confiance de θ comme l'ensemble des valeurs de θ acceptées par cette procédure. Cette procédure est illustrée à la figure C.2. Pour une valeur donnée de θ, la borne inférieure (resp. supérieure) de l'intervalle de probablité $1-\alpha$ pour $\mathcal{T}$ est donnée par le point de la courbe en traits interrompus inférieure (resp. supérieure). Si T représente la valeur de $\mathcal{T}$ calculée sur base de l'échantillon, on voit que l'intervalle de confiance $[\theta_{\inf}, \theta_{\sup}]$ est obtenue à partir des points d'intersection dela droite $\mathcal{T} = T$ et ces deux courbes. Cette approche, illustrée ici dans le cas où θ est un paramètre scalaire, peut évidemment se généraliser au cas de paramètres vectoriels. On parle alors de régions de confiance.

L'estimation par région de confiance consiste donc à choisir un sous-ensemble de valeurs $\Theta_0 \subset \Theta$.

Estimation Bayesienne. Dans cette approche on considère que le paramètre à estimer est une variable aléatoire, caractérisée par une certaine distribution de probabilité a priori $p(\theta)$. On se sert ensuite de l'échantillon pour "conditionner" cette distribution de probabilité, ce qui donne par application de la formule de Bayes

$$p(\theta|x_1,\ldots,x_n) = \frac{p(\theta)p(x_1,\ldots,x_n|\theta)}{\int_\Theta p(x_1,\ldots,x_n|\theta)p(\theta)d\theta}. \tag{C.17}$$

Lorsque les échantillons sont indépendants on peut remplacer dans cette formule le terme $p(x_1,\ldots,x_n|\theta)$ par $\prod_{i=1}^n p(x_i|\theta)$.

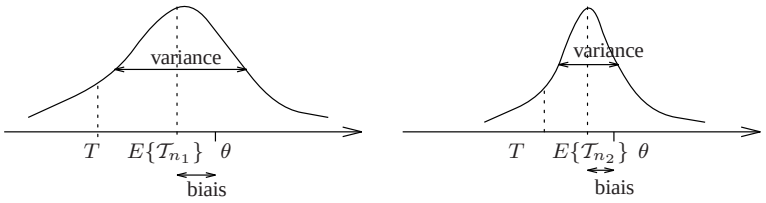

Figure C.3. Biais et variance d'un estimateur

L'estimation Bayesienne consiste donc à construire une loi de probabilité sur l'ensemble Θ. Il est évident qu'à partir d'une telle loi on peut extraire une région de confiance (p.ex. telle que $\int_{\Theta_0} p(\theta|x_1,\ldots,x_n)d\theta = 1-\beta$) ou un estimateur ponctuel (p.ex. tel que $p(\theta_0|x_1,\ldots,x_n)$ soit maximale). Mais ces régions ou valeurs ponctuelles dépendent du choix de $p(\theta)$ et sont donc en général différentes de celles obtenues par les deux méthode précédentes.

Nous allons nous limiter ici à l'estimation ponctuelle, en faisant cependant remarquer que le paramètre θ peut très bien désigner un vecteur, voire une fonction définie sur un ensemble quelconque.

Estimation paramétrique. Dans ce qui suit nous allons nous intéresser à une famille de problèmes d'estimation qui se posent comme suit : on dispose d'un échantillon $x_1,\ldots,x_n$ i.i.d. selon une loi discrète ou continue $F(\mathcal{X})$ qui appartient à une famille supposée connue de lois indicées par un ensemble (fini) de paramètres $\theta_1,\theta_2,\ldots,\theta_M$ inconnus. Cela veut dire qu'une fois que les valeurs de ces paramètres sont toutes connues (et seulement dans ce cas) on dispose d'une connaissance complète de la loi $F(\mathcal{X})$. Lorsque nous en aurons besoin, nous mettrons en évidence ce fait en notant cette loi $F_{(\theta_1,\theta_2,\ldots,\theta_M)}(\mathcal{X})$.

Dans de nombreux cas, on ne souhaite pas d'information concernant certains des paramètres inconnus. On les désignera par le terme générique de paramètres fantômes, et dans ce qui suit nous noterons par θ le sous-ensemble (éventuellement réduit à un seul élément) des paramètres $\theta_1,\theta_2,\ldots,\theta_M$ qu'on souhaite estimer.

C.4.1 Qualités des estimateurs ponctuels

Soit θ le paramètre à estimer, et soit $\mathcal{T}$ un estimateur ponctuel, c'est-à-dire une fonction des $\mathcal{X}_i$ de notre échantillon dont le domaine de valeurs est compatible avec les valeurs acceptables pour θ (par exemple si θ est une variance il faut que $\mathcal{T}$ soit au moins positif; si θ est la probabilité d'un événement il faut que $\mathcal{T} \in [0,1]\ldots$).

Liberté. Un estimateur $\mathcal{T}$ de θ est dit libre si la loi de probabilité de $\mathcal{T}$ est indépendante des valeurs des paramètres fantômes. Un cas particulier trivial d'estimateur libre se présente lorsqu'il n'y a pas de paramètres fantômes.

Par contre, nous avons vu que la distribution d'échantillonnage de la moyenne empirique dépend en général de la variance de $\mathcal{X}$. Par conséquent, si celle-ci est inconnue elle constitue un paramètre fantôme influent, et la moyenne empirique n'est alors pas un estimateur libre.

Convergence. La première qualité qu'on demande la plupart du temps à un estimateur est que si $n \to \infty$ alors $\mathcal{T} \to \theta$. On dit que l'estimateur est convergent ou "consistent" (anglicisme).

Pour un paramètre quelconque on peut construire un très grand nombre d'estimateurs convergents. Par exemple, les estimateurs décrits à la section précédente sont tous convergents. La question se pose donc de choisir parmi tous ces estimateurs ceux qui auront de bonnes performances pour des tailles finies d'un échantillon, ce qui conduit à la notion de précision.

Biais et variance. La figure C.3 représente graphiquement la distribution d'échantillonnage d'un estimateur $\mathcal{T}$ de θ, pour deux tailles de l'échantillon. Pour plus de clarté, nous notons par $\mathcal{T}_n$ l'estimateur obtenu à partir d'une taille d'échantillon n.

L'erreur d'estimation $\mathcal{T}_n - \theta$ (variable aléatoire) se décompose en deux termes élémentaires $\mathcal{T}_n - E\{\mathcal{T}_n\}$ et $E\{\mathcal{T}_n\} - \theta$, où $E\{\mathcal{T}_n\}$ est l'espérance de l'estimateur. $\mathcal{T}_n - E\{\mathcal{T}_n\}$ représente les fluctuations aléatoires de $\mathcal{T}_n$ autour de sa moyenne, tandis que $E\{\mathcal{T}_n\} - \theta$ est assimilable à une erreur systématique. Cette quantité s'appelle le biais.

Si on mesure la précision d'un estimateur par son erreur quadratique moyenne, définie par

$$E\{(\mathcal{T}_n - \theta)^2\} \tag{C.18}$$

alors on peut décomposer celle-ci en un terme relatif à la variance et un terme relatif au biais. En effet, on a

$$\begin{aligned} E\{(\mathcal{T}_n - \theta)^2\} &= E\{[(\mathcal{T}_n - E\{\mathcal{T}_n\}) + (E\{\mathcal{T}_n\} - \theta)]^2\} \\ &= E\{(\mathcal{T}_n - E\{\mathcal{T}_n\})^2\} + 2E\{(\mathcal{T}_n - E\{\mathcal{T}_n\})(E\{\mathcal{T}_n\} - \theta)\} \\ &\quad + E\{(E\{\mathcal{T}_n\} - \theta)^2\} \\ \Rightarrow E\{(\mathcal{T}_n - \theta)^2\} &= V\{\mathcal{T}_n\} + [E\{\mathcal{T}_n\} - \theta]^2. \end{aligned} \tag{C.19}$$

En effet,

$$2E\{(\mathcal{T}_n - E\{\mathcal{T}_n\})(E\{\mathcal{T}_n\} - \theta)\} = 2(E\{\mathcal{T}_n\} - \theta)E\{(\mathcal{T}_n - E\{\mathcal{T}_n\})\} = 0$$

car $E\{(\mathcal{T}_n - E\{\mathcal{T}_n\}\} = 0$, et

$$E\{(E\{\mathcal{T}_n\} - \theta)^2\} = [E\{\mathcal{T}_n\} - \theta]^2,$$

puisque $E\{\mathcal{T}_n\} - \theta$ est une constante.□

Cette formule est d'une importance capitale. En effet, elle montre que l'erreur quadratique moyenne d'un estimateur est composée des deux termes suivants :

Biais. C'est l'erreur quadratique de l'estimateur moyen obtenu de la manière suivante : on prélève un premier échantillon de taille n ($x_1^1, \ldots, x_n^1$) et on calcule $T_n^1 = T(x_1^1, \ldots, x_n^1)$, on répète cette procédure un nombre infini de fois et on calcule la moyenne $\overline{T}_n$ des valeurs T_n^i obtenues (la loi des grands nombres assure que cette moyenne converge vers $E\{\mathcal{T}_n\}$). Le résultat de cette procédure donne ce que nous appelerons dans la suite l'estimateur moyen pour un échantillon de taille n. L'erreur quadratique de cet estimateur moyen est le biais au carré.

Variance. C'est l'écart quadratique moyen des modèles construits ci-dessus par rapport à l'estimateur moyen.

Nous verrons ci-dessous, lorsque nous parlerons de la régression linéaire (et de manière plus approfondie au cours d'apprentissage automatique), que dans les problèmes d'estimation fonctionnelle (θ est alors une fonction $\theta(\cdot)$ d'un certain nombre de variables), ces deux grandeurs varient généralement de façon antagoniste, la variance augmentant lorsque le biais diminue et réciproquement. Dans de telles situations on est amené à un compromis entre biais et variance.

Discussion.

Lorsque la vraie valeur de θ est inconnue, ni le biais, ni la variance, ni même l'erreur quadratique moyenne d'estimation que nous avons définie plus haut, ne sont directement accessibles à l'expérimentateur, à moins de disposer de tailles d'échantillons énormes, ou d'utiliser des techniques de ré-échantillonnage que nous discuterons brièvement à la fin de ce chapitre.

Dans le cas fréquent en pratique où on estime l'espérance mathématique d'une variable aléatoire (problème de régression) à partir d'un échantillon $x_1, \ldots, x_n$ de cette v.a., on utilise pour cette raison souvent une autre mesure, accessible à l'expérimentateur, appelée erreur quadratique empirique moyenne (MSE) et définie de la manière suivante :

$$\text{MSE} \triangleq \frac{1}{n} \sum_{i=1}^{n} |x_i - T_n(x_1, \ldots, x_n)|^2. \tag{C.20}$$

Nous savons que la valeur $T = E\{\mathcal{X}\}$ minimise l'espérance mathématique de cette grandeur (cf appendice B). C'est pourquoi, certaines méthodes d'estimation consistent à choisir dans un ensemble de valeurs candidates la valeur de T_n qui minimise l'erreur quadratique définie par (C.20), pour *apprendre* la valeur de de T_n; l'échantillon $x_1, \ldots, x_n$ porte alors le nom d'échantillon d'apprentissage.

Désignons par T^{MSE} l'estimée ainsi obtenue. Comme dans la formule (C.20) les échantillons x_i et l'estimateur $T_n^{\text{MSE}}(x_1, \ldots, x_n)$ ne sont pas indépendants, il n'est pas possible en général de déduire l'espérance mathématique de MSE, et encore moins sa distribution de probabilité.

Par contre, si on dispose d'un échantillon de test indépendant de taille m, $x'_1, \ldots, x'_m$, on peut calculer la grandeur suivante

$$\text{MSE}(TS) \triangleq \frac{1}{m} \sum_{i=1}^{m} |x'_i - T_n^{\text{MSE}}(x_1, \ldots, x_n)|^2. \tag{C.21}$$

A la limite, lorsque $m \to \infty$ cette grandeur tend vers

$$E_{\mathcal{X}}\{(\mathcal{X} - T_n^{\text{MSE}}(x_1, \ldots, x_n))^2\} \tag{C.22}$$

où l'espérance mathématique est prise par rapport aux échantillons de test. Cette grandeur donnera une évaluation non pas de l'estimateur en général, mais bien de la valeur estimée à partir de l'échantillon $x_1, \ldots, x_n$. Si nous prenons l'espérance mathématique de cette grandeur par rapport à la distribution des échantillons d'apprentissage nous aurons

$$E\{(\mathcal{X} - \mathcal{T}_n)^2\} = E_{\mathcal{X}_1,\ldots,\mathcal{X}_n}\{\big(E_{\mathcal{X}}\{(\mathcal{X} - \mathcal{T}_n(\mathcal{X}_1, \ldots, \mathcal{X}_n))^2\}\big)\}, \tag{C.23}$$

où nous avons explicité sur quelles variables les opérations d'espérance mathématique portent. Un calcul assez similaire au précédent donne le résultat suivant

$$E\{(\mathcal{X} - \mathcal{T}_n)^2\} = V\{\mathcal{X}\} + V\{\mathcal{T}_n\} + [E\{\mathcal{T}_n\} - E\{\mathcal{X}\}]^2. \tag{C.24}$$

Le nouveau terme qui vient s'ajouter aux deux précédents est la variance de la variable aléatoire $\mathcal{X}$. Nous verrons, au cours d'*Apprentissage Inductif Appliqué*, que ce terme porte aussi le nom d'erreur résiduelle (et est parfois qualifié de "bruit"). Ce terme ne dépendant pas de notre estimateur, est en effet inévitable. Il s'agit de l'incertitude concernant la valeur prise par la variable aléatoire $\mathcal{X}$ lors d'une expérience dans laquelle on connaîtrait parfaitement la valeur de $E\{\mathcal{X}\}$. Il s'agit donc de l'incertitude sur $\mathcal{X}$ que notre estimateur $\mathcal{T}$ de $E\{\mathcal{X}\}$ ne peut de toutes façons pas résoudre, même s'il est de biais et de variance négligeables.

Lorsqu'on parle d'erreur d'estimation, il faut donc bien faire attention à ce à quoi on fait référence : si on fait référence à l'estimation d'un paramètre tel que la moyenne, qui caractérise la distribution d'une variable aléatoire on doit comptabiliser biais et variance de l'estimateur. Si par contre on parle de l'estimation de la variable aléatoire elle-même, il faut en outre ajouter la variance de celle-ci. Dans les sous-sections suivantes nous allons nous focaliser sur l'estimation de paramètres.

Estimateurs sans biais et de variance minimale. On dit que l'estimateur est sans biais si $E\{\mathcal{T}\} = \theta$. On peut définir des estimateurs sans biais pour une classe assez large de familles de distributions.

Si, parmi tous les estimateurs sans biais il en existe un dont la variance est inférieure ou égale à celle des autres, pour une famille de lois $F(\mathcal{X})$, on dit que cet estimateur est de variance minimale, ou encore efficace[1]. Notons d'emblée qu'un estimateur est en général efficace seulement pour une famille réduite de distributions $F(\mathcal{X})$. On montre que la moyenne empirique est un estimateur efficace pour une loi Gaussienne.

On dit qu'un estimateur est asymptotiquement efficace si lorsque la taille de l'échantillon devient infinie il tend vers un estimateur efficace.

Notons qu'un estimateur sans biais et de variance minimale n'est pas nécessairement le plus précis. En d'autres termes, il existe parfois des estimateurs biaisés dont la variance est inférieure à celle d'un estimateur efficace, et ceci suffisamment pour compenser leur biais.

On montre que si un estimateur efficace existe il est unique (presque partout).

Exhaustivité. Soit θ un paramètre scalaire de la distribution de probabilités de $\mathcal{X}$, et soit $x_1, \ldots, x_n$ un échantillon aléatoire i.i.d. selon la loi de probabilités $p_\theta(\cdot)$ où nous mettons en évidence la dépendance de cette loi vis-à-vis du paramètre θ. Comme nous l'avons souligné ci-dessus, la loi de $\mathcal{X}$ peut dépendre d'autres paramètres, qui agissent alors comme des paramètres fantômes. Par exemple, si nous supposons que la loi est une loi exponentielle, nous savons que le paramètre λ la caractérise entièrement. Mais si nous faisons référence à l'estimation de la moyenne d'une loi Gaussienne de variance inconnue, la variance sera un paramètre fantôme.

On dit que $\mathcal{T}$ est une statistique exhaustive si toute l'information concernant θ contenue dans l'échantillon est aussi contenue dans $\mathcal{T}$. Autrement dit, $\mathcal{T}$ est exhaustive si

$$p(\mathcal{X}_1, \ldots, \mathcal{X}_n | \mathcal{T}) \tag{C.25}$$

est indépendante de θ.

Notons par

$$\mathcal{L}(\mathcal{X}_1, \ldots, \mathcal{X}_n; \theta_0) \tag{C.26}$$

la densité probabilité si $\mathcal{X}$ est continue (resp. la loi de probabilité, si $\mathcal{X}$ est discrète) d'observer l'échantillon $\mathcal{X}_1, \ldots, \mathcal{X}_n$, sous l'hypothèse que $\theta = \theta_0$. Cette grandeur porte le nom de vraisemblance, et en anglais *likelihood*.

Soit alors $\mathcal{T}$ une statistique (fonction de $x_1, \ldots, x_n$) et soit $g(\mathcal{T}; \theta_0)$ la densité (resp. loi) de celui-ci. On peut montrer que $\mathcal{T}$ est exhaustif si on peut factoriser la vraisemblance de la manière suivante :

$$\mathcal{L}(\mathcal{X}_1, \ldots, \mathcal{X}_n; \theta) = g(\mathcal{T}(\mathcal{X}_1, \ldots, \mathcal{X}_n); \theta) h(\mathcal{X}_1, \ldots, \mathcal{X}_n). \tag{C.27}$$

L'importance de la notion d'exhaustivité d'une statistique provient du théorème suivant :

S'il existe une statistique exhaustive $\mathcal{T}$ et un estimateur efficace, ce dernier peut s'écrire sous la forme d'une fonction de $\mathcal{T}$ seulement.

C.4.2 Méthode du maximum de vraisemblance

Observons tout d'abord que lorsque les échantillons sont i.i.d. la vraisemblance se factorise en

$$\mathcal{L}(\mathcal{X}_1, \ldots, \mathcal{X}_n; \theta) = \prod_{i=1}^{n} \mathcal{L}(\mathcal{X}_i; \theta). \tag{C.28}$$

La méthode du maximum de vraisemblance consiste à choisir comme estimateur de θ la valeur θ^* qui maximise la vraisemblance. C'est donc la valeur du paramètre qui fait paraître l'échantillon comme le plus probable.

Appliquons ce principe à l'estimation de la moyenne μ d'une loi $p_\mu(\mathcal{X})$. Prenons le logarithme de (C.28) :

$$\log \mathcal{L}(x_1, \ldots, x_n; \mu) = \sum_{i=1}^{n} \log p_\mu(x_i), \tag{C.29}$$

qui doit être maximale. En prenant la dérivée par rapport à μ on obtient une équation en μ

$$\sum_{i=1}^{n} \frac{\partial \log p_\mu(x_i)}{\partial \mu} = 0, \tag{C.30}$$

dont la solution (pas nécessairement unique) fournit un estimateur dit au maximum de vraisemblance.

Suggestion : montrer que si la loi de $\mathcal{X}$ est une loi Gaussienne, l'estimateur au maximum de vraisemblance de l'espérance mathématique μ est $\overline{\mathcal{X}}$. Peut-on en déduire des conditions plus générales sous lesquelles $\overline{\mathcal{X}}$ est encore une estimée au maximum de vraisemblance ? Chercher un contre-exemple de loi où $\overline{\mathcal{X}}$, a contrario, n'est pas une estimée au maximum de vraisemblance.

Notons, pour finir qu'un estimateur au maximum de vraisemblance n'est pas nécessairement non-biaisé. Cependant, on montre qu'il est convergent et asymptotiquement sans biais et efficace.

Relation entre l'estimation au maximum de vraisemblance et l'estimation Bayesienne. Notons tout d'abord que la vraisemblance de l'échantillon est identique à la probabilité conditionnelle de celui-ci étant donné la valeur de θ. On peut donc écrire pour l'estimation Bayesienne

$$p(\theta | x_1, \ldots, x_n) = \frac{p(\theta) \mathcal{L}(x_1, \ldots, x_n; \theta)}{\int_\Theta \mathcal{L}(x_1, \ldots, x_n; \theta) p(\theta) d\theta}. \tag{C.31}$$

Il s'en suit que si $p(\theta)$ est indépendante de θ (c'est-à-dire uniforme sur Θ) on peut écrire

$$p(\theta|x_1,\ldots,x_n) = \frac{\mathcal{L}(x_1,\ldots,x_n;\theta)}{h(x_1,\ldots,x_n)} \tag{C.32}$$

où $h(x_1,\ldots,x_n) = \int_\Theta \mathcal{L}(x_1,\ldots,x_n;\theta)d\theta$ est indépendant de θ. En d'autres termes, la valeur de θ qui maximise la vraisemblance est aussi celle qui maximise la probabilité conditionnelle $p(\theta|x_1,\ldots,x_n)$.

Si $p(\theta)$ n'est pas uniforme, alors, dans la démarche Bayesienne, elle a pour effet de renforcer la probabilité a posteriori des valeurs de θ les plus probables a priori.

Voyons comment cet effet se manifeste lorsque n varie. Prenons le logarithme des deux membres de (C.31) et supposons que les échantillons sont i.i.d. selon la loi $p_\theta(\mathcal{X})$. On a

$$\frac{1}{n}\log p(\theta|x_1,\ldots,x_n) \propto \frac{1}{n}\log p(\theta) + \frac{1}{n}\sum_{i=1}^{n}\log p_\theta(x_i), \tag{C.33}$$

où nous n'avons retenu que les termes qui dépendent de θ et divisé par la taille de l'échantillon n. Lorsque $n \to \infty$ le premier terme du membre de droite tend vers zéro. Le second terme tend, si la limite existe, vers la grandeur suivante

$$E\{\log p_\theta(\mathcal{X})\} = \int p_{\theta^*}(\mathcal{X})\log p_\theta(\mathcal{X})dx \tag{C.34}$$

si $\mathcal{X}$ est distribuée selon la loi (densité) p_{θ^*}.

Nous verrons au cours de théorie de l'information que la grandeur suivante

$$D(p_{\theta^*}||p_\theta) \triangleq \int p_{\theta^*}(\mathcal{X})\log\frac{p_{\theta^*}(\mathcal{X})}{p_\theta(\mathcal{X})}dx \tag{C.35}$$

désigne la distance de Kullback-Leibler entre les deux distributions. Si cette grandeur existe, elle est positive, et nulle si, et seulement si $p_{\theta^*}(\mathcal{X}) = p_\theta(\mathcal{X})$, autrement dit si, et seulement si $\theta = \theta^*$. On peut écrire (C.34) en fonction de cette grandeur

$$E\{\log p_\theta(\mathcal{X})\} = -D(p_{\theta^*}||p_\theta) - H_d(\mathcal{X}) \tag{C.36}$$

où

$$H_d(\mathcal{X}) \triangleq -\int p_{\theta^*}(\mathcal{X})\log p_{\theta^*}(\mathcal{X})dx \tag{C.37}$$

désigne l'entropie différentielle de la v.a. $\mathcal{X}$ (voir également cours de théorie de l'information).

On peut donc tirer les conclusions suivantes : si les distributions de probabilités indicées par θ admettent des entropies, alors pour n suffisamment grand on aura $\forall\theta$ tel que $p(\theta) \neq 0$

$$\frac{1}{n}\log p(\theta|x_1,\ldots,x_n) \propto -D(p_{\theta^*}||p_\theta) - H_d(\mathcal{X}) + \epsilon \tag{C.38}$$

avec $\epsilon \to 0$ lorsque $n \to \infty$. Cela veut dire que lorsque la taille des échantillons devient infinie, et pour autant que $p(\theta) > 0$ sur Θ, la valeur de θ^* qui maximise la probabilité a posteriori tend vers θ^*, indépendamment de la forme de la distribution $p(\theta)$. La solution la plus probable au sens Bayesien devient donc identique à la solution au maximum de vraisemblance. De plus, en prenant en compte le comportement asymptotique du dénominateur de (C.31), on peut se convaincre que la loi $p(\theta|x_1,\ldots,x_n)$ se concentre de plus en plus fortement autour de son maximum.

Asymptotiquement, les approches Bayesiennes et du maximum de vraisemblance tombent donc d'accord. On en déduit que l'approche Bayesienne est surtout intéressante lorsque n est petit, car elle permet de prendre en compte de façon cohérente l'information a priori disponible sur θ, information dont l'impact sur le résultat sera d'autant plus forte que l'échantillon est petit.

C.4.3 Minimisation du risque

Pour terminer nous allons brièvement décrire la méthode de minimisation du risque qui est de plus en plus utilisée notamment dans les problèmes de régression.

Dans le cas scalaire, elle consiste à définir une fonction de risque $R(\mathcal{X}, \theta)$ qui mesure l'écart entre la variable aléatoire scalaire et un paramètre censé estimer la valeur la plus réaliste de celle-ci. Vue sous cet angle, l'estimation revient alors à choisir le paramètre θ tel que l'espérance du risque

$$E\{R(\mathcal{X}, \theta)\} \tag{C.39}$$

soit minimale. Une méthode très générale pour atteindre cet objectif à partir d'un échantillon consiste à choisir la valeur du paramètre qui minimise le risque empirique, c'est-à-dire la valeur T_n^* de θ qui minimise

$$\frac{1}{n}\sum_{i=1}^{n} R(x_i, \theta). \tag{C.40}$$

On montre, sous des hypothèses très générales ([Vap95]) que cette méthode est consistente, c'est-à-dire que lorsque $n \to \infty$, $T_n^* \to \theta^*$, où θ^* est la valeur qui minimise l'espérance du risque (C.39).

Suggestion : montrer que si le risque est défini par l'écart quadratique $R(\mathcal{X}, \theta) = (\mathcal{X} - \theta)^2$ alors θ^* est égal à $E\{\mathcal{X}\}$ et $T_n^* = \overline{\mathcal{X}}$.

C.4.4 Robustesse

Dans la démarche classique, mise en évidence dans ce qui précède, on fait des hypothèses sur la forme des lois et sur le mécanisme de production des échantillons, et à partir de ces hypothèses on déduit des estimateurs "optimaux" du point de vue de certaines de leurs qualités.

Un des problèmes majeurs de ces estimateurs est qu'ils sont souvent très sensibles aux écarts vis-à-vis des hypothèses de départ : par exemple si la loi s'écarte de celle supposée ou bien si certaines observations sont aberrantes (c'est-à-dire ne proviennent pas de cette loi), les estimateurs peuvent donner des résultats très peu précis.

La statistique robuste, que nous n'aborderons pas dans cet appendice vise explicitement au développement d'estimateurs qui sont peu sensibles à ce type d'écarts, c'est-à-dire des estimateurs qui continueront à donner des estimées assez précises si on s'écarte des hypothèses. Evidemment, cette robustesse est en général obtenue au prix d'une moins bonne précision dans la cas où les hypothèses sont vérifiées ("no free lunch theorem").

C.5 RE-ECHANTILLONNAGE, SONDAGE, ET SIMULATION

Dans ce qui précède, nous avons supposé que l'échantillon était donné, et discuté les différentes manières d'exploiter un tel échantillon. Dans cette section nous abordons très brièvement le très vaste domaine relatif à la constitution des échantillons. Nous nous limiterons essentiellement à une brève description informelle des différentes techniques en renvoyant le lecteur intéressé à des ouvrages spécialisés.

C.5.1 Ré-échantillonnage

Le ré-échantillonnage consiste à construire à partir d'un échantillon donné un certain nombre de nouveaux échantillons obtenus par tirage aléatoire. Ces techniques sont particulièrement utiles pour étudier le biais et la variance des estimateurs et dans certains cas permettent de les réduire. Ces techniques, assez intensives du point de vue des calculs sont devenues très populaires récemment, suite aux développements informatiques. Elles ont cependant été découvertes il y a de nombreuses années.

Bootstrap. Soit $\mathcal{T}$ un estimateur. Si on connaît la loi $F(x)$ on peut en principe calculer analytiquement ou numériquement sa distribution d'échantillonnage. Si on dispose d'une bonne approximation de $F(x)$, disons $F_n^*(x)$ obtenue à partir d'un échantillon de taille n suffisamment grande on peut obtenir une distribution d'échantillonnage approchée en utilisant $F_n^*(x)$ à la place.

En particulier, la méthode de *bootstrap* ("chausse-pied") consiste à sonder $F_n^*(x)$, en tirant k échantillons de taille n avec remise, à partir de l'échantillon de départ, pour chacun desquels on calcule la valeur de T_i. Cette procédure donne donc une distribution empirique, approximation de la loi de $F_{\mathcal{T}_n}$.

En particulier, si $\mathcal{T}$ est un estimateur non-biaisé de θ mais de variance assez élevée, la moyenne empirique des estimées de Bootstrap $\overline{T_i}$ fournit encore un estimateur non-biaisé de θ, mais de variance plus faible.

Jacknife. Le but ici est de diminuer le biais d'un estimateur. Soit un estimateur biaisé et un échantillon de taille n. Soit T_n la valeur de l'estimateur sur l'échantillon complet, et soit T_{-i} la valeur obtenue en utilisant ce même échantillon à l'exception de la i-ème observation. Désignons par

$$T_i^* = nT_n - (n-1)T_{-i}$$

et formons le nouvel estimateur

$$T_n^J = \frac{1}{n}\sum_{i=1}^{n} T_i^*.$$

On montre que sous certaines conditions ce nouvel estimateur est moins biaisé que $\mathcal{T}$. En particulier, on montre que si

$$E\{\mathcal{T}\} = \theta + \frac{a}{n}$$

alors $E\{\mathcal{T}_J\} = \theta$. Dans ce cas particulier l'estimateur de Jacknife est donc non-biaisé.

Suggestion : à titre d'exercice on peut se convaincre que l'estimateur de Jacknife de la variance appliqué à $\mathcal{S}^2$ donne $\mathcal{S}^2_{n-1}$ qui est non biaisé.

C.5.2 Sondage

Le but des méthodes de sondage est de constituer des échantillons avec comme objectif l'utilisation ultérieure de l'échantillon pour l'estimation. Plusieurs raisons conduisent à l'utilisation de ces méthodes.

1. Le sondage peut être appliqué à un échantillon existant de taille trop importante pour être traité, dans le but d'en extraire un échantillon de taille acceptable. C'est fréquemment le cas dans le domaine de l'analyse de grandes bases de données (plusieurs millions d'observations, parfois).
2. Un autre raison d'appliquer le sondage est simplement de constituer un échantillon. Par exemple, dans le domaine du contrôle de qualité en fabrication, on prélève régulièrement des pièces à différents stades du processus, puis on les examine et on analyse les performances moyennes.
3. Une troisième raison d'appliquer le sondage est de pouvoir influencer la manière dont les observations sont sélectionnées, notamment dans le but de permettre des estimations aussi précises que possible avec un minimum d'observations. Ceci est par exemple le cas dans les enquêtes d'opinion et les sondages électoraux, où il est nécessaire pour des raisons de coût de limiter le nombre de personnes interrogées.

Nous allons simplement expliquer une méthode classique de sondage aléatoire qui repose sur la notion de stratification de la population.

Sondage stratifié. Soit $\mathcal{F}$ la grandeur dont nous souhaitons estimer l'espérance.

L'approche est basée sur la connaissance a priori d'une partition de l'espace Ω en régions où la variance de $\mathcal{F}$ est relativement faible.

Soit en effet $\{\Omega_1, \ldots, \Omega_K\}$ une telle partition, c'est-à-dire telle que

$$\forall\, i \neq j \;:\; \Omega_i \cap \Omega_j = \emptyset, \tag{C.41}$$

et

$$\Omega = \bigcup_{i=1}^{K} \Omega_i. \tag{C.42}$$

Supposons qu'a priori on connaisse les probabilités $P(\Omega_i)$ et désignons par $E_{P_i}(\mathcal{F})$ l'espérance conditionnelle de $\mathcal{F}$ dans la classe Ω_i, et par $\sigma^2_{i,\mathcal{F}}$ la variance conditionnelle de $\mathcal{F}$.

Notons que les $E_{P_i}(\mathcal{F})$ peuvent être estimées par l'analyse d'observations indépendantes obtenues pour chacun des sous-ensembles Ω_i de la population. Ayant estimé les grandeurs $E_{P_i}\{\mathcal{F}\}$, on peut alors reconstruire une estimée de $E\{\mathcal{F}\}$ au moyen de la formule suivante

$$\hat{E}\{\mathcal{F}\} = \sum_{i=1}^{K} P(\Omega_i)\hat{E}_{P_i}\{\mathcal{F}\}. \tag{C.43}$$

Comme les termes de cette somme sont estimés de manière indépendante, la variance de l'estimée reconstituée est obtenue par la formule suivante :

$$\sigma^2_{\hat{E}\{\mathcal{F}\}} = \sum_{i=1}^{K} P^2(\Omega_i)\sigma^2_{\hat{E}_{P_i}\{\mathcal{F}\}}. \tag{C.44}$$

La question qui se pose alors, est d'allouer de façon optimale un nombre total N d'observations aux différentes régions Ω_i, c'est-à-dire de manière à minimiser la variance $\sigma^2_{\hat{E}\{\mathcal{F}\}}$.

Par ailleurs, si on utilise N_i observations pour estimer les $E_{P_i}(F)$, alors

$$\sigma^2_{\hat{E}_{P_i}\{\mathcal{F}\}} = N_i^{-1}\sigma^2_{i,\mathcal{F}}. \tag{C.45}$$

On peut montrer que dans ces conditions l'optimum est réalisé par le schéma d'allocation suivant

$$N_i = N\frac{P(\Omega_i)\sigma_{i,\mathcal{F}}}{\sum_{i=1}^{K} P(\Omega_i)\sigma_{i,\mathcal{F}}}, \tag{C.46}$$

et que la variance de l'estimée $\hat{E}_P(F)$ vaut alors

$$\sigma^2_{\hat{E}_P\{\mathcal{F}\}} = N^{-1}\left(\sum_{i=1}^{K} P(\Omega_i)\sigma_{i,\mathcal{F}}\right)^2. \tag{C.47}$$

Remarques. A. Si les K variances conditionnelles sont uniformes (disons égales à $\sigma^2_{K,F}$), ce schéma consiste à allouer à chaque région un nombre d'observations proportionnel à sa probabilité a priori. L'équation (C.47) devient alors

$$\sigma^2_{\hat{E}_P(F)} = N^{-1}\sigma^2_{K,F}, \tag{C.48}$$

ce qui veut dire que le gain est exprimé par le rapport

$$\text{Gain} = \frac{\sigma^2_F}{\sigma^2_{K,F}}. \tag{C.49}$$

B. Signalons enfin que même si les variances conditionnelles ne sont pas uniformes, le schéma d'échantillonnage qui consiste à allouer un nombre d'échantillons à chaque région proportionnel à sa probabilité n'est jamais moins bon que la méthode de base de sondage non stratifié.

C. En corrolaire, le schéma d'allocation optimal conduit toujours à une variance inférieure ou égale à celle de la méthode de base. La méthode sera d'autant plus efficace que le second membre de l'équation (C.47) sera faible. Tout l'art consiste donc à identifier une partition de Ω en régions à faible variance. Nous verrons au cours d'*Apprentissage Inductif Appliqué* que c'est ce qui est fait par les méthodes de construction d'arbres de régression.

C.5.3 Méthode de Monte-Carlo

Le problème posé est le suivant. Etant donné un espace d'entrée Ω sur lequel une mesure de probabilité P est définie, et étant donné une fonction $\boldsymbol{f}(\cdot)$ définie sur Ω à valeurs dans $\mathbb{R}^n$, il s'agit d'estimer l'espérance mathématique

$$E_P(\boldsymbol{F}) = \int_{\Omega} \boldsymbol{f}(\omega)dP(\omega), \tag{C.50}$$

où $\boldsymbol{F}$ désigne la variable aléatoire vectorielle induite par la fonction $\boldsymbol{f}(\cdot)$.

Afin d'alléger les notations, nous allons dans la suite restreindre notre discussion au cas d'une variable aléatoire scalaire. Nous utiliserons les notations f et F (resp. $\boldsymbol{f}$ et $\boldsymbol{F}$) pour représenter une variable aléatoire scalaire (resp. vectorielle).

Remarquons que dans le cas qui nous intéresse, l'espace (Ω, P) est de structure mixte continue/discrète, ce qui justifie les notations un peu lourdes que nous utilisons. D'un point de vue pratique (simulation sur ordinateur) l'espace est évidemment discret et fini, même si sa taille est énorme.

Notons également que pour la suite il est nécessaire de supposer que les deux premiers moments de F sont finis, ce qui est le cas en pratique puisque les valeurs de f sont bornées (les grandeurs physiques qui nous intéressent sont en effet bornées).

Dans ce problème, la paire (Ω, P) constitue le modèle stochastique et $\boldsymbol{f}$ le modèle physique. Tous les deux sont en pratique des approximations plus ou moins bonnes de la réalité.

C.5.3.1 Méthode de Monte-Carlo de base. Lorsque la fonction $f(\cdot)$ possède une structure simple (p.ex. linéarité), il est possible de calculer l'espérance mathématique par voie analytique (ainsi que les autres caractéristiques de la distribution de F). Lorsque ce n'est pas le cas, on a recours à la méthode de Monte-Carlo.

La méthode de Monte-Carlo de base consiste à générer un échantillon aléatoire d'observations $\omega_1, \ldots, \omega_N \in \Omega$ indépendants et distribués selon P, puis à calculer les valeurs correspondantes $f(\omega_i)$, et enfin à approximer l'espérance mathématique de F par la moyenne d'échantillon

$$\mu_F^{(N)} = N^{-1} \sum_{i=1}^{N} f(\omega_i). \tag{C.51}$$

Le théorème central limite assure que cette estimée est non-biaisée, et que sa distribution d'échantillonnage est asymptotiquement Gaussienne de variance

$$\sigma^2_{\mu_F^{(N)}} = N^{-1} \sigma_F^2. \tag{C.52}$$

Au bruit aléatoire près, la précision de la méthode croît donc de façon monotone mais assez lente, en fonction du nombre d'échantillons.

En pratique on utilise une version itérative, qui génère une suite d'échantillons ω_k de longueur non fixée a priori, et calcule à chaque pas les valeurs de $\mu_F^{(k)}$ et $\sigma^2_{\mu_F^{(k)}}$. Le process est arrêté lorsque $\sigma^2_{\mu_F^{(k)}} \leq \epsilon$, où ϵ est un seuil de précision fixé à priori en fonction du niveau de précision recherché.

Notons que des formules de mise à jour permettent de calculer les valeurs de $\mu_F^{(k)}$ et $\sigma^2_{\mu_F^{(k)}}$ efficacement.

Le seuil de précision est normalement fixé en fonction de critères pratiques, et la précision intrinsèque du modèle utilisé doit évidemment être prise en compte pour fixer ϵ.

Notons qu'en général l'exploitation du modèle stochastique est peu consommatrice en temps CPU, et c'est le calcul de la fonction f qui est en général la partie contraignante.

Note. *Dans certains cas pratiques ce n'est pas la simulation mais bien l'expérimentation physique qui permet d'observer les valeurs de la fonction $f(\cdot)$.*

L'expérimention physique pouvant être coûteuse en temps et en moyens, certaines méthodes de réduction de la variance gardent leur intérêt dans la mesure où elles peuvent permettre de réduire les coûts d'expérimenation. Par exemple, l'utilisation de modèles de simulation sur ordinateur permet dans ce cas de troquer une partie de l'effort d'expérimenation physique pour de la simulation numérique.

C.5.3.2 Réduction de la variance. La réduction de la variance vise à augmenter la précision pour un nombre donné d'échantillons simulés. En d'autres termes, elle vise à réduire le nombre de simulations nécessaires pour obtenir le niveau de précision recherché.

L'hypothèse de départ sous-jacente aux méthodes de réduction de la variance, c'est qu'on dispose d'informations partielles a priori sur la variable aléatoire F. Si on ne dispose d'aucune information a priori aucune réduction de la variance n'est possible; si on dispose de la connaissance totale de F a priori, aucune simulation de Monte-Carlo n'est nécessaire. Entre ces deux extrêmes, les méthodes de réduction de la variance peuvent opérer, de manière plus ou moins efficace en fonction de la nature des informations dont on dispose et de particularités du problème.

L'information a priori peut être obtenue de diverses façons : construction par voie analytique de modèles approchés se prêtant au traitement analytique; simulation du processus et extraction de connaissances à l'aide de techniques d'analyse de données; le bon sens physique et l'expérience humaine.

Quoiqu'il en soit, dans l'approche de Monte-Carlo les simulations successives peuvent être exploîtées pour accroître l'expérience qui devient alors de plus en plus riche. Cette expérience peut alors être exploîtée pour réduire progressivement le nombre de simulations nécessaires dans le processus d'estimation.

Tout d'abord, décrivons deux méthodes classiques de réduction de la variance : la première nécessite une modification de la stratégie de tirage aléatoire; pas la seconde.

C.5.3.3 Echantillonnage stratifié.

L'échantillonnage stratifié décrit précédemment peut évidemment être appliqué pour réduire le nombre de simulations nécessaires, ou de manière équivalente, réduire la variance des estimées pour un même nombre de simulations.

C.5.3.4 Variables de contrôle.

Cette approche ne nécessite pas de modification du schéma d'échantillonnage. Elle peut donc en principe se combiner avec la méthode précédente.

L'idée principale est de se servir d'une fonction auxiliaire $g(\cdot)$ (appelée variable de contrôle) fortement corrélée à $f(\cdot)$ et telle que $E_P\{\mathcal{G}\}$ soit facile à estimer. En d'autres termes $g(\cdot)$ doit être une bonne approximation de $f(\cdot)$, à une transformation linéaire près.

Dans ce cas, quel que soit $\beta \in \mathbb{R}$ la variable aléatoire

$$\mathcal{H} = \mathcal{F} - \beta(\mathcal{G} - E_P\{\mathcal{G}\}) \tag{C.53}$$

possède la même espérance mathématique que $\mathcal{F}$, et on peut donc estimer $E_P\{\mathcal{F}\}$ en appliquant la technique de Monte-Carlo de base à $\mathcal{H}$.

Selon la valeur choisie pour β la variance de cette estimée sera plus ou moins importante, et on montre que la valeur optimale de β^* qui minimise la variance est

$$\beta^* = \frac{\text{Cov}(\mathcal{F}, \mathcal{G})}{\sigma^2_{\mathcal{G}}}. \tag{C.54}$$

La variance de $\mathcal{H}$ vaut alors

$$\sigma^2_{\mathcal{H}} = (1 - \rho^2(\mathcal{F}, \mathcal{G}))\sigma^2_{\mathcal{F}}. \tag{C.55}$$

Si $g(\cdot)$ est une très bonne approximation de $f(\cdot)$, on aura $\beta^* \approx 1$ et $\rho^2(\mathcal{F}, \mathcal{G}) \approx 1$. Intuitivement, si $g(\cdot)$ est une bonne approximation de $f(\cdot)$, la différence entre $g(\cdot)$ et $f(\cdot)$ est petite dans les régions les plus probables de Ω, et donc la variance de cette différence doit également être petite.

En pratique, pour qu'on puisse facilement calculer $E_P\{\mathcal{G}\}$, il suffit que $g(\cdot)$ soit facile à calculer et on peut alors utiliser la méthode de Monte-Carlo de base avec un nombre d'échantillons très grand pour estimer $E_P\{\mathcal{G}\}$ et $\sigma^2_{\mathcal{G}}$. La difficulté réside dans le calcul (l'estimation) de β^* dont le terme relatif à la covariance fait appel au calcul de $f(\cdot)$.

En pratique on peut cependant soit postuler la valeur a priori (au risque de se tromper), soit estimer grossièrement la covariance dans une phase de rodage.

C.6 REGRESSION ET MODELES LINEAIRES

C.6.1 Régression simple

Etant donné deux v.a. $\mathcal{X}$ et $\mathcal{Y}$, la recherche d'une fonction $f(\cdot)$ telle que $f(\mathcal{X})$ soit aussi proche que possible de $\mathcal{Y}$ en moyenne quadratique à déjà été abordée dans l'appendice B. Nous savons la $f(\mathcal{X}) = E\{\mathcal{Y}|\mathcal{X}\}$ réalise le minimum de $E\{(\mathcal{Y} - f(\mathcal{X}))^2\}$.

Nous avons donc ici à faire à un schéma d'estimation fonctionnelle, puisque nous souhaitons trouver une fonction $f(\mathcal{X})$ qui fournisse une bonne approximation en moyenne quadratique de $E\{\mathcal{Y}|\mathcal{X}\}$.

La démarche statistique classique consiste alors à introduire des hypothèses sur la distribution conjointe de $(\mathcal{X}, \mathcal{Y})$ et à en déduire des estimateurs optimaux. Nous n'allons pas insister ici sur cette approche; nous allons

simplement rappeler le fait, déjà mis évidence dans l'appendice B, que si $(\mathcal{X}, \mathcal{Y})$ sont conjointenement Gaussiennes, alors $E\{\mathcal{Y}|\mathcal{X}\}$ est une fonction linéaire de $\mathcal{X}$ et de plus la variance conditionnelle est alors indépendante de $\mathcal{X}$. Cependant, il ne s'agit de toute évidence pas d'un condition nécessaire, car il suffit que l'on puisse écrire

$$\mathcal{Y} = \alpha + \beta\mathcal{X} + \epsilon(\mathcal{X}), \tag{C.56}$$

où $\epsilon(\mathcal{X})$ est une variable aléatoire centrée (de moyenne nulle) pour presque tout $\mathcal{X}$, pour que l'espérance conditionnelle de $\mathcal{Y}$ soit une fonction linéaire de $\mathcal{X}$.

La méthode des moindres carrés consiste à choisir, étant donné un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$, les valeurs β^* et α^* qui minimisent l'erreur quadratique totale d'estimation sur cet échantillon, i.e. telles que

$$\text{MSE}(\alpha^*, \beta^*) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\alpha^* + \beta^* x_i))^2, \tag{C.57}$$

soit minimal. Nous renvoyons le lecteur aux ouvrages de statistique de base (p.ex. [Sap90]) pour une discussion détaillée de cette méthode.

C.6.2 Régression multiple

Le problème de la régression multiple s'obtient par généralisation du précédent en utilisant p variables d'entrée au lieu d'une seule. Soit un échantillon de n observations $\omega_1, \ldots, \omega_n$, caractérisées par p valeurs $x_1(\omega_i), \ldots, x_p(\omega_i)$ et une grandeur de sortie $y(\omega_i)$. Le but est alors d'estimer, au moyen d'une fonction linéaire, l'éspérance conditionnelle

$$E\{\mathcal{Y}|\mathcal{X}_1, \ldots, \mathcal{X}_p\} \approx b_0 + \sum_{i=1}^{p} b_i\mathcal{X}_i. \tag{C.58}$$

A nouveau, la méthode des moindres carrés consiste à choisir les valeurs de $b_0^*, b_1^*, \ldots, b_p^*$ de telle manière à ce que

$$\text{MSE}(b_0, \ldots, b_p) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + \sum_{i=1}^{p} b_i x_i))^2, \tag{C.59}$$

soit minimale. Cette fonction étant une fonction quadratique en termes des paramètres b_i, son gradient (par rapport aux b_i) est une fonction linéaire de ces paramètres. L'annulation de ce gradient est une condition nécessaire et suffisante d'optimalité, puisque la fonction est au moins semi-définie positive (et donc au moins convexe). La condition d'annulation du gradient donne lieu à un système de $p+1$ équations linéaires dont les $p+1$ variables sont les b_i. Si ce système est non singulier, sa solution est unique et identique à la solution au sens de moindres carrés. Sinon, il y aura un ensemble de solutions optimales (équivalentes) qui forme un sous-espace linéaire de $\mathbb{R}^{p+1}$.

On montre que pour que la solution au sens des moindres carrées soit unique, il faut (et il suffit) que les $p+1$ vecteurs colonnes

$$\boldsymbol{x}_0 == \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \text{ et } \boldsymbol{x}_i = \begin{bmatrix} x_i(\omega_1) \\ x_i(\omega_2) \\ \vdots \\ x_i(\omega_n) \end{bmatrix}$$

soient linéairement indépendants, ce qui implique que les p variables $\mathcal{X}_i$ soient non-constantes et linéairement indépendantes, et en nombre strictement inférieur à n.

Comme ci-dessus, nous renvoyons le lecteur intéressé aux ouvrages classiques de statistique pour plus de détails concernant cette approche. On peut évidemment justifier l'utilisation du modèle linéaire multiple de la même manière que ci-dessus pour des variables $(\mathcal{Y}, \mathcal{X}_1, \ldots, \mathcal{X}_p)$ conjointement Gaussiennes ou, plus généralement, une situation où $E\{\mathcal{Y}|\mathcal{X}_1, \ldots, \mathcal{X}_p\}$ est effectivement une fonction linéaire.

Notes

1. Nous attirons l'attention du lecteur sur le fait qu'il y a en réalité une différence entre la notion d'efficacité, telle que définie par les statisticiens, et la notion de variance minimale.

D CALCUL VECTORIEL ET MATRICIEL

D.1 INTRODUCTION

Cet appendice collationne une certain nombre de résultats utiles (voire indispensables) dans les cadre des différents enseignements faisant appel aux méthodes stochastiques.

La plupart des questions couvertes sont en principe connues par les étudiants ingénieurs. Nous avons donc volontairement adopté un style très synthétique, de manière à encourager les étudiants qui auraient oublié ces notions à consulter leurs notes de cours de candidature.

Nous commençons par un bref rappel sur les espaces linéaires euclidiens de dimension finie quelconque. Nous donnons ensuite quelques définitions et propriétés relatives aux fonctions convexes définies sur de tels espaces.

Ensuite, nous couvrons quelques éléments relatifs au calcul matriciel, en particulier les propriétés des matrices réelles symétriques et définies positives. Nous avons inclus dans cette partie une discussion des matrices de Toeplitz, fréquemment rencontrées dans les applications au traitement de signal statistique.

D.2 ESPACES EUCLIDIENS

La plupart des notions introduites ici sur $\mathbb{R}^p$ peuvent être généralisées au cas de vecteurs de nombres complexes. Nous en fournirons à l'appendice E une généralisation plus large encore.

D.2.1 Définitions

On rappelle que $\forall p \geq 1$, $\mathbb{R}^p$ désigne l'ensemble des p-uples ordonnés de nombres réels. On convient de représenter ces p-uples par des vecteurs colonnes de dimension p, qui seront désignés par une lettre minuscule grasse (e.g. $\boldsymbol{x}$), avec éventuellement en exposant la dimension p de l'espace, si nécessaire (e.g. $\boldsymbol{x}^p$). Nous utiliserons indifféremment le terme point ou vecteur de $\mathbb{R}^p$.

On désigne par $\mathbf{0}$ le vecteur nul, dont toutes les composantes sont nulles.

Soient $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^p$ trois vecteurs. Nous désignerons par $x_i, y_i, z_i \in \mathbb{R}$ leur i-ème composante. Deux vecteurs sont égaux (par définition) si toutes leurs composantes sont égales.

On définit alors, à partir du produit de nombres réels, la multiplication d'un vecteur par un nombre réel de la manière suivante

$$\lambda \boldsymbol{x} = \boldsymbol{y} \Leftrightarrow \forall i = 1, \ldots, p \ : y_i = \lambda x_i, \tag{D.1}$$

et, à partir de l'addition de nombres réels, l'addition de deux vecteurs par

$$\boldsymbol{x}+\boldsymbol{y} = \boldsymbol{z} \Leftrightarrow \forall i = 1, \ldots, p \ : z_i = x_i + y_i. \tag{D.2}$$

On a

$$\mathbf{0} = 0\boldsymbol{x}, \forall \boldsymbol{x} \in \mathbb{R}^p, \tag{D.3}$$

et

$$\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}, \forall \boldsymbol{x} \in \mathbb{R}^p. \tag{D.4}$$

Plus généralement, on définit la notion de combinaison linéaire de n vecteurs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ par

$$\boldsymbol{z} = \sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i, \tag{D.5}$$

où les λ_i sont des nombres réels quelconques.

On désigne par $\boldsymbol{e}_i, \forall i = 1, \ldots, p$, le i-ème vecteur *unitaire* défini par $e_{i,j} = \delta_{i,j}$, c'est-à-dire dont toutes les composantes sont nulles sauf la i-ème qui vaut 1. On a, remarquablement,

$$\boldsymbol{x} = \sum_{i=1}^{p} x_i \boldsymbol{e}_i. \tag{D.6}$$

D.2.2 Produit scalaire, norme et distance

On définit le *produit scalaire* de deux vecteurs $\boldsymbol{x}, \boldsymbol{y}$ le nombre réel noté $\boldsymbol{x}^T\boldsymbol{y}$ et défini par

$$\boldsymbol{x}^T\boldsymbol{y} \triangleq \sum_{i=1}^{p} x_i y_i. \tag{D.7}$$

Nous verrons à la section suivante que dans cette notation, héritée du calcul matriciel, $\boldsymbol{x}^T$ désigne le vecteur transposé, et que le produit scalaire est un cas particulier du produit de deux matrices.

Le produit scalaire est commutatif par définition. Il est également distributif par rapport aux combinaisons linéaires. Evidemment, on a

$$\mathbf{0}^T\boldsymbol{x} = \boldsymbol{x}^T\mathbf{0} = 0. \tag{D.8}$$

Par ailleurs, si deux vecteurs non nuls sont tels que leur produit scalaire est nul, on dit qu'ils sont orthogonaux.

Nous verrons à la section suivante que le produit scalaire peut-être généralisé à une forme bi-linéaire du type $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}$ où $\boldsymbol{A}$ est une matrice $p \times p$ symétrique et définie positive.

On appelle module, ou norme euclidienne d'un vecteur, le nombre réel positif ou nul

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}} = \sqrt{\sum_{i=1}^{p} x_i^2}. \tag{D.9}$$

On a

$$\|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = \mathbf{0}, \tag{D.10}$$

$$\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|, \tag{D.11}$$

et, l'inégalité triangulaire

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|, \tag{D.12}$$

et l'inégalité de Schwarz

$$|\boldsymbol{x}^T\boldsymbol{y}| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|. \tag{D.13}$$

On dit que la norme euclidienne est *induite* par le produit scalaire. On peut généraliser cette notion à la norme d'ordre m par

$$\|\boldsymbol{x}\|_m = \sqrt[m]{\sum_{i=1}^{p} |x_i|^m}, \tag{D.14}$$

$\forall m \geq 1$. En particulier on a,

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{p} |x_i|, \tag{D.15}$$

qu'on appelle aussi norme de Manhattan, et

$$\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|, \tag{D.16}$$

la norme euclidienne, et

$$\|\boldsymbol{x}\|_\infty = \lim_{m\to\infty} \sqrt[m]{\sum_{i=1}^{p} |x_i|^m} = \max_{i=1,\ldots,p} |x_i|, \tag{D.17}$$

qu'on appelle la norme "infinie".

On définit la distance (induite par une certaine norme) entre deux vecteurs $\boldsymbol{x}, \boldsymbol{y}$, par la norme du vecteur $\boldsymbol{x}-\boldsymbol{y}$:

$$d_m(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\triangle}{=} |\boldsymbol{x} - \boldsymbol{y}|_m. \tag{D.18}$$

On rappelle qu'une distance, pour mériter ce nom doit satisfaire aux propriétés suivantes :

1. $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$
2. $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$
4. $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y}), \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^p$.

On appelle distance euclidienne la distance induite par la norme euclidienne.

Généralisation aux vecteurs complexes. Nous généralisons le produit scalaire aux vecteurs de nombres complexes par

$$\boldsymbol{x}^* \boldsymbol{y} \stackrel{\triangle}{=} \sum_{i=1}^{n} \overline{x_i} y_i, \tag{D.19}$$

de cette façon la norme euclidienne d'un vecteur complexe reste définie par le produit scalaire de ce vecteur par lui-même et deux vecteurs sont orthogonaux si leur produit scalaire est nul. La notation * est, elle aussi, héritée du calcul matriciel.

Nous attirons l'attention du lecteur sur le fait que certains auteurs adoptent la définition suivante pour le produit scalaire

$$\boldsymbol{y}^* \boldsymbol{x} \stackrel{\triangle}{=} \sum_{i=1}^{n} \overline{y_i} x_i, \tag{D.20}$$

ce qui ne change cependant rien au niveau des principes. La définition (D.19) que nous adoptons ici est celle qui est utilisée dans la plupart de livres de physique.

Notons également que dans l'annexe E nous défnirons le produit scalaire sur des espaces linéaires plus généraux que $\boldsymbol{C}^p$ ou $\mathbb{R}^p$. Nous utiliserons alors la notation $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ pour désigner celui-ci.

D.2.3 Dépendance et indépendance linéaire

On dit que les vecteurs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ sont linéairement indépendants si

$$\boldsymbol{0} = \sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i \Rightarrow \lambda_i = 0, \forall i = 1, \ldots, n; \tag{D.21}$$

sinon ils sont dits linéairement dépendants. Dans ce cas, il existe un ensemble de valeurs de λ_i non toutes nulles, telles que la combinaison linéaire soit le vecteur nul.

On se convaincra facilement que les vecteurs $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ sont linéairement indépendants.

Tout ensemble de vecteurs qui comprend le vecteur nul est linéairement dépendant. Tout ensemble de vecteurs qui comprend un sous-ensemble linéairement dépendant est linéairement dépendant. Si des vecteurs sont linéairement indépendants et deviennent linéairement dépendants lorsqu'on leur adjoint un autre vecteur, alors ce dernier est une combinaison linéaire (unique) des premiers.

Il est capital de se rendre compte que si un vecteur est combinaison linéaire de vecteurs alors cette combinaison linéaire est unique si, et seulement si les vecteurs sont linéairement indépendants.

Enfin, *parmi les combinaisons linéaires de q vecteurs, il y en a au plus q linéairement indépendantes.* De façon équivalente, $q+1$ combinaisons linéaires de q vecteurs sont linéairement dépendantes.

Il s'en suit que le nombre maximum de vecteurs linéairement indépendants de $\mathbb{R}^p$ est au plus p, puisque tous les vecteurs de cet espace peuvent s'écrire sous forme d'une combinaison linéaire des p vecteurs unitaires. Par conséquent, puisque ces derniers sont en nombre égal à p et linéairement indépendants, ce nombre maximum vaut exactement p. On dit que la dimension de l'espace $\mathbb{R}^p$ est p.

D.2.4 Sous-espaces linéaires

On appelle sous-espace linéaire de $\mathbb{R}^p$ tout ensemble de vecteurs qui contient toutes les combinaisons linéaires de ses vecteurs. On dit que l'espace est *fermé* sous les deux opérations suivantes : produit par un scalaire et addition vectorielle. Tout sous-espace contient le vecteur nul. Par exemple, le singleton composé du vecteur nul est un sous-espace linéaire (on le dit trivial).

Soient $\boldsymbol{x}_i, \forall i = 1, \ldots, q$. Alors, l'ensemble des combinaisons linéaires de ces vecteurs est un sous-espace vectoriel : on parle du sous-espace *engendré* par ces vecteurs et on le notera $\mathcal{L}_{\{\boldsymbol{x}_i\}}$.

On appelle *dimension* d'un sous-espace vectoriel $\mathcal{L}$ le nombre maximum (s'il existe) de vecteurs linéairement indépendants : on le note $\dim(\mathcal{L})$. On déduit de ce qui précède que tout sous-espace de $\mathbb{R}^p$ a une dimension inférieure ou égale à p. La dimension du sous-espace trivial est nulle. Si la dimension d'un sous-espace vaut p, alors ce sous-espace est identique à $\mathbb{R}^p$.

L'intersection de deux sous-espaces linéaires est encore un sous-espace linéaire, éventuellement trivial. L'union de deux sous-espaces linéaires n'est en général pas un sous-espace linéaire. Le complément d'un sous-espace linéaire n'est jamais un sous-espace linéaire.

Etant donné deux sous-espaces linéaires $\mathcal{L}_1$ et $\mathcal{L}_2$, on définit leur somme, notée $\mathcal{L}_1 + \mathcal{L}_2$ par

$$\mathcal{L}_1 + \mathcal{L}_2 \stackrel{\triangle}{=} \{\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y} | \boldsymbol{x} \in \mathcal{L}_1, \boldsymbol{y} \in \mathcal{L}_2\}. \tag{D.22}$$

Notons que $\mathcal{L}_1 + \mathcal{L}_2$ est un sous-espace linéaire, mais il est possible que la décomposition ne soit pas unique. Si la décomposition est unique, cela veut dire que toute paire de vecteurs non-nuls $\boldsymbol{x}$ et $\boldsymbol{y}$ sont indépendants. On utilise alors le terme de somme directe et on utilise la notation $\mathcal{L}_1 \oplus \mathcal{L}_2$. Dans ce cas, l'intersection $\mathcal{L}_1 \cap \mathcal{L}_2 = \{\boldsymbol{0}\}$ et $\dim(\mathcal{L}_1 \oplus \mathcal{L}_2) = \dim(\mathcal{L}_1) + \dim(\mathcal{L}_2)$.

On appelle *base* d'un sous-espace linéaire tout ensemble de vecteurs linéairement indépendants en nombre égal à la dimension. Il s'en suit que tout vecteur du sous-espace peut s'écrire sous la forme d'une combinaison linéaire des vecteurs d'une base. De plus, cette combinaison linéaire est unique.

Etant donné q vecteurs linéairement indépendants d'un sous-espace de dimension $r > q$, il est possible de leur adjoindre $r - q$ vecteurs linéairement indépendants de façon à former une base.

D.2.5 Vecteurs et sous-espaces orthogonaux

Rappelons que deux vecteurs non-nuls sont orthogonaux par définition si leur produit scalaire est nul. Donc, les vecteurs unitaires $\boldsymbol{e}_i$ sont orthogonaux deux à deux et linéairement indépendants. Un vecteur est normé si sa norme vaut 1. Tout vecteur non-nul peut évidemment être normé en le multipliant par l'inverse de sa norme.

Plus généralement, des vecteurs orthogonaux deux à deux sont linéairement indépendants. Inversément, étant donné un ensemble $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}$ de q vecteurs linéairement indépendants, il est possible de construire q combinaisons linéaires de ces vecteurs, orthogonales deux à deux et normées. Par conséquent, l'existence d'une base implique l'existence d'une base de vecteurs orthonormés (orthogonaux et normés).

L'ensemble des vecteurs orthogonaux à ces q vecteurs est encore un sous-espace vectoriel : on le notera $\mathcal{L}^{\perp}_{\boldsymbol{x}_i}$ et nous noterons par $\mathcal{L}_{\boldsymbol{x}_i}$ l'espace engendré par les $\boldsymbol{x}_i$.

On a

$$\mathbb{R}^p = \mathcal{L}_{\{\boldsymbol{x}_i\}} \oplus \mathcal{L}^{\perp}_{\{\boldsymbol{x}_i\}}, \tag{D.23}$$

et la somme des dimensions d'un sous-espace engendré et d'un sous-espace orthogonal relatifs à un même ensemble de vecteurs vaut p :

$$\dim(\mathcal{L}_{\{\boldsymbol{x}_i\}}) + \dim(\mathcal{L}^{\perp}_{\{\boldsymbol{x}_i\}}) = p. \tag{D.24}$$

Tous les vecteurs du premier sont orthogonaux à tous les vecteurs du second. Tout vecteur peut donc s'écrire de manière unique sous la forme d'une somme de deux vecteurs appartenant respectivement à chacun de ces espaces. La concaténation de bases orthogonales de ces deux espaces fournit une base orthogonale de $\mathbb{R}^p$.

D.3 FONCTIONNELLES, APPLICATIONS ET OPÉRATEURS LINÉAIRES

Avant de passer au calcul matriciel, nous donnons ici une motivation pour la définition de ce types d'objets. Nous allons voir que les matrices apparaissent sous la forme d'une représentation d'applications linéaires. Ces notions seront définies de manière plus générale et circonstanciée à l'annexe F.4.

D.3.1 Fonctionnelles et produit scalaire

Une *fonctionnelle linéaire* (on dit aussi une forme linéaire) est une application $f(\cdot)$ de $\mathbb{R}^p$ dans $\mathbb{R}$ (resp. de C^p dans C) telle que

$$f(\sum_i \lambda_i \boldsymbol{x}_i) = \sum_i \lambda_i f(\boldsymbol{x}_i).$$

Etant données deux fonctionnelles linéaires $f(\cdot)$ et $f'(\cdot)$ on définit la somme et la multiplication par un scalaire de la manière suivante :

$$(f + f')(\boldsymbol{x}) = f(\boldsymbol{x}) + f'(\boldsymbol{x}) \tag{D.25}$$

$$(\lambda f)(\boldsymbol{x}) = \lambda f(\boldsymbol{x}), \tag{D.26}$$

et ces deux opération définissent (on peut aisément le démontrer) un structure d'espace linéaire (voir F.4).

On dit que cet espace est l'espace dual de $\mathbb{R}^p$ (resp. C^p).

Représentation. Il est clair que la linéarité implique que la fonctionnelle est entièrement déterminée par la donnée des $f_i = f(\boldsymbol{e}_i), \forall i$, autrement dit par la donnée d'un vecteur de dimension p $\boldsymbol{f}$, et on a $\boldsymbol{f}^t\boldsymbol{x} = f(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^p$. On dit que le vecteur $\boldsymbol{f}$ représente la fonctionnelle dans la base $\boldsymbol{e}_i$. Deux fonctionnelles sont alors égales si et seulement si leurs vecteurs de représentation sont identiques. L'ensemble des fonctionnelles est donc en bijection avec l'espace sur lequel les fonctionnelles sont définies.

D.3.2 Applications linéaires

Une *application linéaire* de $\mathbb{R}^n$ dans $\mathbb{R}^m$ est une application $T(\cdot)$ qui est telle que

$$T(\sum_i \lambda_i \boldsymbol{x}_i)) = \sum_i \lambda_i T(\boldsymbol{x}_i).$$

On peut définir, similairment à ce qui a été fait au pragraphe précédent, la somme et la multiplication par un scalaire d'applications linéaires, et se convaincre que l'ensemble des applications linéaires forme encore un espace vectoriel de dimension nm.

Représentation. Il est claire qu'un application linéaire est entièrement définie par la donnée des n vecteurs images (à m dimensions) des vecteurs de la base $\boldsymbol{e}_i$ de $\mathbb{R}^n$. Ces vecteurs définissent un tableau (ou une matrice) $n \times m$.

D.3.3 Opérateur linéaire

Une application linéaire de $\mathbb{R}^p$ dans lui-même est désignée par le nom d'opérateur linéaire. Il donne lieu à une matrice carrée d'ordre p.

D.4 FONCTIONS CONVEXES DANS UN ESPACE EUCLIDIEN

Une fonction ($\phi(\cdot) : \mathbb{R} \to \mathbb{R}$) est dite **convexe** (au sens restreint du terme) sur un intervalle $[a, b]$, si $\forall x_1, x_2 \in [a, b]$ et $\forall \lambda \in [0, 1]$ on a

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2). \tag{D.27}$$

Elle est dite **strictement convexe**, si elle est convexe et si l'égalité n'est satisfaite que lorsque $\lambda \in \{0, 1\}$ ou $x_1 = x_2$.

Si $\lambda \in [0, 1]$, on dit que $\lambda x_1 + (1 - \lambda)x_2$ est une combinaison linéaire convexe. Un ensemble C est dit convexe, s'il contient toutes ses combinaisons linéaires convexes. Dans $\mathbb{R}$ tout intervalle est convexe, et tout ensemble convexe est un intervalle (éventuellement égal à $\mathbb{R}$).

Une fonction $\phi(\cdot)$ est dite **concave** (resp. strictement concave) si $-\phi(\cdot)$ est convexe (resp. strictement convexe).

Au sens large on dit qu'une fonction est convexe si elle est concave ou convexe au sens restreint du terme. On convient alors d'indiquer par un des symboles ($\cap$ ou $\cup$) le sens de la concavité.

La notion d'ensemble convexe se généralise aux sous-ensembles d'un espace linéaire quelconque E. Un sous-ensemble de $\mathbb{R}^p$ est dit convexe s'il contient l'ensemble de ses combinaisons linéaires convexes. Tout sous-espace linéaire est un sous-ensemble convexe. La notion de fonction convexe sur un tel ensemble se généralise immédiatement aussi. On a les deux propriétés importantes suivantes

Toute fonction convexe sur un ensemble convexe est continue à l'intérieur de cet ensemble.

Toute fonction convexe sur un ensemble convexe borné y admet un minimum global.

On en déduit immédiatement que

Toute fonction concave sur un ensemble convexe est continue à l'intérieur de cet ensemble.

Toute fonction concave sur un ensemble convexe borné y admet un maximum global.

Les fonctions convexes (ou concaves) jouent un rôle particulièrement important dans les applications des probabilités, et en particulier en théorie de l'information.

Revenant au cas simple des fonctions définies sur $\mathbb{R}$, on a les deux propriétés suivantes.

Critère. *Si $\phi(\ldots)$ admet une dérivée seconde qui est positive ou nulle sur un intervalle (resp. négative ou nulle), alors cette fonction est convexe (resp. concave).*

D.5 RAPPELS DE CALCUL MATRICIEL

Nous ne donnons ici qu'un rappel très synthétique sur le calcul matriciel en nombres réels, en indiquant quelques généralisations aux cas de matrices de nombres complexes. Le livre [Gan66] constitue une très bonne référence à laquelle on pourra faire appel en cas de nécessité.

D.5.1 Définitions et notations

On appelle matrice réelle (resp. complexe) $n \times m$ un tableau de nm nombres réels (resp. complexes) organisés comme suit

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}, \tag{D.28}$$

c'est-à-dire en n lignes et m colonnes. Nous utiliserons la plupart du temps des lettres latines grasses (généralement majuscules) pour désigner des matrices (p.ex. $\boldsymbol{A}, \boldsymbol{B}, \ldots$), et nous désignerons l'élément d'une matrice $\boldsymbol{A}$ figurant à la i-ème ligne et la j-ème colonne par $a_{i,j}$, c'est-à-dire en utilisant la lettre latine minuscule correspondante. Nous utiliserons également la notation $(\boldsymbol{A})_{i,j}$ pour désigner l'élément i, j de $\boldsymbol{A}$.

Inversément, étant donné une collection de valeurs réelles (resp. complexes) $\{x_{i,j} : i = 1, \ldots, n; j = 1, \ldots, m\}$ nous désignerons par $[x_{i,j}]$ la matrice $\boldsymbol{X}$ formée par ces valeurs, où il est sous-entendu que i varie dans $\{1, \ldots, n\}$ et j dans $\{1, \ldots, m\}$. Par ailleurs, étant donné une collection de matrices $\boldsymbol{A}_{i,j}$ de dimensions $n_i \times m_j$,

avec $i = 1, \ldots, n; j = 1, \ldots, m$, nous pouvons former une nouvelle matrice de dimension $(\sum_i n_i) \times (\sum_j m_j)$ obtenue par concaténation des éléments des matrices $\boldsymbol{A}_{i,j}$ comme suit

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{1,1} & \boldsymbol{A}_{1,2} & \cdots & \boldsymbol{A}_{1,m} \\ \boldsymbol{A}_{2,1} & \boldsymbol{A}_{2,2} & \cdots & \boldsymbol{A}_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{A}_{n,1} & \boldsymbol{A}_{n,2} & \cdots & \boldsymbol{A}_{n,m} \end{bmatrix}. \tag{D.29}$$

Insistons sur le fait que les éléments de cette matrice sont les éléments des matrices constitutives et non pas ces matrices.

Inversément, étant donné une matrice $\boldsymbol{A}$ de dimension $n \times m$ il est possible d'en extraire une sous-matrice constituée des éléments de certaines colonnes et lignes de $\boldsymbol{A}$. Pour la convenance, nous noterons cette matrice $\boldsymbol{A}_{I,J}$ où $I = \{i_1, \ldots, i_k\}$ $(k \leq n, i_i < i_{i+1})$ et $J = \{j_1, \ldots, j_l\}$ $(l \leq m, j_i < j_{i+1})$ désignent les ensembles d'indices (en ordre croissant) de lignes et colonnes extraites, et $k \times l$ est la dimension de la matrice extraite.

Insistons sur le fait que l'ensemble des matrices réelles (resp. complexes) de dimension $p \times 1$ n'est pas identique à l'ensemble des vecteurs $\mathbb{R}^p$ (resp. $\boldsymbol{C}^p$). Cependant, nous ne distinguerons pas, la plupart du temps, ces deux ensembles. Nous noterons par $\mathbb{R}^{n \times m}$ l'ensemble de toutes les matrices réelles de dimension $n \times m$ (resp. $\boldsymbol{C}^{n \times m}$).

En particulier nous noterons par $\boldsymbol{a}_{.,j}$ le vecteur (colonne) de $\mathbb{R}^n$ formé par les éléments de la j-ème colonne de $\boldsymbol{A}$ et par $\boldsymbol{a}_{i,.}$ le vecteur (colonne) de $\mathbb{R}^m$ formé par les éléments de la i-ème ligne.

Une matrice de dimension $n \times n$ est dite carrée d'ordre n.

Deux matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ de même dimension sont par définition identiques si tous leurs éléments le sont

$$\boldsymbol{A} = \boldsymbol{B} \Leftrightarrow \forall i, j \ : a_{i,j} = b_{i,j}. \tag{D.30}$$

Matrices associées. Etant donné une matrice $\boldsymbol{A}$ de dimension $n \times m$, on appelle transposée de $\boldsymbol{A}$, qu'on note en abrégé $\boldsymbol{A}^T$, la matrice de dimension $m \times n$ définie par

$$(\boldsymbol{A}^T)_{i,j} = (\boldsymbol{A})_{j,i}. \tag{D.31}$$

Etant donné une matrice complexe $\boldsymbol{A}$ on peut définir la matrice complexe conjuguée notée $\overline{\boldsymbol{A}}$ dont les éléments sont les complexes conjugués (nous notons par $\overline{\alpha}$ le complexe conjugué du nombre complexe α) des éléments de $\boldsymbol{A}$. Si la matrice est réelle on a évidemment $\overline{\boldsymbol{A}} = \boldsymbol{A}$. On définit également la matrice adjointe par

$$\boldsymbol{A}^* \triangleq \overline{\boldsymbol{A}^T} = \overline{\boldsymbol{A}}^T.$$

D.5.2 Espaces vectoriels de matrices

Notation: $\mathbb{R}^{n \times m}$ et $C^{n \times m}$

Nous pouvons munir $\mathbb{R}^{n \times m}$ (resp. $\boldsymbol{C}^{n \times m}$) des opérations internes d'addition (élément par élément) et de multiplication par un nombre réel (resp. un nombre complexe), ce qui confère à $\mathbb{R}^{n \times m}$ (resp. $\boldsymbol{C}^{n \times m}$) une structure d'espace vectoriel. On a

$$\boldsymbol{C} = \alpha \boldsymbol{A} + \beta \boldsymbol{B} \Leftrightarrow \forall i, j \ : c_{i,j} = \alpha a_{i,j} + \beta b_{i,j}. \tag{D.32}$$

On a

$$(\alpha \boldsymbol{A} + \beta \boldsymbol{B})^T = \alpha \boldsymbol{A}^T + \beta \boldsymbol{B}^T, \tag{D.33}$$
$$\overline{(\alpha \boldsymbol{A} + \beta \boldsymbol{B})} = \overline{\alpha} \overline{\boldsymbol{A}} + \overline{\beta}\, \overline{\boldsymbol{B}}, \tag{D.34}$$
$$(\alpha \boldsymbol{A} + \beta \boldsymbol{B})^* = \overline{\alpha} \boldsymbol{A}^* + \overline{\beta} \boldsymbol{B}^*. \tag{D.35}$$

Dans cet espace on appelle matrice nulle, la matrice notée $\mathbf{0}$ dont tous les élements sont nuls. On a

$$\mathbf{0} = 0\boldsymbol{A}, \forall \boldsymbol{A}. \tag{D.36}$$

D.5.3 Multiplication de deux matrices

On dit que la matrice de $\boldsymbol{C}$ de dimension $n \times m$ est le produit d'une matrice $\boldsymbol{A}$ de dimension $n \times p$ et d'une matrice $\boldsymbol{B}$ de dimension $p \times m$ si

$$c_{i,j} = \sum_{k=1}^{p} a_{i,k} b_{k,j}, \tag{D.37}$$

$\forall i = 1, \ldots n \forall j = 1, \ldots m$. En notation abrégée on écrit alors $\boldsymbol{C} = \boldsymbol{AB}$.

On voit que le produit scalaire de deux vecteurs de $\mathbb{R}^p$ apparaît comme cas particulier du produit de deux matrices $1 \times p$ et $p \times 1$, ce qui justifie la notation $\boldsymbol{x}^T \boldsymbol{y}$ du chapitre précédent.

Il est important de noter que le produit de matrices n'est défini que si ces deux matrices sont compatibles du point de vue de leur dimensions. On a

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T \boldsymbol{A}^T, \tag{D.38}$$

mais

$$\overline{\boldsymbol{AB}} = \overline{\boldsymbol{A}}\,\overline{\boldsymbol{B}} \tag{D.39}$$

et donc

$$(\boldsymbol{AB})^* = \boldsymbol{B}^* \boldsymbol{A}^*. \tag{D.40}$$

D'autre part, le produit matriciel est distributif par rapport à l'addition matricielle

$$\boldsymbol{C}(\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{CA} + \boldsymbol{CB}, \tag{D.41}$$

associatif

$$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C} \tag{D.42}$$

et on a

$$\alpha(\boldsymbol{AB}) = (\alpha \boldsymbol{A})\boldsymbol{B} = \boldsymbol{A}(\alpha \boldsymbol{B}). \tag{D.43}$$

D.5.4 Déterminants

Sauf mention explicite du contraire, nous supposons dans cette section que les matrices sont carrées.

Permutations. Soit $\{1, \ldots, n\}$ l'ensemble des n premiers entiers, et soient $\nu_1, \ldots, \nu_n$ et $\mu_1, \ldots, \mu_n$ deux suites formées de ces nombres dans un certain ordre. Alors, l'opération P par laquelle ν_i devient μ_i s'appelle une permutation des indices $1, \ldots, n$; on la note en abrégé $P(\nu_i \to \mu_i)$. La permutation $P(\mu_i \to \mu_i)$ qui laisse inchangés les indices porte le nom de permutation identité, et sera notée I dans la suite.

Il est clair, que par un réarrangement simultané des indices ν_i et μ_i on ne change pas la permutation. On peut donc toujours s'arranger pour que $\nu_i = i, \forall i \leq n$. Nous utiliserons donc ci-dessous la notation simplifiée $P(\mu_i)$ pour désigner la permutation $P(i \to \mu_i)$.

La seule opération entre permutations est le produit et en particulier l'inversion. Le produit de deux permutations P et Q est la permutation obtenue en appliquant d'abord la permutation Q et au résultat de celle-ci la permutation P. Notons qu'en général on a $PQ \neq QP$. L'inverse d'une permutation P est une permutation notée P^{-1} telle que $PP^{-1} = P^{-1}P = I$; elle s'obtient en intervertissant le rôle des indices : $\mu_i \leftrightarrow \nu_i$.

Un *cycle* de longueur $p \leq n$ est une permutation $\nu_i \to \mu_i$ telle que $\mu_p = \nu_1$ et $\mu_i = \nu_{i+1}, \forall i < p$ et $\nu_i = \mu_i, \forall i = p+1, \ldots, n$. On appelle *transposition* un cycle de longueur 2 : une transposition est donc une permutation qui permute deux indices i et j et laisse inchangés les autres. On montre que toute permutation est le produit d'un certain nombre de transpositions. Cette décomposition n'est pas unique, mais la parité du nombre de transposition est unique. On l'appelle *parité* de la permutation et on appelle *signe* de la permutation P le nombre $\text{sign}(P)$ qui vaut $(-1)^k$ où k vaut 0 pour une permutation paire et 1 pour une permutation impaire (ou de manière équivalente, k représente le nombre de transpositions dans une représentation de P.)

Sur un ensemble de n il existe exactement $n!$ permutations différentes, la moitié sont des permutations paires, et l'autre moitié sont des permutations impaires.

Nous désignerons ci-dessous par $\mathcal{P}(1, \ldots, n)$ l'ensemble des permutations des n premiers indices, et par $P(\mu_i)$ une permutation quelconque.

Définition du déterminant. On appelle déterminant d'une matrice carrée $\boldsymbol{A} = [a_{i,j}]$ d'ordre n, le nombre

$$\text{dtm}\boldsymbol{A} = |\boldsymbol{A}| = \sum_{P \in \mathcal{P}(1,\ldots,n)} \text{sign}(P(\mu_i)) \prod_{i=1}^{n} a_{i,\mu_i}. \tag{D.44}$$

Vu ce que nous venons de dire sur $\mathcal{P}(1,\ldots,n)$, cette somme contient $n!$ termes, dont $\frac{n!}{2}$ sont précédés du signe $-$ et $\frac{n!}{2}$ sont précédés du signe $+$. Chaque terme est le produit de n éléments de A; il contient un élément et un seul de chaque ligne et colonne.

Cas particuliers. Pour $n = 1$, il n'existe qu'un seule permutation (la permutation identité, qui est paire) : on a $|\boldsymbol{A}| = a_{1,1}$.

Pour $n = 2$, on a $|\boldsymbol{A}| = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$.

Matrices associées. On a

$$|\boldsymbol{A}^T| = |\boldsymbol{A}|, \tag{D.45}$$

et

$$|\overline{\boldsymbol{A}}| = \overline{|\boldsymbol{A}|}, \tag{D.46}$$

et par conséquent

$$|\boldsymbol{A}^*| = \overline{|\boldsymbol{A}|}. \tag{D.47}$$

Matrices triangulaires composées. Considérons une matrice carrée et bloc triangulaire inférieure suivante

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{1,1} & 0 & \cdots & 0 \\ \boldsymbol{A}_{2,1} & \boldsymbol{A}_{2,2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{A}_{n,1} & \boldsymbol{A}_{n,2} & \cdots & \boldsymbol{A}_{n,n} \end{bmatrix}. \tag{D.48}$$

dont les sous-matrices diagonale $\boldsymbol{A}_{i,i}$ sont toutes carrées et toutes les matrices $\boldsymbol{A}_{i,j}, \forall j > i$ sont nulles. Alors on a

$$|\boldsymbol{A}| = \prod_{i=1}^{n} |\boldsymbol{A}_{i,i}|. \tag{D.49}$$

Il suffit de démontrer la propriété pour le cas particulier où $n = 2$. Le cas général s'en déduit immédiatement par plusieurs applications de ce cas particulier. Considérons alors la matrice

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{1,1} & 0 \\ \boldsymbol{A}_{2,1} & \boldsymbol{A}_{2,2} \end{bmatrix}, \tag{D.50}$$

et soient r et p respectivement l'ordre des matrices $\boldsymbol{A}_{1,1}$ et $\boldsymbol{A}_{2,2}$, avec $r + p = n$. Si nous appliquons la définition (D.44) du déterminant, nous obtenons

$$|\boldsymbol{A}| = \sum_{P \in \mathcal{P}(1,\ldots,n)} \text{sign}(P(\mu_i)) a_{1,\mu_1} \ldots a_{r,\mu_r} a_{r+1,\mu_{r+1}} \ldots a_{n,\mu_n}, \tag{D.51}$$

et nous voyons que seules les permutations qui consistent à choisir les r premier facteurs dans les r premières colonnes conduisent à des termes non-nuls. Alors, $\mu_1, \ldots, \mu_r$ est nécessairement une permutation de $1, \ldots, r$ et $\mu_{r+1}, \ldots, \mu_n$ une permutation de $r+1, \ldots, n$. Comme de plus le signe de cette permutation est le produit des signes des deux sous-permutations, cela revient à dire que tout terme "utile" peut s'écrire sous la forme

$$(\text{sign}(P(\mu_1, \ldots, \mu_r)) a_{1,\mu_1} \ldots a_{r,\mu_r})(\text{sign}(P(\mu_{r+1}, \ldots, \mu_n)) a_{r+1,\mu_{r+1}} \ldots a_{n,\mu_n}),$$

et on a donc bien $|\boldsymbol{A}| = |\boldsymbol{A}_{1,1}||\boldsymbol{A}_{2,2}|$. □

Comme les déterminants sont invariants par rapport à la transposition, cette propriété est également vraie pour une matrice bloc triangulaire supérieure.

Enfin, pour une matrice triangulaire inférieure, c'est-à-dire telle que $a_{i,j} = 0, \forall j > i$, on a

$$|\boldsymbol{A}| = \prod_{i=1}^{n} a_{i,i}. \tag{D.52}$$

Cette propriéte est évidemment également applicable à une matrice triangulaire supérieure.

Pour terminer, signalons que lorsqu'on permute deux rangées d'une matrice le déterminant de celle-ci change de signe (toutes les permutations changent de parité). On en déduit immédiatement que si deux rangées sont identiques alors le déterminant est nul.

Mineurs. On appelle mineur $A_{i,j}$ de l'élément $a_{i,j}$ d'une matrice carrée $\boldsymbol{A}$ d'ordre n le déterminant de la nouvelle matrice obtenue en supprimant la i-ème ligne et la j-ème colonne de $\boldsymbol{A}$ multiplié par $(-1)^{i+j}$. C'est aussi le déterminant obtenu en remplaçant $a_{i,j}$ par 1 dans $\boldsymbol{A}$ et en annulant les autres élements de la i-ème ligne, ou de la j-ème colonne ou des deux.

Si nous désignons par $\boldsymbol{r}$ le vecteur composé des éléments d'une rangée (ligne ou colonne) de $\boldsymbol{A}$ et par $\boldsymbol{m}$ le vecteur composé des mineurs correspondants, alors on montre que

$$\boldsymbol{r}^T \boldsymbol{m} = |\boldsymbol{A}|, \tag{D.53}$$

en d'autres termes, le déterminant d'une matrice dépend linéairement des éléments d'une quelconque de ces rangées. La démonstration est assez immédiate. En effet, étant donné la définition (D.44) chaque terme de la somme qui constitue le déterminant contient un et un seul facteur de la rangée considérée. Il doit donc nécessairement pouvoir s'écrire sous la forme

$$|\boldsymbol{A}| = \sum_i r_i m_i, \tag{D.54}$$

où nous avons mis en évidence les éléments de la rangée. Les m_j ne dépendent pas de la valeurs des éléments de la rangée. Et pour calculer leur valeur, il suffit de considérer le cas particulier où $r_i = \delta_{i,j}$ puisque dans ce cas on a $|\boldsymbol{A}| = \sum_i \delta_{i,j} m_i = m_j$. On obtient précisément le mineur de l'élément r_j. □

Ce résultat est fondamental, car il nous permet de déduire un certain nombre d'autres de façon immédiate.

Tout d'abord, on en déduit que si une rangée est nulle, alors le déterminant est nul (il suffit de développer le déterminant à partir de cette rangée).

Ensuite, si nous prenons deux rangées parallèles (et différentes) $\boldsymbol{r}$ et $\boldsymbol{r}'$ et les mineurs correspondants $\boldsymbol{m}$ et $\boldsymbol{m}'$, alors on a

$$\boldsymbol{r}^T \boldsymbol{m}' = \boldsymbol{r}'^T \boldsymbol{m} = 0. \tag{D.55}$$

En effet, montrons que le premier produit scalaire est nul. Prenons la matrice obtenue à partir de $\boldsymbol{A}$ en remplaçant la rangée $\boldsymbol{r}'$ par $\boldsymbol{r}$. Le déterminant de cette matrice est nul puisqu'elle dispose de deux rangées parallèles identiques. Mais ce déterminant est aussi égal à $\boldsymbol{r}^T \boldsymbol{m}'$. □

On déduit également de (D.53) que si une rangée est une combinaison linéaire d'un certain nombre de vecteurs, alors le déterminant peut également s'écrire sous la forme d'une combinaison linéaire de déterminants :

$$\boldsymbol{r} = \sum_i \alpha_i \boldsymbol{r}_i \Rightarrow |\boldsymbol{A}| = \sum_i \alpha_i \boldsymbol{r}_i^T \boldsymbol{m} \tag{D.56}$$

et $\boldsymbol{r}_i^T \boldsymbol{m}$ est évidemment le déterminant de la matrice obtenue à partir de $\boldsymbol{A}$ en remplaçant la rangée $\boldsymbol{r}$ par le vecteur $\boldsymbol{r}_i$.

De cette dernière propriété on déduit que le déterminant ne change pas si on ajoute à une rangée une combinaison linéaire des autres rangées.

Enfin, on en déduit qu'un déterminant est nul si, et seulement s'il existe une combinaison linéaire entre ses lignes (et donc forcément aussi entre ses colonnes). S'il existe une combinaison linéaire entre les rangées alors il est possible d'écrire une rangée comme combinaison linéaire des autres. La propriété (D.56) implique alors que le déterminant est nul car chaque terme du membre de droite de (D.56) est nul. Nous ne démontrerons pas la réciproque.

Déterminant du produit de deux matrices semblables. Soient $\boldsymbol{A}$ et $\boldsymbol{B}$ deux matrices $n \times m$ quelconques. Alors la matrice $\boldsymbol{AB}^T$ est une matrice $n \times n$ et son déterminant vaut

$$|\boldsymbol{AB}^T| = \sum_{k_1,\ldots,k_n \text{ parmi } 1,\ldots,m} |[\boldsymbol{a}_{\cdot,k_i}]||[\boldsymbol{b}_{\cdot,k_i}]| \tag{D.57}$$

où $k_1, \ldots, k_n$ est une combinaison de n indices de colonnes parmi m, $[\boldsymbol{a}_{\cdot,k_i}]$ (resp. $[\boldsymbol{b}_{\cdot,k_i}]$) désigne la matrice carrée formée des n colonnes correspondantes de $\boldsymbol{A}$ (resp. $\boldsymbol{B}$), et où la somme est effectuée sur toutes les combinaisons possibles de n colonnes parmi m. Nous ne démontrerons pas cette propriété qui porte le nom de "théorème de Binet-Cauchy".

En particulier, si $m < n$ la somme s'étend sur un nombre nul de combinaisons et vaut zéro. Cela correspond à la situation où n les rangées de $|\boldsymbol{AB}^T|$ sont des combinaisons linéaires de $m < n$ vecteurs, et sont donc linéairement dépendantes.

Si $m = n$, la somme se réduit à un seul terme et la relation devient pour deux matrices carrées :

$$|\boldsymbol{AB}^T| = |\boldsymbol{A}||\boldsymbol{B}^T|, \tag{D.58}$$

dont on déduit évidemment que

$$|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|. \tag{D.59}$$

Rang d'une matrice. Soient $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ p vecteurs de dimension n. Soit $\boldsymbol{X}$ la matrice $p \times n$ dont les lignes sont ces vecteurs. Désignons par $\boldsymbol{y}_j, \forall j = 1, \ldots, n$ les p colonnes de cette matrice, et par $[\boldsymbol{y}_k i]$ une matrice formée par les p colonnes $k_i, i = 1, \ldots, p$ de la matrice $\boldsymbol{X}$

Alors le déterminant $p \times p$

$$|\boldsymbol{XX}^*| = |[x_i x_j^*]| \tag{D.60}$$

est nul, si et seulement si les vecteurs sont linéairement dépendants, ce qui est vrai si, et seulement si tous les déterminants formés de p colonnes de la matrice $\boldsymbol{X}$ sont nuls.

Montrons d'abord que ces deux conditions sont équivalentes. On a, par application du théorème de Binet-Cauchy

$$|\boldsymbol{XX}^*| = \sum_{k_1,\ldots,k_p \text{ parmi } 1,\ldots,n} |[\boldsymbol{y}_{k_i}]||[\overline{\boldsymbol{y}_{k_i}}]| = \sum_{k_1,\ldots,k_p \text{ parmi } 1,\ldots,n} ||[\boldsymbol{y}_{k_i}]||^2, \tag{D.61}$$

dont l'équivalence des deux propriétés découle immédiatement.

Montrons que les deux conditions sont nécessaires. En effet, s'il existe une combinaison linéaire entre les vecteurs $\boldsymbol{x}_i$ tous les déterminants $|[\boldsymbol{y}_{k_i}]|$ sont nuls.

Montrons que les deux conditions sont suffisantes pour garantir la dépendance linéaire. Puisque $|\boldsymbol{XX}^*|$ est nul. Cela veut dire qu'il existe une relation linéaire entre les lignes de cette matrice, ce qui veut dire $\exists \alpha_1, \ldots, \alpha_p$ non tous nuls tels que

$$\sum_i \alpha_i \boldsymbol{x}_i \boldsymbol{x}_j{}^* = 0, \forall j = 1, \ldots, p \tag{D.62}$$

ce qui implique évidemment que

$$\sum_j \overline{\alpha_j} \sum_i \alpha_i \boldsymbol{x}_i \boldsymbol{x}_j{}^* = 0 \tag{D.63}$$

ce qui est équivalent à

$$||\sum_i \alpha_i \boldsymbol{x}_i|| = 0 \tag{D.64}$$

ce qui implique

$$\sum_i \alpha_i \boldsymbol{x}_i = \boldsymbol{0}, \tag{D.65}$$

ce qui prouve bien l'existence d'une combinaison linéaire (pas n'importe laquelle) entre les vecteurs $\boldsymbol{x}_i$. □

On en déduit les deux propriétés suivantes :

1. Des vecteurs dont le nombre dépasse la dimension sont linéairement dépendants.

2. Le nombre maximum de lignes et de colonnes linéairement indépendantes d'une matrice rectangulaire est le même.

La dernière propriété conduit à la définition du rang d'une matrice : le *rang* d'une matrice est par définition le nombre maximal de vecteurs lignes linéairement indépendants. Ce nombre est idendique au nombre maximal de vecteurs colonnes linéairement indépendants, et il est par conséquent majoré par $\min(n, m)$.

D.5.5 Matrices carrées

Dans cette section nous ne considérons que des matrices carrées d'ordre n, n étant quelconque.

Nous avons déjà défini la matrice nulle, dont tous les éléments sont nuls.

La matrice identité est la matrice $\boldsymbol{I} = [\delta_{i,j}]$ dont les éléments diagonaux sont égaux à 1 et les éléments non diagonaux sont tous nuls.

Rappelons que le produit de deux matrices d'ordre n donne une matrice d'ordre n, et que ce produit n'est pas en général commutatif. Par ailleurs, on a évidemment

$$\boldsymbol{AI} = \boldsymbol{IA} = \boldsymbol{A}, \tag{D.66}$$

où $\boldsymbol{A}$ est une matrice d'ordre n quelconque et $\boldsymbol{I}$ la matrice identité d'ordre n. Cette propriété justifie le nom de cette dernière matrice. Toute matrice commute aussi avec $\mathbf{0}$ et on a

$$\boldsymbol{A0} = \boldsymbol{0A} = \mathbf{0}. \tag{D.67}$$

Toute matrice commute évidemment avec elle-même et le produit de matrices est associatif à condition de respecter l'ordre des facteurs. On note $\boldsymbol{A}^m$ la m-ème puissance d'une matrice, c'est-à-dire le produit

$$\boldsymbol{A}^m = \overbrace{\boldsymbol{A} \cdots \boldsymbol{A}}^{m} , \tag{D.68}$$

et à partir de là on définit des polynômes de matrices. On trouve que les polynômes d'une même matrice commutent.

Le produit de matrices diagonales est la matrice diagonale formée par les produits des éléments diagonaux de même position. Nous noterons $\Delta(a_1, \ldots, a_n)$, et en abrégé $\Delta(a_i)$ une matrice diagonale et on vérifiera qu'on a bien

$$\Delta(a_i)\Delta(b_i) = \Delta(a_i b_i) = \Delta(b_i)\Delta(a_i). \tag{D.69}$$

Remarquons que toute matrice ne commute pas nécessairement avec toute matrice diagonale. En effet soit $\Delta(a_1, \ldots, a_n)$ une matrice diagonale et $\boldsymbol{B}$ une autre matrice. L'élément i, j de la matrice Δ peut alors s'écrire $a_i\delta_{i,j}$, et on a

$$(\Delta\boldsymbol{B})_{i,j} = \sum_{k=1}^{n} a_i \delta_{i,k} b_{k,j} = a_i b_{i,j}, \tag{D.70}$$

ce qui veut dire que la i-ème ligne est multipliée par a_i. Par ailleurs

$$(\boldsymbol{B}\Delta)_{i,j} = \sum_{k=1}^{n} b_{i,k} a_k \delta_{k,j} = b_{i,j} a_j. \tag{D.71}$$

ce qui veut dire que la i-ème colonne est multipliée par a_i. Le produit commute ssi $b_{i,j}a_j = a_i b_{i,j}, \forall i, j$, c'est-à-dire ssi $b_{i,j} = 0, \forall i, j | a_i \neq a_j$.

Une matrice d'ordre n peut être vue comme un *opérateur* linéaire sur l'espace des vecteurs de dimension n. En effet, on définit l'action d'une matrice $\boldsymbol{A}$ sur un vecteur $\boldsymbol{x}$ par le nouveau vecteur noté $\boldsymbol{Ax}$ définit par

$$(\boldsymbol{Ax})_i = \sum_{j=1}^{n} a_{i,j} x_j, \tag{D.72}$$

ce qui peut être vu comme un cas particulier du produit de deux matrices respectivement de dimensions $n \times n$ et de dimension $n \times 1$. On voit que l'action d'une matrice sur un vecteur produit une combinaison linéaire des colonnes de cette matrice. En effet,

$$\boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^{n} x_j \boldsymbol{a}_{\cdot,j}. \tag{D.73}$$

On a en particulier $\boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}$ et $\boldsymbol{0}\boldsymbol{x} = \boldsymbol{0}$ pour tout vecteur $\boldsymbol{x}$.

On montre qu'une matrice définit ainsi un transformation effectivement linéaire de l'espace linéaire de vecteurs, ce qui veut dire que l'action d'une matrice sur une combinaison linéaire de vecteurs est bien la combinaison linéaire des actions de cette matrice sur les vecteurs constitutifs de la combinaison linéaire :

$$\boldsymbol{A}(\sum_k \alpha_k \boldsymbol{x}_k) = \sum_k \alpha_k \boldsymbol{A}\boldsymbol{x}_k. \tag{D.74}$$

On montre immédiatement que

$$\boldsymbol{x}^*(\boldsymbol{A}\boldsymbol{y}) = (\boldsymbol{A}^*\boldsymbol{x})^*\boldsymbol{y}. \tag{D.75}$$

Grandeurs numériques attachées à une matrice. Nous avons déjà défini le déterminant d'une matrice à la section précédente.

La trace d'une matrice est définie comme la somme de ses éléments diagonaux :

$$\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{i,i}. \tag{D.76}$$

On a $\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A}^T) = \overline{\text{tr}(\overline{\boldsymbol{A}})} = \overline{\text{tr}(\boldsymbol{A}^*)}$ et on a

$$\text{tr}(\sum_k \alpha_k \boldsymbol{A}_k) = \sum_k \alpha_k \text{tr}(\boldsymbol{A}_k), \tag{D.77}$$

et

$$\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{B}\boldsymbol{A}). \tag{D.78}$$

Polynôme caractéristique. Par définition, le polynôme caractéristique d'une matrice est le polynôme en λ obtenu en développant le déterminant

$$|\boldsymbol{A} - \lambda \boldsymbol{I}|. \tag{D.79}$$

Il s'agit d'un polynôme de degré n (n étant l'ordre de la matrice) qui possède donc exactement n racines complexes (lorsqu'on les compte avec leur multiplicité) et m racines complexes distinctes ($m \leq n$). On les appelle les valeurs propres de $\boldsymbol{A}$. Le polynôme caractéristique peut donc s'écrire sous la forme

$$p(\lambda) = \prod_{i=1}^{m} (\lambda_i - \lambda)^{m_i} \tag{D.80}$$

où les λ_i sont toutes différentes et m_i désigne la multiplicité de la valeur propre λ_i.

En fonction de ces valeurs propres $\lambda_i, i = 1, \ldots, n$ on a

$$|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i, \tag{D.81}$$

et

$$\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i. \tag{D.82}$$

Il est évident que les valeurs propres d'une matrice diagonale sont précisément les éléments diagonaux de cette matrice. En particulier, le polynôme caractéristique de la matrice $\boldsymbol{I}$ est $p(\lambda) = (1 - \lambda)^n$, et 1 est la seule valeur propre de cette matrice.

On montre que si $p(x)$ est un polynôme, alors les valeurs propres de $p(\boldsymbol{A})$ sont les valeurs $p(\lambda_i)$, où λ_i sont les valeurs propres de $\boldsymbol{A}$. En particulier, on a donc

$$|p(\boldsymbol{A})| = \prod_{i=1}^{n} p(\lambda_i). \tag{D.83}$$

Cette formule a comme cas particulier la formule (D.81) si on prend le polynôme $p(x) = x$. D'autre part elle implique que si $p(\boldsymbol{A})$ est singulière alors il existe au moins une valeurs propre de $\boldsymbol{A}$ qui soit racine de $p(x)$. En particulier si $p(\boldsymbol{A}) = \boldsymbol{0}$, alors les valeurs propres de $p(\boldsymbol{A})$ sont toutes nulles, et donc $p(\lambda_i) = 0, \forall i$. Donc tout polynôme qui est annulé par une matrice doit avoir les valeurs propres distinctes de cette matrice comme racines.

Matrices inverses. On appelle matrice inverse d'une matrice $\boldsymbol{A}$, si elle existe, une matrice notée $\boldsymbol{A}^{-1}$ telle que

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}. \tag{D.84}$$

On montre que si la matrice inverse existe alors elle est unique, et on a

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}, \tag{D.85}$$

et on a $|\boldsymbol{A}| \neq 0$ et $|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}$. De plus, la condition $|\boldsymbol{A}| \neq 0$ est une condition suffisante d'existence de l'inverse.

Seules les matrices de rang égal à n admettent par conséquent une inverse. En particulier, l'inverse d'une matrice diagonale $\Delta(a_i)$ existe si et seulement si tous les a_i sont non nuls et est alors la matrice diagonales $\Delta(a_i^{-1})$.

Formule de Frobenius-Schur. Cette formule relie l'inverse d'une matrice composée aux inverses de matrices qui la composent. Soit une matrice composée

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}, \tag{D.86}$$

où $\boldsymbol{A}$ et $\boldsymbol{D}$ sont carrées.

Alors si $\boldsymbol{A}$ et $\boldsymbol{F} = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$ sont invertibles on a

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{F}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{F}^{-1} \\ -\boldsymbol{F}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{F}^{-1} \end{bmatrix}, \tag{D.87}$$

et si $\boldsymbol{D}$ et $\boldsymbol{G} = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ sont invertibles on a

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{G}^{-1} & -\boldsymbol{G}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{G}^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{G}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix}. \tag{D.88}$$

Notons qu'il peut arriver en pratique que ni $\boldsymbol{A}$ et $\boldsymbol{F}$ ni $\boldsymbol{D}$ et $\boldsymbol{G}$ ne soient simultanémement invertibles alors que la matrice composée l'est. Mais si ces couples de matrices sont conjointement invertibles alors la matrice composée l'est forcément, et son inverse est donnée par la (ou les) formules de Frobenius-Schur qui s'appliquent.

Transformation d'une matrice. Soit $\boldsymbol{S}$ une matrice invertible et $\boldsymbol{A}$ une matrice quelconque. Alors on appelle transformée de $\boldsymbol{A}$ par $\boldsymbol{S}$ la nouvelle matrice

$$\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}. \tag{D.89}$$

On montre que le déterminant et la trace d'une matrice restent invariants lors d'une telle transformation (conséquence du fait que ces deux grandeurs appliquées à un produit de matrices sont invariants lorsqu'on commute les facteurs), et par conséquent que la matrice reste invertible (resp. non invertible) par transformation.

Plus généralement, les relations en matrices formées de polynômes et de combinaisons linéaires restent également invariantes. En particulier, si une matrice est invertible alors la transformée de son inverse est l'inverse de sa transformée.

Comme la transformée laisse invariant le déterminant on en déduit que

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = |\boldsymbol{S}^{-1}(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{S}| = |\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S} - \lambda \boldsymbol{I}|, \tag{D.90}$$

ce qui veut dire que le polynôme caractéristique et donc l'ensemble des valeurs propres reste également invariants.

Il s'en suit que s'il existe une matrice $\boldsymbol{S}$ telle que $\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$ soit diagonale (on dit que $\boldsymbol{A}$ est diagonalisable), alors les éléments diagonaux de $\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$ sont les valeurs propres de $\boldsymbol{A}$.

Vecteurs propres. On dit que $\boldsymbol{x} \neq \boldsymbol{0}$ est un vecteur propre d'une matrice $\boldsymbol{A}$ si, et seulement s'il existe λ tel que

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{D.91}$$

Cela est équivalent à dire que

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = \boldsymbol{0} \tag{D.92}$$

ce qui implique qu'il existe une combinaison linéaire entre les colonnes de $\boldsymbol{A} - \lambda \boldsymbol{I}$ et donc que le déterminant de cette matrice est nul. Cela n'est donc possible que si λ est une valeur propre de $\boldsymbol{A}$.

On a les propriétés importantes suivantes :

1. l'ensemble des vecteurs propres d'une valeur propre donnée forme un sous-espace linéaire, la dimension de ce sous-espace ne peut dépasser la multiplicité de la valeur propre. Il peut être inférieur à cette multiplicité.
2. les vecteurs propres de valeurs propres différentes sont linéairement indépendants.
3. en conclusion, le nombre de vecteurs propres linéairement indépendants (toutes valeurs propres confondues) est inférieur ou égal à $\sum_{i=1}^{m} m_i = n$.

Si $\boldsymbol{S}$ est invertible et $\boldsymbol{x}$ est un v.p. de valeur propre λ de $\boldsymbol{A}$, alors $\boldsymbol{S}^{-1}\boldsymbol{x}$ est un vecteur propre de valeur propre λ de $\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$. Inversément, si $\boldsymbol{y}$ est un vecteur propre de v.p. λ de $\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$, alors $\boldsymbol{S}\boldsymbol{y}$ est un vecteur propre de $\boldsymbol{A}$ de valeur propre λ. Comme les vecteurs $\boldsymbol{x}_i$ sont linéairement indépendants si, et seulement si les vecteurs $\boldsymbol{S}^{-1}\boldsymbol{x}_i$ le sont, le nombre de vecteurs propres linéairement indépendants de chaque valeur propre ne change pas.

Tous les vecteurs $\boldsymbol{e}_i$ sont des vecteurs propres des matrices diagonales $\Delta(a_i)$, de valeur propre $\lambda_i = a_i$.

On déduit de ce qui précède que s'il existe une matrice $\boldsymbol{S}$ qui transforme $\boldsymbol{A}$ en une matrice diagonale alors les éléments diagonaux de cette dernière sont les valeurs propres de $\boldsymbol{A}$ et le nombre de vecteurs propres linéairement indépendants de chaque valeur propre est exactement égal à la multiplicité de celle-ci. Inversément, si chaque valeur propre admet m_i vecteurs propres linéairement indépendants, alors il existe $\boldsymbol{S}$ qui transforme $\boldsymbol{A}$ en une matrice diagonale. En effet, on peut alors choisir n vecteurs propres linéairement indépendants $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ de valeurs propres respectives $\lambda_1, \ldots, \lambda_n$ et la matrice $\boldsymbol{S}$ dont les colonnes sont ces vecteurs est invertible et fait le travail :

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{-1}\boldsymbol{A}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{-1}(\lambda_1 \boldsymbol{x}_1, \ldots, \lambda_n \boldsymbol{x}_n), \tag{D.93}$$

or

$$(\lambda_1 \boldsymbol{x}_1, \ldots, \lambda_n \boldsymbol{x}_n) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\Delta(\lambda_1, \ldots, \lambda_n). \tag{D.94}$$

et par conséquent

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{-1}\boldsymbol{A}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \Delta(\lambda_1, \ldots, \lambda_n), \tag{D.95}$$

comme le produit de matrices est associatif.

La matrice $\boldsymbol{S}$ qui diagonalise la matrice $\boldsymbol{A}$ est donc formée par des vecteurs propres linéairement indépendants de $\boldsymbol{A}$.

Si la matrice est réelle, les coefficients du polynôme caractéristique sont tous réels et les valeurs propres sont donc soit réelles soit viennent par paires complexes conjuguées. D'autre part, si $\boldsymbol{x}$ est vecteur propre de valeur propre λ alors $\overline{\boldsymbol{x}}$ est vecteur propre de la valeur propre de $\overline{\lambda}$. On en déduit que si une valeur propre est réelle, alors si Re$\boldsymbol{x}$ est non nul (resp. Im$\boldsymbol{x}$)) il est vecteur propre en même temps que $\boldsymbol{x}$ de cette valeur propre.

Montrons que si les $\boldsymbol{x}_i$ sont k vecteurs propres complexes linéairement indépendants d'une même valeur propre réelle, alors il est possible de construire un ensemble de k vecteurs propres réels également linéairement indépendants. En effet, si c'est le cas, alors le rang $\rho(\boldsymbol{A} - \lambda \boldsymbol{I})$ est inférieur à $n - k$. Comme cette matrice est réelle, il doit donc exister au moins k vecteurs réels linéairement indépendants tels que $(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{y}_i = \boldsymbol{0}$. □

Polynôme minimum. Par définition, le polynôme minimum d'une matrice est le quotient de son polynôme caractéristique par la plus grand commun diviseur des mineurs de la matrice $\boldsymbol{A} - \lambda \boldsymbol{I}$. L'adjectif "minimum" est justifié par le fait que tout polynôme qui est annulé par $\boldsymbol{A}$ est multiple du polynôme minimum. Pour une matrice diagonale on trouve directement que le polynôme minimum s'écrit

$$m(\lambda) = \prod_{j=1}^{m} (\lambda - a_{i_j}), \tag{D.96}$$

où les a_{i_j} désignent les m valeurs diagonales distinctes de cette matrice.

On montre que toute matrice annule son polynôme minimum, et donc a fortiori son polynôme caractéristique. On montre également que toutes les valeurs propres de $\boldsymbol{A}$ sont des racines de son polynôme minimum. Par conséquent, si le polynôme caractéristique n'a que des racines simples il coïncide avec le polynôme minimum.

La propriété la plus importante est la suivante :

La condition nécessaire et suffisante pour qu'une matrice puisse se diagonaliser est que son polynôme minimum n'ait que des racines simples.

D.5.6 Matrices hermitiennes et unitaires

Une matrice complexe $\boldsymbol{H}$ est dite hermitienne si $\boldsymbol{H}^* = \boldsymbol{H}$. Dans le cas réel, on dit que la matrice est symétrique. Le matrices de variance-covariance forment un exemple de matrice réelle symétrique. La somme de deux matrices hermitiennes est une matrice hermitienne. Le déterminant d'une matrice hermitienne est réel. On a $\forall \boldsymbol{x}, \boldsymbol{y}$ vecteurs

$$\boldsymbol{x}^*(\boldsymbol{H}\boldsymbol{y}) = (\boldsymbol{x}^*\boldsymbol{H})\boldsymbol{y} = (\boldsymbol{H}\boldsymbol{x})^*\boldsymbol{y}. \tag{D.97}$$

On a également

$$\overline{\boldsymbol{x}^*\boldsymbol{H}\boldsymbol{x}} = (\boldsymbol{x}^T\overline{\boldsymbol{H}}\overline{\boldsymbol{x}}) = (\boldsymbol{x}^T\overline{\boldsymbol{H}}\overline{\boldsymbol{x}})^T = \boldsymbol{x}^*\boldsymbol{H}^*\boldsymbol{x} = \boldsymbol{x}^*\boldsymbol{H}\boldsymbol{x}, \tag{D.98}$$

et par conséquent $\boldsymbol{x}^*\boldsymbol{H}\boldsymbol{x}$ est réel quel que soit $\boldsymbol{x}$. Supposons alors que $\boldsymbol{x}$ est un vecteur propre de valeur propre λ, alors donc $\boldsymbol{H}\boldsymbol{x} = \lambda\boldsymbol{x}$. On déduit de l'égalité précédente que

$$\boldsymbol{x}^*(\boldsymbol{H}\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|^2 = \overline{\lambda}\|\boldsymbol{x}\|^2, \tag{D.99}$$

autrement dit, λ est réelle.

Une matrice complexe $\boldsymbol{U}$ est dite unitaire si $\boldsymbol{U}^* = \boldsymbol{U}^{-1}$. Les lignes ou bien les vecteurs colonnes d'une telle matrice forment une base orthonormée de C^p. Dans le cas réel on dit que la matrice est orthogonale. Le produit de deux matrices unitaires est une matrice unitaire. Le déterminant d'une matrice unitaire est de module égal à un. On a vis-à-vis du produit scalaire

$$(\boldsymbol{U}\boldsymbol{x})^*(\boldsymbol{U}\boldsymbol{y}) = \boldsymbol{x}^*\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{y} = \boldsymbol{x}^*\boldsymbol{y}. \tag{D.100}$$

On en déduit que le module des valeurs propres vaut 1. En effet, on a en appliquant la propriété précédente à un vecteur propre de $\boldsymbol{U}$

$$\lambda\overline{\lambda}\boldsymbol{x}^*\boldsymbol{x} = \boldsymbol{x}^*\boldsymbol{x} \Rightarrow \lambda\overline{\lambda} = 1. \tag{D.101}$$

La transformée d'une matrice hermitienne (resp. unitaire) par une matrice unitaire reste une matrice hermitienne (resp. unitaire).

Les deux types de matrices partagent la propriété suivante :

Les vecteurs propres de valeurs propres distinctes sont orthogonaux.

Il suffit d'appliquer les égalités (D.97) (resp. D.100) à deux vecteurs propres de valeurs propres distinctes. Par exemple dans le cas d'une matrice hermitienne on trouve, en supposant que $\boldsymbol{H}\boldsymbol{x} = \lambda\boldsymbol{x}$ et $\boldsymbol{H}\boldsymbol{y} = \mu\boldsymbol{y}$ et $\lambda \neq \mu$:

$$(\lambda\boldsymbol{x})^*\boldsymbol{y} = \boldsymbol{x}^*(\mu\boldsymbol{y}) \Rightarrow (\overline{\lambda} - \mu)\boldsymbol{x}^*\boldsymbol{y} = 0, \tag{D.102}$$

ce qui compte tenu du fait que $\overline{\lambda} = \lambda \neq \mu$ implique que

$$\boldsymbol{x}^*\boldsymbol{y} = 0. \tag{D.103}$$

On montre que toute matrice hermitienne ou unitaire peut être diagonalisée au moyen d'une matrice unitaire. En admettant cette propriété sans démonstration, nous pouvons voir que cette matrice est composée de vecteurs propres orthogonaux, obtenus de la manière suivante : pour chaque valeur propre λ_i de multiplicité m_i on construit à partir de m_i vecteurs propres linéairement indépendants (on sait qu'ils existent car la matrice est diagonalisable) un ensemble orthonormé. Puisque les vecteurs propres de valeurs propres différentes sont orthogonaux, l'ensemble formé par tous ces vecteurs propres, toutes valeurs propres confondues est un ensemble orthonormé, qui peut être utilisé directement pour former une matrice unitaire.

Donc, toute matrice symétrique ou orthogonale peut être diagonalisée au moyen d'une matrice orthogonale. Montrons, par exemple que si la matrice est symétrique réelle, alors c'est bien vrai. Si la matrice est symétrique et réelle, elle est hermitienne (et par là de valeurs propres réelles). Or nous savons que pour une matrice réelle, l'existence de k vecteurs propres linéairement indépendants implique l'existence de k vecteurs propres réels linéairement indépendants. Ceux-ci peuvent encore être orthonormés, ce qui implique l'existence d'une matrice orthogonale qui diagonalise $\boldsymbol{A}$.

Exemples de matrices hermitiennes. Toutes les matrices qui ont la forme $\boldsymbol{A}^*\boldsymbol{A}$ où $\boldsymbol{A}$ est une matrice quelconque (pas nécessairement carrée) sont hermitiennes, et si $\boldsymbol{A}$ est réelle, elles sont symétriques.

Toutes les matrices diagonales réelles sont hermitiennes et symétriques.

Toutes les matrices bloc diagonales composées de blocs hermitiens (resp. symétriques) sont hermitiennes (resp. symétriques).

Enfin les matrices de Toeplitz, décrites ci-dessous sont symétriques.

D.5.7 Matrices de Toeplitz

Nous allons faire une petite digression en montrant l'analogie de ce qui vient d'être vu avec certaines notions de théorie des systèmes. Nous en parlerons plus en détails à la fin de cet appendice, lorsque nous analyserons les espaces de Hilbert.

Une matrice $\boldsymbol{T}$ réelle est dite de Toeplitz si $t_{i,j} = t_{k,l}, \forall i,j,k,l||i-j| = |k-l|$. Autrement dit, la valeur d'un élément ne dépend que de la différence (absolue) entre ses indices de lignes et de colonnes. Ce type de matrice est donc de la forme

$$\boldsymbol{H} = \begin{bmatrix} h_0 & h_1 & \cdots & h_n \\ h_1 & h_0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & h_1 \\ h_n & \cdots & h_1 & h_0 \end{bmatrix}, \tag{D.104}$$

et, en particulier est symétrique. Ce type de matrice est fréquemment recontré dans les applications en traitement de signal et en théorie des systèmes, stochastiques notamment. Par exemple, pour une processus stationnaire et un ensemble fini d'instants équidistants $t_i = t_0 + k\Delta t, k = 0, \ldots, n$, la matrice de variance-covariance finidimensionnelle dont les éléments $R_{i,j} \triangleq E\{\mathcal{X}(t_i)\mathcal{X}(t_j)\}$ est une matrice de Toeplitz.

Si nous travaillons avec des signaux en temps discret de durée finie (disons $n+1$ échantillons), nous pouvons représenter un signal par un vecteur $(x_0, \ldots, x_n)^T$. Pour la simplicité supposons que n soit pair et supposons que $\frac{n}{2}+1$ représente l'origine des temps. Alors l'action d'une matrice $\boldsymbol{A}$ sur un vecteur représente l'action d'un système linéaire sur le signal correspondant. La i-ème colonne de $\boldsymbol{A}$ représente alors la réponse impulsionnelle du système linéaire pour une impulsion d'entrée située à l'instant t_{i-1}. En particulier, la colonne centrale représente la réponse impulsionnelle pour une impulsion en $t = 0$. Nous voyons que si la matrice est de Toeplitz, cette réponse est le vecteur $(h_{\frac{n}{2}}, \ldots, h_1, h_0, h_1, \ldots, h_{\frac{n}{2}})$, qui est non-causale (à moins que le système ne soit l'opérateur identité). On voit que les autres colonnes sont simplement décalées dans le temps, avec un effet de bord aux extrémités. Si maintenant on fait tendre $n \to \infty$ cet effet de bord disparaît et l'action de la matrice sur un vecteur est équivalente à l'action d'un système de convolution (linéaire et invariant dans le temps) sur le signal correspondant. On sait que l'action d'un tel système sur les signaux de la forme $(1, \exp(j\omega), \exp(2j\omega), \ldots, \exp(nj\omega))$ (c'est-à-dire $x_k = \exp(kj\omega)$) est simplement de les multiplier par $H(\omega)$ où $H(\omega)$ est la transformée de Fourier de la réponse impulsionnelle. Dans notre terminologie des matrices, cela veut dire que pour $n \to \infty$, les vecteurs de signaux $\boldsymbol{x}$ tels que $x_k = \exp kj\omega$ sont des vecteurs propres de valeur propres $H(\omega)$. L'ensemble des valeurs propres tend donc vers un ensemble continu. On peut montrer (voir es-

paces de Hilbert) que l'ensemble des vecteurs propres forme une base orthonormée d'un espace de signaux (ceux dont la transformée de Fourier existe) et "diagonalisent le système", ce qui se traduit par le fait que la transformée de Fourier d'un signal de sortie est le produit de la transformée de Fourier de l'entrée par la transformée de Fourier de la réponse impulsionnelle du système.

D.5.8 Matrices hermitiennes définies positives

Un matrice hermitienne est définie positive (h.d.p) si toutes ses valeurs propres sont strictement positives. Si aucune valeur propre n'est négative (certaines pouvant être nulles) nous dirons que la matrice est non-définie négative ou semi-definie positive. Si toutes les valeurs propres sont strictement négatives, nous dirons que la matrice est définie négative.

Montrons qu'une matrice est h.d.p. si et seulement si $\forall \boldsymbol{x} \neq 0$,

$$\boldsymbol{x}^* \boldsymbol{H} \boldsymbol{x} \text{ est réel et strictement positif} \tag{D.105}$$

Montrons que $\boldsymbol{H}$ est hermitienne. En effet, puisque $\boldsymbol{x}^* \boldsymbol{H} \boldsymbol{x}$ est réel, on a (en prenant deux vecteurs de la base)

$$\overline{\boldsymbol{x}^*(\boldsymbol{H}\boldsymbol{x})} = \boldsymbol{x}^T \overline{\boldsymbol{H}} \overline{\boldsymbol{x}} = (\boldsymbol{H}^* \boldsymbol{x})^T \overline{\boldsymbol{x}} = \boldsymbol{x}^*(\boldsymbol{H}^* \boldsymbol{x}). \tag{D.106}$$

Donc $\boldsymbol{x}^*((\boldsymbol{H} - \boldsymbol{H}^*)\boldsymbol{x}) = 0$ pour tout $\boldsymbol{x}$. Donc, pour tous $\boldsymbol{x}$ et $\boldsymbol{y}$ on a

$$\begin{aligned}
4\boldsymbol{x}^*((\boldsymbol{H} - \boldsymbol{H}^*)\boldsymbol{y}) &= (\boldsymbol{x}+\boldsymbol{y})^*(\boldsymbol{H} - \boldsymbol{H}^*)(\boldsymbol{x}+\boldsymbol{y}) - && \text{(D.107)}\\
&\quad (\boldsymbol{x}-\boldsymbol{y})^*(\boldsymbol{H} - \boldsymbol{H}^*)(\boldsymbol{x}-\boldsymbol{y}) + && \text{(D.108)}\\
&\quad i(\boldsymbol{x}+i\boldsymbol{y})^*(\boldsymbol{H} - \boldsymbol{H}^*)(\boldsymbol{x}+i\boldsymbol{y}) + && \text{(D.109)}\\
&\quad -i(\boldsymbol{x}-i\boldsymbol{y})^*(\boldsymbol{H} - \boldsymbol{H}^*)(\boldsymbol{x}-i\boldsymbol{y}) && \text{(D.110)}\\
&= 0. && \text{(D.111)}
\end{aligned}$$

En prenant $\boldsymbol{x} = (\boldsymbol{H} - \boldsymbol{H}^*)\boldsymbol{y}$ on en déduit que

$$|(\boldsymbol{H} - \boldsymbol{H}^*)\boldsymbol{y}|^2 = \boldsymbol{0} \tag{D.112}$$

pour tout $\boldsymbol{y}$ et donc la matrice $(\boldsymbol{H} - \boldsymbol{H}^*)$ doit être nulle. Puisque la matrice est hermitienne elle dispose de valeurs propres réelles. Montrons qu'elles sont strictement positives. Soit λ et un vecteur propre (forcément non nul) de celle-ci $\boldsymbol{x}$. On a

$$\boldsymbol{x}^* \boldsymbol{H} \boldsymbol{x} = \lambda |\boldsymbol{x}|^2 > 0 \Rightarrow \lambda > 0. \tag{D.113}$$

La réciproque est presque immédiate : si $\boldsymbol{H}$ est h.d.p. alors elle est diagonalisable par une unitaire $\boldsymbol{U}$. Donc $\forall \boldsymbol{x}$ non nul le vecteur $\boldsymbol{y} = \boldsymbol{U}^* \boldsymbol{x}$ est non nul. Et,

$$\boldsymbol{x}^* \boldsymbol{H} \boldsymbol{x} = (\boldsymbol{U}\boldsymbol{y})^* \boldsymbol{H} (\boldsymbol{U}\boldsymbol{y}) = \boldsymbol{y}^* (\boldsymbol{U}^* \boldsymbol{H} \boldsymbol{U}) \boldsymbol{y} = \sum_{i=1} \lambda_i |y_i|^2 > 0. \tag{D.114}$$

On peut alors se convaincre que pour toute matrice h.d.p., la forme

$$\boldsymbol{x}^* \boldsymbol{H} \boldsymbol{y} \tag{D.115}$$

définit un produit scalaire et de par là induit une norme et une distance sur l'espace des vecteurs complexes. Similairement, toute matrice réelle d'ordre p symétrique définie positive définit un produit scalaire, une norme et une distance sur $\mathbb{R}^p$.

Exemples. Les matrices qui s'écrivent $\boldsymbol{A}^* \boldsymbol{A}$ sont h.d.p. pour autant que leur rang soit maximal. Sinon, elles sont s.d.p.

E STRUCTURES ALGEBRIQUES DISCRETES

E.1 INTRODUCTION

Le but de cet appendice est de collationner les notions de structures algébriques classiquement utilisées en mathématique. La plupart de ces notions sont connues par les étudiants, dans le cadre des corps $\mathbb{R}$ (et C) et des espaces linéaires $\mathbb{R}^p$ (et C^p) (voir notamment appendices D et F.).

Plus spécifiquement, notre but est de discuter ces notions dans le cadre des ensembles finis (structures discrètes), ce qui représente le contexte dans lequel ces notions seront utilisées en théorie de l'information (codage de canal).

Nous introduirons donc tout d'abord de manière abstraite les notions de groupe, anneau et corps.

Ces notions seront alors illustrées d'une part sur les exemples classiques ($\mathbb{R}$ et C), puis, de manière plus détaillée, sur les ensembles finis.

Ensuite, nous discuterons ce que deviennent les stuctures d'espace vectoriel lorsqu'elles sont définies sur des ensembles finis et des corps finis. Notre but est ici de nous convaincre que les notions telles qu'indépendance linéaire, base vectorielle et orthogonalité survivent avec la plupart des propriétés intuitives dont elles font preuve dans les espaces $\mathbb{R}^p$.

Enfin, l'appendice fournit également une introduction aux corps de Gallois, ce qui représente une matière plus ardue mais largement utilisée (dans le cadre de la théorie des codes linéaires notamment).

Note. La matière sur les corps de Gallois ne fait pas partie de la version distribuée pour l'année académique 1998-1999.

E.2 GROUPES COMMUTATIFS

E.2.1 Structure de groupe commutatif

Soit un ensemble G et une opération interne désignée par $+$ (une fonction de $G \times G \rightarrow G$, qui associe à $x, y \in G$ un élément de G désigné par $x + y$). Cette opération définit une structure de groupe commutatif si elle vérifie les axiomes suivants

1. Associativité : $x + (y + z) = (x + y) + z, \forall x, y, z \in G$.

2. Commutativité : $x + y = y + x, \forall x, y \in G$.

3. Elément neutre : $\exists 0 \in G | \forall x \in G \ : x + 0 = 0 + x = x$.

4. Elements opposés : $\forall x \in G \mid \exists - x \in G \ : x + (-x) = (-x) + x = 0$.

Remarques.

1. Souvent on note l'opération binaire par la juxtaposition (xy) et dans ce cas l'élément neutre est désigné par 1 *et l'opposé par* x^{-1} *(appelé* inverse*).*

2. Nous désignerons l'élement neutre par 0 ou 1 selon qu'il s'agit d'un groupe additif ou multiplicatif (voir ci-dessous).

3. Bien qu'un groupe soit défini par l'ensemble G et l'opération $+$, nous désignerons souvent le groupe par G (plutôt que $(G, +)$) lorsqu'il n'y a pas de risque de confusion sur l'opération $+$.

On utilise le terme de groupe si le deuxième axiome n'est pas vérifié. S'il est vérifié on utilise indifféremmment le terme de groupe commutatif ou de groupe Abélien. Dans la suite nous ne considérons que les groupes commutatifs, sauf mention explicite du contraire.

Exemples.

1. L'ensemble $\mathbb{R}$ muni de l'addition est un groupe commutatif. De même, $\mathbb{R}\backslash\{0\}$ muni de la multiplication est un groupe commutatif.

2. L'ensemble Z (entiers positifs et négatifs) muni de l'addition est un groupe commutatif. L'ensemble $\mathbb{N}$ muni de l'addition (ou de la multiplication) entière n'est pas un groupe (non-existence des élements opposés).

3. L'ensemble $\{0, 1\}$ muni de l'opération d'addition *modulo* 2 (notation $\oplus_2$), définie par la table

$\oplus_2$	0	1
0	0	1
1	1	0

est un groupe commutatif (élément neutre 0). L'opposé de 1 est 1 lui-même. Ce groupe est souvent désigné par Z_2.

4. Plus généralement, soit p un nombre entier positif ≥ 2. Alors, l'ensemble $Z_p = \{0, 1, \ldots, p-1\}$ muni de l'opération d'addition *modulo* p (notation $\oplus_p$) définie par

$$x \oplus_p y = \begin{cases} x + y & \text{si } x + y < p \\ x + y - p & \text{si } x + y \geq p \end{cases}$$

est un groupe (élément neutre 0, opposé : $x \neq 0 \Rightarrow -x \bmod p = p - x$). Par souci de concision nous noterons cette opération par $\oplus$, et nous désignerons ce groupe par $(Z_p, \oplus)$. De toute évidence l'exemple 3 est le cas particulier pour $p = 2$.

E.2.2 Sous-groupes

Si $(G, +)$ est un groupe, alors $K \subset G$ défini un sous-groupe $(K, +)$ s'il satisfait aux conditions suivantes

1. Si $x, y \in K$, alors $x + y \in K$ (fermeture).

2. Si $x \in K$, alors $-x \in K$.

3. (par conséquent), $0 \in K$.

Exemples.

1. Dans $(Z, +)$ l'ensemble de tous les entiers multiples d'un nombre entier positif p est un sous-groupe; on le désigne par $(pZ, +)$. En particulier, l'ensemble des nombres pairs $(2Z, +)$ est un sous-groupe.

2. Si K_1 et K_2 sont deux sous-groupes, alors l'ensemble $K_1 + K_2$ défini par $K_1 + K_2 = \{x = x_1 + x_2 | x_1 \in K_1, x_1 \in K_2\}$ est encore un sous-groupe.

3. Le singleton $\{0\}$ qui ne contient que l'élement neutre est un sous-groupe. Il est dit trivial. De même, G lui-même est un sous-groupe. Il est dit non propre.

E.2.3 Cosets

Soit $x \in G$ et K un sous-groupe. On appelle *coset* de x modulo K, le sous-ensemble de G noté $x + K$ défini par

$$x + K = \{x + k | k \in K\}. \tag{E.1}$$

Notons que le coset d'un élément non-nul n'est pas un groupe. Dans $(Z, +)$ l'ensemble des nombres impairs est le coset de 1 modulo Z_2 (le sous-groupe des nombres pairs).

Proposition.

L'ensemble des cosets modulo un sous-groupe K quelconque possède les propriétés suivantes

1. *Tout élement de G appartient à au moins un coset.*
2. *Si deux cosets ont un élément commun, ils sont identiques.*
3. *Deux éléments x et y appartiennent au même coset si et seulement si $x - y \in K$.*
4. *Si le sous-groupe K possède r élements, alors chaque coset possède r élements.*

Par conséquent, un sous-groupe K définit par l'intermédiaire de ses cosets une partition de G en classes d'équivalence qui, lorsque K est fini, ont toutes le même nombre d'éléments.

Notons que si G est lui-même fini, alors K doit l'être aussi. Par conséquent, lorsque G est fini et de taille R, alors tout sous-groupe doit avoir un nombre d'éléments r tel que $rk = R$, pour un entier k fini.

Il s'en suit également que si R est un nombre premier, alors G n'admet que deux sous-groupes : G lui-même, et le sous-groupe trivial. Un tel groupe est dit *primitif* ou *irréductible*.

Par exemple, dans $(Z, +)$ le sous-groupe des nombres pairs définit deux cosets : l'ensemble des nombres pairs et l'ensemble des nombres impairs. Ces classes ne sont pas finies, mais elles ont même cardinalité.

E.2.4 Factorisation

Soit G un groupe et K un sous-groupe propre. Dans chaque coset de K, choisissons (arbitrairement) un élément appelé *tête* de coset. Désignons alors par G/K l'ensemble formé par les têtes de tous les cosets de K ainsi choisis, et définissons sur cet ensemble l'opération suivante : $x \diamond y$ [1] est la tête du coset $x + y + K$.

Alors il est facile de démontrer que la structure $(G/K, \diamond)$ définit un nouveau groupe commutatif. L'élément neutre de ce groupe est la tête du coset $0 + K$.

De façon concrète, ce groupe dépend bien sûr du choix effectué pour les têtes de coset. Cependant, on montre que deux groupes quelconques ainsi construits à partir du même sous-groupe K sont isomorphes. cela veut dire qu'ils peuvent être mis en bijection en respectant l'opération interne : image de la somme = somme des images, l'image de l'élement neutre est l'élément neutre et l'image de l'opposé est l'opposé de l'image.

Par conséquent, la factorisation d'un groupe par un sous-groupe définit essentiellement une seule nouvelle structure de groupe, qu'on note en général $(G/K, +)$ ou simplement G/K.

Exemples.

1. $(Z/2Z, +)$ désigne la factorisation de $(Z, +)$ par le sous-groupe des nombres pairs. Ce groupe est composé de deux éléments (les têtes des cosets des nombres pairs et impairs respectivement). Ce nouveau groupe est isomorphe au groupe $(Z_2, \oplus_2)$.

2. Plus généralement, pour un entier possitif p quelconque, la factorisation de $(Z, +)$ par le sous-groupe $(pZ, +)$ est un groupe isomorphe à Z_p : $Z/pZ = Z_p$.

E.2.5 Congruence modulo p

On dit que deux entiers sont congruents (ou égaux) *modulo* p si leur différence est un multiple entier de p, c'est-à-dire si ces nombres appartiennent au même coset de pZ.

On note cela par

$$x \equiv y,$$

en laissant implicite la valeur de p. Il est clair que cette relation est une relation d'équivalence (symétrique, réflexive et transitive). Elle définit par conséquent une partition de Z. Cette partition est évidemment la même que celle définie par les cosets modulo pZ. Elle partitionne Z en p classes disjointes de même taille.

E.2.6 Produit cartésien de groupes

Soient $(G_1, +_1)$ et $(G_2, +_2)$ deux groupes. Considérons, la structure $(G, +)$ définie par le produit cartésien $G = G_1 \times G_2$, c'est-à-dire l'ensemble $\{(x_1, x_2) \,|x_1 \in G_1, x_2 \in G_2\}$ muni d'une opération d'addition $+$ induite à partir des opérations $+_1$ et $+_2$ par la règle

$$(x_1, x_2) + (y_1, y_2) = (x_1 +_1 y_1, x_2 +_2 y_2).$$

On peut se convaincre que cette opération répond bien à tous les desiderata pour définir une structure de groupe. L'élément neutre de G est $(0_1, 0_2)$, c'est-à-dire la paire définie par les éléments neutres des groupes G_1 et G_2 de départ.

De toute évidence le groupe produit $G_2 \times G_1$ est isomorphe au précédent et il n'y a donc pas de raison de distinguer ces deux objets (on dira que le produit cartésien de groupes est commutatif).

Par induction, on définit alors le produit cartésien d'un nombre quelconque de groupes, et en particulier des groupes de la forme $G^m = G \times \cdots \times G$ définis à partir d'un seul groupe de départ. Les élements de ces groupes sont désignés par le terme de vecteur à m dimensions. Nous reviendrons sur ce type de structure un peu plus loin.

E.3 ANNEAUX ET CORPS

Ces structures sont définies à l'aide de deux opérations internes qu'on désigne habituellement par les termes d'addition $+$ et de multiplication ($\cdot$ ou juxtaposition). Il s'agit de structures algébriques essentiellement décalquées sur le corps des nombres réels.

Notre objectif est ici d'énoncer les propriétés axiomatiques de ces structures, puis de discuter plus en détails le cas de ces structures lorsqu'elles sont définies sur un ensemble *fini*. Elles jouent un rôle très important dans le cadre de la théorie des codes correcteurs d'erreurs.

E.3.1 Corps

Un corps est un ensemble F (terme anglais *field*) muni de deux opérations binaires $+$ et $\cdot$ telles que

1. $(F, +)$ est un groupe commutatif avec l'élement neutre désigné par 0.
2. $(F\backslash\{0\}, \cdot)$ est un groupe commutatif (élément neutre désigné par 1).
3. $\cdot$ est distributive par rapport à $+$: $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$ sur F.

On voit que toute la cohérence entre addition et multiplication est assurée par la simple propriété de distributivité.

Ce type de structure répond aux axiomes suivants

1. associativité pour les deux opération internes
2. commutativité pour les deux opération internes
3. distributivité
4. existence d'élements neutres 0 et 1 pour les deux opérations internes
5. existence d'opposés (par rapport à $+$)
6. existence d'inverses (par rapport à $\cdot$, et sauf pour 0)

E.3.2 Anneaux

Comme de nombreuses structures algébriques répondent à tous ces axiomes sauf au dernier, on réserve le terme d'anneau à ces structures. Un corps est donc un anneau tel que tout élement non nul dispose d'un inverse.

E.3.3 Quelques propriétés élémentaires

1. Dans un corps on peut définir les opérations de soustraction (resp. de division) par l'addition de l'opposé (resp. la multiplication par l'inverse).

2. Comme $(F\backslash\{0\},\cdot)$ est un groupe, cela implique que le produit de deux éléments non-nuls est un élément non-nul. Cependant, la structure ne définit pas explicitement la multiplication par 0. Montrons que la seule façon cohérente d'étendre la multiplication au cas où l'un des deux termes est nul est $x \cdot 0 = 0 \cdot x = 0$. En effet, la distributivité implique que $x(0+y) = x0 + xy, \forall x, y$; par ailleurs comme $0 + y = y$ on doit donc avoir $xy = x0 + xy$ ce qui implique que $x0 = 0$. Une démonstration identique implique aussi que $0x = 0$, quelque soit x non-nul. Si $0x = 0$ alors l'associativité implique que $0(0+y) = 0 \cdot 0 + 0 \cdot y = 0$, et donc que $0 \cdot 0 = 0$. Inversement, ayant ainsi étendu l'opération de multiplication à 0, on en déduit que $xy = 0 \Rightarrow x = 0 \vee y = 0$.

3. Simplification : $[a \neq 0] \wedge [ax = ay] \Rightarrow x = y$. En effet, $ax = ay \Rightarrow a(x-y) = 0 \Rightarrow [x = y] \vee [a = 0]$.

4. $-x = (-1) \cdot x$, car $(1 + (-1)) \cdot x = x + (-1) \cdot x = 0 \cdot x = 0$.

Exemples.

1. $\mathbb{R}$ muni des opérations usuelles forme un corps, de même que C.

2. L'ensemble $\{0, 1\}$ muni de l'addition modulo 2 et de la multiplication $1 \cdot 1 = 1$ (le groupe $(F\backslash\{0\},\cdot)$ est trivial) forme un corps. On a évidemment $0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0$.

3. L'ensemble Z ne forme pas un corps (les deux seuls entiers qui admettent une inverse sont 1 et -1). Il s'agit cependant d'un anneau.

E.3.4 Les anneaux et corps Z_p

Soit un anneau F et p un élément non-nul de F. Nous pouvons construire l'ensemble $F/pF = F_p$ constitué par les têtes des cosets de pF (les multiples de p).

Comment étendre les opérations d'addition et de multiplication de la manière suivante : soient x, y deux éléments de F_p (c'est-à-dire deux têtes (choisies arbitrairement) de deux cosets de pF), alors

$$x + y \qquad \text{est défini comme la tête du coset } x + y + pF \tag{E.2}$$

$$xy \qquad \text{est défini comme la tête du coset } xy + pF \tag{E.3}$$

Il est aisé de se convaincre que cette structure jouit des propriétés requises pour ête un anneau.

Exemple.

Dans Z_p l'addition est définie par l'addition *modulo* p ($\oplus_p$) et de même pour la multiplication (notée $\otimes_p$). C'est-à-dire que si nous choissons les entiers $\{0, 1, \ldots, p-1\}$ comme têtes de cosets, alors $x \otimes_p y$ dans Z_p est obtenu par xy modulo p dans Z. On voit bien que cette définition préserve la distributivité : $x \oplus_p x = 2 \otimes_p x$, $x \oplus_p x \oplus_p x = 3 \otimes_p x \ldots$

Théorème.

Si, et seulement si p est un nombre premier alors Z_p est un corps.

Preuve.

Pour montrer que la condition est suffisante, il suffit de se convaincre que chaque élément non-nul de Z_p dispose d'une inverse lorsque p est premier. Nous allons procéder par induction (le cas de base est trivial, car $1^{-1} = 1$) : supposons avoir la preuve que les éléments $1, 2, \ldots, i-1$ disposent des inverses $1^{-1}, 2^{-1}, \ldots (i-1)^{-1}$ montrons que i doit alors aussi avoir une inverse.

Il suffit d'effectuer la division entière de p par i : on a $p - r = iq$, qui implique $-r = i \otimes q$ (dans Z_p). Comme r est plus petit que i (c'est le reste de la division de p par i) et ne peut être nul (car p est un nombre premier) son inverse doit exister par hypothèse. Soit alors $x = (-1) \otimes q \otimes r^{-1}$. On trouve

$$i \otimes x = i \otimes (-1) \otimes q \otimes r^{-1} = (-1) \otimes (i \otimes q) \otimes r^{-1} = r \otimes r^{-1} = 1.$$

A contrario, si p n'est pas premier alors Z_p ne peut pas être un corps : en effet comme p est non premier il peut s'écrire sous la forme $p = r_1 r_2$ (avec $0 < r_1, r_2 < p$). Mais alors $r_1 \otimes_p r_2 = 0$, ce qui est incompatible avec la structure de corps, sauf si $r_1 = 0$ ou $r_2 = 0$ ce qui contredit notre hypothèse.

Nous sommes donc équipés des corps Z_2, Z_3, Z_5.... Dans la suite, nous allons construire des corps ayant p^k éléments avec p un nombre premier et k un entier positif quelconque (vu ce qui vient d'être dit, ces corps sont différents (non isomorphes à) de Z_{p^k}), sauf si $k = 1$. Il s'agit des corps de Gallois discutés à la section E.5.

Par contre, il n'existe pas de corps fini ayant six éléments. En fait on peut démontrer qu'un corps fini doit nécessairement comporter un nombre d'éléments égal à une puissance entière d'un nombre premier. On montre enfin que les seuls corps finis sont les corps de Gallois.

La structure de corps fini impose donc des restrictions sur le choix de la taille de F.

E.4 ESPACES LINEAIRES

Soit $(F, \oplus, \otimes)$ un corps (dont les éléments sont appelés *scalaires*) et soit L un ensemble (dont les éléments sont appelés *vecteurs*) muni de deux opérations $+$ (interne) et $\cdot$ (externe : multiplication par un scalaire). Cette structure conjointe définit un espace linéaire (les élements de L en sont les vecteurs) si

1. $(L, +)$ est un groupe
2. $t \cdot x$ est un vecteur défini pour tout scalaire t et tout vecteur x.
3. Associativité : $(s \otimes t) \cdot x = s \cdot (t \cdot x)$.
4. Distributivité : $t \cdot (x + y) = (t \cdot x) + (t \cdot y)$ et $(s \oplus t)x = s \cdot x + t \cdot x$.
5. $1 \cdot x = x$ pour tout vecteur x.

De ces axiomes on peut déduire les conséquences suivantes :

1. $0 \cdot x = \mathbf{0}$ (le produit d'un vecteur par le scalaire nul donne le vecteur nul).
2. $(-1) \cdot x = -x$;
3. $t \cdot \mathbf{0} = \mathbf{0}$.

Exemples.

1. F^n (l'ensemble de tous les n-tuples de scalaires du corps F) et muni des opérations $+$ et $\cdot$ induites par celles de F comme suit

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 \oplus y_1, \ldots, x_n \oplus y_n) \quad \text{(E.4)}$$

$$t \cdot (x_1, \ldots, x_n) = (t \otimes x_1, \ldots, t \otimes x_n) \quad \text{(E.5)}$$

est un espace vectoriel linéaire. On convient dans ce cas d'utiliser les mêmes symboles pour désigner les opérations vectorielles et celles sur le corps des scalaires.

2. $\mathbb{R}^p$ est de cette manière un espace vectoriel linéaire défini sur $\mathbb{R}$.

3. Z_p^n est de cette manière un espace vectoriel linéaire défini sur Z_p (p étant premier). Cet espace définit en réalité l'ensemble des mots construits sur un alphabet de p symboles, ce dernier étant muni d'une structure de corps. En particulier, lorsque $p = 2$ on parle de mots binaires de longueur n.

E.4.1 Sous-espaces linéaires

Un sous-ensemble K d'un espace linéaire L est dit *sous-espace* si toute combinaison linéaire de vecteurs de K appartient encore à K (on peut avoir $K = L$ avec cette définition). En particulier, pour tout espace le singleton $\{\mathbf{0}\}$ forme un sous-espace (dit trivial). Ce sous-espace est aussi dit minimal, car il doit évidemment être inclus dans tout autres sous-espace de L.

Les sous-espaces linéaires de Z_2^n forment ce qu'on appellera l'ensemble des codes (binaires) linéaires de longueur n. Voyons comment on peut construire de tels codes.

Espace engendré.

Soit $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ un ensemble de vecteurs. Alors, l'ensemble des combinaisons linéaires de ces vecteurs est nécessairement un sous-espace linéaire de L. Il s'agit du plus petit sous-espace linéaire qui contient ces vecteurs. On dit que les vecteurs engendrent ce sous-espace.

Indépendance linéaire et base vectorielle.

Les vecteurs $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ sont dits (par définition) linéairement indépendants si aucun d'entre-eux ne peut s'exprimer sous la forme d'une combinaison linéaire des autres.

Les vecteurs $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ forment une base d'un sous-espace linéaire s'ils engendrent cet espace et sont linéairement indépendants. En particulier, ils forment une base de L s'ils sont linéairement indépendants et engendrent L. S'il existe une base avec un nombre fini de vecteurs alors l'espace est dit de dimension finie (voir appendice F.4 pour une discussion plus fine de ce concept).

Les espaces F^n définis à partir d'un corps F sont tous de dimension égale à n. Ils sont en effet engendrés par les vecteurs e_i (composantes $\delta_{i,j}$), quel que soit le corps F et ces vecteurs sont linéairement indépendants quel que soit le corps F.

On montre que si L est défini sur un corps fini, alors il est de dimension finie si et seulement s'il est fini. (C'est le cas pour F^n si F est fini).

On peut d'ailleurs mettre un espace de dimension finie n défini sur le corps F en bijection avec F^n. En effet, soit L cet espace et $\{e_1, \ldots, e_n\}$ une base de L. Alors, il est facile de montrer que $\forall x \in L$ il existe une combinaison linéaire unique

$$x = \sum_{i=1}^{n} t_i e_i.$$

Inversement, pour tout choix des t_i, on a $\sum_{i=1}^{n} t_i e_i \in L$ et les t_i identifient un vecteur unique (linéarité). Donc la donnée d'un vecteur est équivalente à la donnée d'un n-tuple de scalaires de F. Dans cette correspondance bijective, l'addition vectorielle se traduit par l'addition dans F^n et la multiplication par un scalaire à celle sur F^n. Si R est le nombre d'éléments de F, le nombre d'élements de F^n et donc de vecteurs de L vaut R^n.

La morale de cette histoire est que si nous nous focalisons sur les espaces de dimension finie définis sur un corps F, nous pouvons aussi bien limiter notre intérêt aux espaces F^n.

Dimension.

Soit L un espace de dimension k (il dispose d'un base composée de k vecteurs). Alors

1. Chaque base de L dispose exactement de k vecteurs.
2. Chaque ensemble composé de k vecteurs linéairement indépendants forme une base.
3. k est le nombre maximal de vecteurs linéairement indépendants.
4. Chaque sous-espace propre de L a une dimension strictement inférieure à k (ce qui implique aussi que si L' est un sous-espace de L de dimension k, il doit être identique à L).

E.4.2 Orthogonalité

Nous restreignons notre étude aux espaces F^n définis sur et à partir d'un corps fini $(F, \oplus, \otimes)$. Essayons de voir si la notion d'orthogonalité peut encore être utilisée comme usuellement.

Produit scalaire.

Etant donnés deux vecteurs $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in F^n$ nous définissons leur produit scalaire par :

$$\langle x, y \rangle = x_1 \otimes y_1 \oplus \cdots \oplus x_n \otimes y_n.$$

On dit que deux vecteurs sont orthogonaux si leur produit scalaire est nul.

Ce produit scalaire définit donc une application de $F^n \times F^n$ dans F, et vérifie (c'est assez évident) les propriétés suivantes

1. $\langle \mathbf{0}, y \rangle = 0$.
2. $\langle x, y \rangle = \langle y, x \rangle, \forall x, y \in F^n$.
3. $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle, \forall x, y, z \in F^n$.

4. $\langle x, t \cdot y\rangle = t \otimes \langle x, y\rangle, \forall x, y \in F^n, \forall t \in F.$

Malheureusement, nous n'avons pas en général

$$\langle x, x\rangle = 0 \Rightarrow [x = \mathbf{0}],$$

et le produit scalaire ne définit donc pas nécessairement une norme sur F^n. D'ailleurs, la structure de corps fini est d'une certaine manière incompatible avec la structure d'ordre qui est nécessaire à la définition d'une norme.

Par exemple, bien qu'on puisse définir une structure d'ordre sur un corps fini (p.ex. sur Z_p on peut utiliser l'ordre "naturel"), il n'est pas possible de définir une structure d'ordre qui reste compatible (au sens usuel) avec les opérations d'addition et de multiplication. Il faudrait pour cela que (par exemple)

$$x, y > 0 \Rightarrow x + y > 0,$$

et

$$x > y \Rightarrow x - y > 0,$$

quels que soient $x, y \in F$.

Afin de bien clarifier cette distinction qu'il faut faire par rapport à la notion usuelle de produit scalaire prenons comme exemple Z_2^2, et prenons deux vecteurs de $x, y \in Z_2^2$. On trouve que

$$\begin{aligned}
\langle (0,0),(0,0)\rangle &= 0 && \text{(E.6)}\\
\langle (0,0),(0,1)\rangle &= 0 && \text{(E.7)}\\
\langle (0,0),(1,0)\rangle &= 0 && \text{(E.8)}\\
\langle (0,0),(1,1)\rangle &= 0 && \text{(E.9)}\\
\langle (0,1),(0,1)\rangle &= 1 && \text{(E.10)}\\
\langle (0,1),(1,0)\rangle &= 0 && \text{(E.11)}\\
\langle (0,1),(1,1)\rangle &= 1 && \text{(E.12)}\\
\langle (1,0),(1,0)\rangle &= 1 && \text{(E.13)}\\
\langle (1,0),(1,1)\rangle &= 1 && \text{(E.14)}\\
\langle (1,1),(1,1)\rangle &= 0 && \text{(E.15)}
\end{aligned}$$

On voit donc en particulier que certains vecteurs non-nuls peuvent être orthogonaux à eux-mêmes, ce qui est impossible lorsque $F = \mathbb{R}$.

Complément orthogonal.

Soit L un sous-espace linéaire de F^n. Le complément orthogonal de L (noté $L^\perp$) est par définition, l'ensemble des vecteurs de F^n orthogonaux à tous les vecteurs de L.

Par exemple, dans Z_2^2 considérons l'ensemble $L = \{(0,0),(1,1)\}$. Cet ensemble est un sous-espace linéaire, comme on peut le vérifier aisément. Il est identique (!) à son complément orthogonal.

Proposition.

On a néanmoins le résultat important suivant pour un sous-espace L de F^n :

1. *$L^\perp$ est lui-même un sous-espace linéaire.*

2. *Si x est orthogonaux à chacun des vecteurs d'une base de L, alors il figure dans $L^\perp$, ce qui veut dire en fait que[2] quelle que soit la base de L, $L^\perp$ se réduit exactement aux vecteurs qui sont orthogonaux à tous les vecteurs de cette base.*

3. *Si k est la dimension de L, $n - k$ est la dimension de $L^\perp$.*

4. *Corollairement, $(L^\perp)^\perp = L$.*

E.4.3 Matrices

Nous introduisons ici les matrices non pas comme opérateurs mais parce que nous nous intéressons aux solutions de systèmes homogènes d'équations linéaires. Ce type de systèmes est utilisé notamment en théorie des codes linéaires et nous verrons que l'ensemble des solutions de tels systèmes forme un sous-espace linéaire (complément orthogonal de l'espace engendré par les vecteurs qui définissent les coefficients des différentes équations).

Un tel système peut s'écrire en abrégé sous la forme

$$\langle x, a_1 \rangle = 0 \quad \text{(E.16)}$$
$$\ldots \langle x, a_m \rangle = 0 \quad \text{(E.17)}$$

où les a_i sont des vecteurs de F^n et x désigne l'inconnue.

Soit k_i l'indice du premier élément non-nul du vecteur a_i. Nous pouvons évidemment réarranger ces équations de façon à ce que les vecteurs a_i soient triés par ordre croissant de k_i. Dans ce cas, les vecteurs a_k nuls viennent en dernière position, et peuvent être éliminés. Comme l'ensemble de solutions ne change pas lorsque nous ajoutons (ou retranchons) à un des vecteurs a_i un multiple d'un des autres vecteurs, on peut s'arranger pour que les indices k_i soient en ordre strictement décroissant. La propriété principale de ce réarrangement est de conduire à un ensemble de vecteurs linéairement indépendants.

Le rang du système est par définition le nombre de vecteurs linéairement indépendants de ce système. Il est évident que ce nombre ne varie pas lors des transformations indiquées ci-dessus. Donc le nombre de vecteurs (non nuls) qui subsistent est égal au rang du système, et les vecteurs solutions du système sont orthogonaux aux vecteurs qui subsistent : ils forment le complément orthogonal de l'espace engendré par les vecteurs du système.

Un système de rang m peut donc être représenté par une matrice de m vecteurs linéairement indépendants de F^n.

E.5 CORPS DE GALLOIS

Note. Cette section ne figure pas dans la version distribuée en 1998-99.

Notes

1. Nous utilisons un symbole différent pour l'opération interne définie sur G/K pour bien la distinguer de l'opération interne $+$ définie sur G.

2. Compte tenu de la définition de $L^{\perp}$.

F ESPACES LINEAIRES TOPOLOGIQUES

F.1 INTRODUCTION

Les deux sections suivantes couvrent une matière non enseignée en candidature ingénieur, à savoir des éléments de topologie et d'espaces vectoriels linéaires généraux.

Notre but est de fournir une introduction aux espaces de Hilbert, mais nous en profitons pour fournir au lecteur intéressé quelques éléments de topologie abstraite, théorie mathématique qui soutend les notions de convergence et de continuité de fonctions. Nous espérons ainsi mettre en évidence comment les notions géométriques et analytiques classiques dans $\mathbb{R}^p$ apparaissent comme des cas particuliers de notions plus générales.

F.2 ELEMENTS DE TOPOLOGIE

Nous indiquons tout d'abord la démarche générale qui conduit à la définition d'une topologie sur un ensemble quelconque, puis nous particulariserons aux espaces métriques. Nous encourageons le lecteur réellement intéressé par les questions abordées dans les sections qui suivent de consulter [Rom75].

Soit Ω un ensemble dit universel. Une topologie $\mathcal{T}$ sur Ω est un ensemble de parties (ou sous-ensembles[1]) dites *ouverts* de Ω (i.e. $\mathcal{T} = \{\dots V_\alpha \dots\} \subset 2^\Omega$) qui jouit des propriétés axiomatiques suivantes :

1. $\emptyset \in \mathcal{T}$,
2. $\Omega \in \mathcal{T}$,
3. Une union quelconque $\bigcup_\beta V_\beta$ d'ouverts est encore un ouvert.
4. Une intersection finie d'ouverts est encore un ouvert.

Le couple $(\Omega, \mathcal{T})$ est appelé espace topologique.

Nous verrons que si on remplaçait la condition 4. par "une intersection quelconque d'ouverts est encore un ouvert" on impose une restriction trop forte à l'ensemble $\mathcal{T}$, qui ne conduirait qu'à des topologies relativement inintéressantes (par exemple la topologie discrète décrite ci après).

Exemples.

1. Il est clair que 2^Ω (c'est-à-dire l'ensemble de toutes les parties de Ω) définit une topologie. Cette topologie s'appelle la topologie discrète.

2. La topologie formée des seuls ouverts Ω et $\emptyset$ est appelée topologie indiscrète.

3. Nous verrons ci-dessous que la notion de boule ouverte dans $\mathbb{R}^p$ au sens d'une certaine distance définit une topologie : les ouverts y sont les sous-ensembles de $\mathbb{R}^p$ qui contiennent au moins une boule ouverte centrée en chacun de leurs points.

F.2.1 Comparaison et construction de topologies

Il est clair que pour un ensemble universel donné, il est possible de définir plusieurs topologies (voire une infinité). Etant donné deux topologies $\mathcal{T}_1$ et $\mathcal{T}_2$ définies sur un même Ω, on dit que $\mathcal{T}_1$ est plus faible que $\mathcal{T}_2$ si $\mathcal{T}_1 \subset \mathcal{T}_2$. On voit que la topologie indiscrète est la plus faible et la topologie discrète est la plus forte. Cependant, cette relation ne fournit qu'un ordre partiel sur l'ensemble de topologies. Comme nous le verrons, plus une topologie est forte, plus elle structure fortement l'ensemble universel du point de vue des notions de convergence et de continuité.

Notion de base topologique. On appelle base topologique tout sous-ensemble $\mathcal{B} \subset \mathcal{T}$, tel que

$$\forall V \in \mathcal{T} \ : V = \bigcup_{\beta} B_\beta, \tag{F.1}$$

avec $B_\beta \in \mathcal{B}, \forall \beta$. Il s'agit donc d'un ensemble particulier d'ouverts qui permet de reconstruire tous les autres ouverts par union, non-nécessairement dénombrable.

Une sous-base topologique $\mathcal{S}$ de $\mathcal{T}$ est un sous-ensemble de $\mathcal{T}$, tel qu'il existe une base $\mathcal{B}$ qui puisse être obtenue par intersections finies de parties de $\mathcal{S}$.

L'idée derrière ces notions est de synthétiser avec un "petit" nombre d'ouverts la topologie. Nous verrons ci-dessous que dans un espace métrique, l'ensemble des boules ouvertes est une base.

Construction d'une topologie. A partir d'un ensemble quelconque de parties $\mathcal{A}$ de Ω, on peut construire la topologie la plus petite qui contienne tous les ensembles de $\mathcal{A}$ de la manière suivante. Il suffit d'ajouter à $\mathcal{A}$ les ensembles Ω et $\emptyset$ (s'ils n'y figuraient pas déjà), ensuite nous déclarons cet ensemble comme étant une sous-base.

$\mathcal{T}$ est alors obtenue en formant d'abord toutes les intersections finies des ensembles ainsi obtenus, ce qui donne une base, puis en formant toutes les unions des ensembles de cette base.

Par exemple, nous pouvons construire une topologie sur $\mathbb{R}$ en partant de l'ensemble de tous les intervalles ouverts $]a, b[$. Nous verrons ci-dessous qu'il s'agit de la topologie naturelle induite par la distance "module de la différence" dans $\mathbb{R}$.

Topologie relative. Soit $A \subset \Omega$. Alors l'ensemble $\mathcal{T}_A \triangleq \{V \cap A \ : V \in \mathcal{T}\}$ définit une topologie sur A. On parle de la topologie relative de A (ou projection de la topologie sur A).

Topologie produit. Soient $(\Omega, \mathcal{T})$ et $(\Omega', \mathcal{T}')$ deux espaces topologiques. Alors on appelle espace topologique produit $(\Omega \times \Omega', \mathcal{T} \times \mathcal{T}')$, où $\mathcal{T} \times \mathcal{T}'$ est l'ensemble composé de toutes les unions possibles d'ensembles qui peuvent s'écrire sous la forme $G = A \times B$, avec $A \in \mathcal{T}$ et $B \in \mathcal{T}'$.

Notons que cela ne veut pas dire que

$$\forall G \in \mathcal{T} \times \mathcal{T}' \ : \exists A \in \mathcal{T}, B \in \mathcal{T}' | G = A \times B.$$

Ces ensembles figurent bien dans la topologie produit, mais celle-ci contient bien d'autres ensembles qui ne peuvent pas s'écrire sous cette forme. La figure F.1 montre des exemples d'ouverts dans $\mathbb{R}^2$ obtenus à partir du produit cartésien de deux topologies sur $\mathbb{R}$ induites par les intervalles ouverts. On voit que les ouverts élémentaires G_i sont bien obtenus par produit cartésien du type $A_i \times B_j$. Mais, par exemple l'ouvert $G_1 \cup G_2 \cup G_3$ ne peut pas être mis sous cette forme.

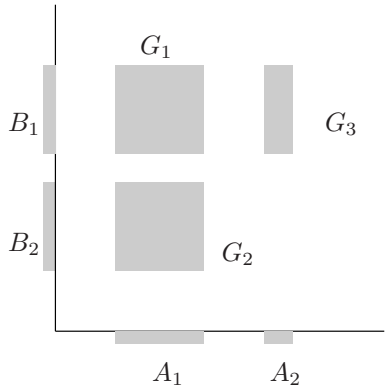

Figure F.1.

F.2.2 Voisinage d'un point

Un *voisinage* d'un point $\omega \in \Omega$ est un ouvert qui contient ce point. Les unions quelconques et les intersections finies de voisinages d'un même point sont encore des voisinages de ce point. Parfois on utilise une définition légèrement plus générale, en appelant voisinage tout ensemble qui contient un ouvert contenant le point.

On a la caractérisation importante suivante des ouverts : les ouverts sont les ensembles de points qui contiennent au moins un voisinage de chacun de leurs points. En effet, tout point d'un ouvert V est certainement inclus dans un tel voisinage (V lui-même). D'autre part, si un ensemble V possède la propriété alors on peut écrire

$$V = \bigcup_{\omega \in V} V(\omega), \tag{F.2}$$

où $V(\omega)$ désigne un voisinage de ω qui est inclus dans V. Comme il s'agit d'une union d'ouverts V doit être ouvert.□

F.2.3 Points isolés et points d'accumulation

Etant donné un sous-ensemble $A \subset \Omega$, un point $\omega \in A$ est dit un point isolé de A s'il existe un voisinage $V(\omega)$ de ω tel que $A \cap V(\omega) = \{\omega\}$.

A contrario, étant donné un sous-ensemble $A \subset \Omega$, un point $\omega \in \Omega$ est dit un point d'accumulation de A si tout voisinage V de ce point est tel que $(A \cap V(\omega)) - \{\omega\} \neq \emptyset$ (l'intersection comprend au moins un point différent de ω). Un point d'accumulation de A n'appartient pas nécessairement à A.

F.2.4 Ensembles fermés

L'ensemble F est dit fermé s'il contient tous ses points d'accumulation, ou de manière équivalente, s'il est le complémentaire d'un ensemble ouvert.

Montrons que ces deux caractérisations sont en effet équivalentes.

D'une part si V est ouvert, alors $\neg V$ doit contenir tous ses points d'accumulation. Sinon, supposons que ω soit un point d'accumulation de $\neg V$ qui appartienne à V; comme V est ouvert et que ω est un point d'accumulation de $\neg V$ on devrait avoir $V \cap \neg V \neq \emptyset$, ce qui est impossible.

D'autre part, soit F un ensemble qui contient tous ses points d'accumulation. $\neg F$ doit être ouvert car il doit contenir pour chacun de ces points au moins un voisinage. Sinon, cela voudrait dire qu'il existe un point de $\neg F$ tel que chacun de ses voisinages intersecte F, et ce point serait alors un point d'accumulation de F exclu de celui-ci.□

On a alors les propriétés duales suivantes pour les ensembles fermés : $\emptyset$ et Ω sont fermés, toute union finie de fermés est fermée, et toute intersection de fermés est fermée.

En particulier, on voit que certains ensembles peuvent être à la fois ouverts et fermés.

On montre qu'un ensemble fermé est exactement formé par ses points d'accumulation et ses points isolés.

Un ensemble fermé qui ne contient que ses points d'accumulation (et ne possède donc aucun point isolé) et dit parfait.

F.2.5 Intérieur, fermeture, frontière

L'intérieur A° d'un ensemble A est le plus grand ouvert inclus dans cet ensemble, ou de manière équivalente, l'union de tous les ouverts inclus dans cet ensemble. A est ouvert si, et seulement s'il est égal à son intérieur.

De façon duale, la fermeture $\bar{A}$ d'un ensemble est l'intersection de tous les fermés contenant cet ensemble. Il s'agit donc du plus petit fermé qui contienne notre ensemble. Un point appartient à la fermeture si, et seulement si tous ses voisinages coupent A. La fermeture de A est l'union de A et de ses points d'accumulation.

Enfin, la frontière $\dot{A}$ d'un ensemble est définie par

$$\dot{A} = \bar{A} - A^\circ; \tag{F.3}$$

il s'agit d'un ensemble fermé.

F.2.6 Convergence

Une des raisons principales de munir un ensemble d'une topologie est de pouvoir définir la notion de convergence, qui comme on le sait joue un rôle important aussi bien en théorie qu'en pratique.

Une suite ω_n d'un espace topologique est convergente vers la limite ω, si et seulement si

$$\forall V(\omega), \exists K > 0 \mid \forall n \geq K \ : \omega_n \in V(\omega). \tag{F.4}$$

On dit que la suite finit par entrer et rester dans tout voisinage de ω, et on le note par $\lim_{n\to\infty} \omega_n = \omega$. Une suite est dite convergente, s'il existe ω tel que la suite converge vers celui-ci. On peut dire aussi, de manière équivalente, que tout voisinage d'une limite contient presque tous les points de la suite (à l'exception d'un nombre fini).

De façon générale, il ne faut pas confondre la notion de limite d'une suite avec la notion de point d'accumulation. Mais, on peut dire que si la suite contient un nombre infini de points distincts et est convergente, alors une limite est un point d'accumulation de l'ensemble des points de la suite. En général, la limite n'est pas nécessairement unique.

F.2.7 Continuité

Soit $f(\cdot)$ une fonction d'un espace topologique $(\Omega, \mathcal{T})$ vers un espace topologique $(\Omega', \mathcal{T}')$. Cette fonction est dite continue (relativement aux deux topologies), si et seulement si

$$\forall V' \in \mathcal{T}' \ : f^{-1}(V') = \{\omega \in \Omega | f(\omega) \in V'\} \in \mathcal{T}, \tag{F.5}$$

en d'autres mots, si l'image inverse d'un ouvert est un ouvert.

On a la propriété importante suivante qui lie continuité et convergence :

L'image d'un suite convergente par une fonction continue est encore une suite convergente. La suite image converge vers l'image de toute limite de la suite de départ.

F.2.8 Types d'espaces topologiques

Espaces de Hausdorff. Un espace est dit de Hausdorff, si $\forall \omega_1 \neq \omega_2 \in \Omega \ : \exists V(\omega_1), V(\omega_2) | V(\omega_1) \cap V(\omega_2) = \emptyset$. En d'autres mots, dans un tel espace il est possible de séparer les points en les emballant par des voisinages.

Montrons que dans un espace de Hausdorff toute suite convergente a une limite unique.

En effet, supposons qu'il existe une suite ω_n et deux points limites distincts ω et ω'. Alors il existe deux voisinages $V(\omega)$ et $V(\omega')$ disjoints et il existe aussi $K(V(\omega))$ et $K'(V(\omega')$ tels que $\forall m \geq \max\{K, K'\} \ : \omega_n \in V(\omega) \cap V(\omega')$, ce qui est en contradiction avec le fait ces voisinages sont disjoints. □

En anticipant sur la suite, mentionnons que tout espace métrique est un espace de Hausdorff (voir ci-dessous).

Espace connexe. Un espace topologique est dit connexe s'il est impossible de le séparer en deux ouverts disjoints non vides; en d'autres mots, si

$$\forall V_1, V_2 \ : V_1 \cup V_2 = \Omega \Rightarrow V_1 \cap V_2 \neq \emptyset.$$

Densité et espaces séparables. Un ensemble $A \subset \Omega$ est dit dense si $\bar{A} = \Omega$. Un espace topologique est dit séparable si et seulement s'il contient un sous-ensemble dénombrable dense.

Par exemple, on sait que dans $\mathbb{R}$ tout nombre réel peut être obtenu (et est en fait défini par cette propriété) comme la limite d'une suite de nombre rationnels. Il s'en suit que la fermeture de l'ensemble des nombres rationnels est égale à $\mathbb{R}$. Comme les rationnels sont dénombrables $\mathbb{R}$ est séparable. Le produit cartésien d'ensembles dénombrables étant encore dénombable, les vecteurs de $\mathbb{R}^p$ à composantes rationnelles sont dénombrables, et ils permettent évidemment de construire les vecteurs de $\mathbb{R}^p$ par passage à la limite. Donc $\mathbb{R}^p$ est aussi séparable. Cette propriété reste encore vraie si nous considérons l'ensemble des suites définies sur $\mathbb{R}$.

Si un espace topologique contient une base topologique dénombrable, il est séparable.

En effet, soit B_i les ensembles de la base, et soit $\omega_i \in B_i$ une suite de points choisis dans ces ensembles. Cet ensemble doit être dense car, si nous donnons un voisinage (non vide) $V(\omega)$ d'un point quelconque de Ω, ce voisinage s'écrit nécessairement sous la forme d'une union de certains des B_i non-vides (propriété de la base). Soit alors $B_j \subset V(\omega)$ un de ceux-ci : on en déduit que $\omega_j \in V(\omega)$. Donc, tout voisinage de ω coupe l'ensemble $\{\ldots \omega_i \ldots\}$ et ω est donc un point d'accumulation de cet ensemble dénombrable. □

Espace et ensembles compacts. Un espace est dit compact, si tout ensemble d'ouverts qui couvre l'espace (dont l'union vaut l'espace) contient un sous-ensemble fini qui couvre encore l'espace.

Un ensemble $A \subset \Omega$ est dit compact, s'il est compact dans sa topologie relative.

Si un espace est compact, alors tout ensemble fermé l'est aussi automatiquement.

Dans un espace de Hausdorff tout ensemble compact est fermé.

Exemples.

1. Dans la topologie discrète tout ensemble est à la fois ouvert et fermé. Tout point d'un ensemble est un point isolé de cet ensemble. Aucun point n'est un point d'accumulation d'un ensemble quelconque A. Un espace muni de sa topologie discrète est : Hausdorff, non connexe, séparable seulement s'il est dénombrable.

2. Dans la topologie indiscrète, toute suite est convergente et tout point de Ω est une limite de toute suite. Cela confirme bien que la limite n'est pas nécessairement unique. Ce type d'espace : n'est pas de Hausdorff, est non connexe mais séparable (tout singleton est dense).

F.2.9 Espaces métriques et topologie naturelle

Un espace métrique est un espace muni d'une distance (une application $d(\cdot,\cdot)$ de $\Omega \times \Omega$ dans $\mathbb{R}$ qui vérifie les axiomes de la distance donnés dans l'appendice D). Une mesure de distance induit une topologie de la manière suivante.

On définit tout d'abord la notion de boule ouverte de rayon $\epsilon > 0$ centrée en un point ω par

$$S_\epsilon(\omega) = \{\omega' \in \Omega | d(\omega, \omega') < \epsilon\}. \tag{F.6}$$

Il s'agit donc des points de Ω ϵ-près de ω, au sens de la distance $d(\cdot,\cdot)$.

Soit V une partie d'un espace métrique $(\Omega, d(\cdot,\cdot))$. Alors V est un ensemble ouvert au sens de la topologie induite par $d(\cdot,\cdot)$ si, et seulement s'il contient une boule centrée en chacun de ses points. La topologie induite par $d(\cdot,\cdot)$ est l'ensemble des parties de Ω qui répondent à cette définition. En particulier, toute boule est un ouvert. On vérifie aisément que l'ensemble d'ouverts ainsi définis répond bien aux 4 axiomes de définition d'une topologie. On appelle cette topologie la topologie *naturelle* de l'espace métrique.

Inversément, on dit qu'un espace topologique est métrisable s'il existe une distance qui induit la topologie. Tous les espaces topologiques ne sont pas métrisables (par exemple, la topologie indiscrète n'est pas métrisable). Mais, si une topologie est métrisable alors il existe en fait une infinité de distances qui induisent cette topologie.

Une définition équivalente d'un ensemble ouvert est la suivante : *un ensemble est ouvert si et seulement s'il est une union (éventuellement infinie) de boules ouvertes.*

Théorème. *Tout espace métrique est un espace de* Hausdorff.

En effet, soient ω_1, ω_2 deux points distinct de cet espace. On a

$$\omega_1 \neq \omega_2 \Rightarrow d(\omega_1, \omega_2) = \eta > 0. \tag{F.7}$$

Soient alors les boules de rayon $S_{\frac{\eta}{3}}(\omega_1)$ et $S_{\frac{\eta}{3}}(\omega_2)$, deux ouverts qui contiennent respectivement ω_1 et ω_2. Ces deux boules ne peuvent pas avoir d'intersection, car si nous supposons que $\omega \in S_{\frac{\eta}{3}}(\omega_1) \cap S_{\frac{\eta}{3}}(\omega_2)$, on en déduit que

$$d(\omega_1, \omega_2) \leq d(\omega_1, \omega) + d(\omega, \omega_2) \leq \frac{2}{3}\eta, \tag{F.8}$$

ce qui contredit (F.7).□

Il s'en suit que dans tout espace métrique, les limites des suites convergentes sont uniques.

Notons que la réciproque de ce théorème est fausse, car il existe des espaces de Hausdorff non métrisables.

Dans un espace métrique tout ensemble fini est un ensemble fermé composé uniquement de points isolés.

Exemples.

1. $\mathbb{R}^p$ muni d'une distance induite par une des normes d'ordre $m \geq 1$ est un espace métrique.

2. Soit $[a, b]$ (avec a et b finis) un intervalle borné de $\mathbb{R}$ muni de sa topologie usuelle définie par la distance $|x-y|$. Alors on peut définir l'ensemble des fonctions à valeurs réelles (ou complexes) et continues sur $[a, b]$ par rapport à cette topologie, et le munir de la distance

$$d_\infty(f, g) = \sup_{x\in[a,b]} \{|f(x) - g(x)|\}, \tag{F.9}$$

dont on peut se convaincre qu'il s'agit bien d'une distance. On désigne cet espace par $C[a, b]$. Par exemple, la suite de fonctions $f_n(x) = x/n$ est convergente au sens de la topologie induite par cette distance, et converge dans cet espace vers la fonction constante $f(x) = 0$. Nous en profitons pour insister ici sur la différence entre la notion de convergence dans un espace fonctionnel (dont nous venons de donner un exemple) et la convergence ponctuelle. La convergence ponctuelle signifie que $\forall x \in [a, b]$, la suite $f_n(x)$ de $\mathbb{R}$ converge, i.e.

$$\forall x \in [a, b]\ \exists f(x) | \forall \epsilon \exists K(x, \epsilon)\ : n \geq K(x, \epsilon) \Rightarrow |f_n(x) - f(x)| \leq \epsilon. \tag{F.10}$$

L'important ici est de noter que K peut être fonction de x, et que rien ne garantit en principe que les valeurs $f(x)$ caractérisent une fonction continue sur $[a, b]$. Il est d'ailleurs facile de construire des exemples de fonctions qui convergent ainsi ponctuellement vers une fonction discontinue. Par exemple la suite x^n converge ponctuellement sur $[0, 1]$ vers la fonction $f(x)$ qui vaut 1 en $x = 1$ et est nulle partout ailleurs dans $[0, 1]$.

Terminons, en disant que si la convergence ponctuelle est uniforme (si $K(\epsilon)$ peut être choisi indépendant de x), alors la convergence fonctionnelle s'en déduit.

3. La distance de Dirac (aussi appelée distance discrète) définie par

$$\delta(\omega_1, \omega_2) = \delta_{\omega_1,\omega_2}, \tag{F.11}$$

(est égale à un quels si $\omega_1 = \omega_2$ et vaut zéro sinon) induit la topologie discrète. En effet, dans cette topologie tout singleton est un ouvert (car il est égal à une boule ouverte de rayon égal à 1), et donc tout ensemble est un ouvert.

Continuité dans les espaces métriques. La définition générale de la continuité d'une fonction par rapport à deux topologies est que l'image inverse d'un ouvert est un ouvert. Montrons que cette condition est équivalente, dans le cas où les deux topologies sont métriques à la condition suivante :

Si $f(\cdot)$ est une fonction d'un espace métrique Ω dans un autre espace métrique Ω', alors elle continue si, et seulement si, $\forall \omega_0 \in \Omega$ et $\forall S_\epsilon(f(\omega_0))$ il existe $\delta > 0$ tel que $\omega \in S_\delta(\omega_0) \Rightarrow f(\omega) \in S_\epsilon(f(\omega_0))$.

En effet, si la fonction est continue alors l'image inverse d'un ouvert est un ouvert. Donc l'image inverse de $S_\epsilon(f(\omega_0))$ est un ouvert qui contient ω_0. Celui-ci contient alors une boule ouverte centrée sur ω_0, et l'image de cette boule est incluse dans $S_\epsilon(f(\omega_0))$.

Inversément, supposons que la fonction vérifie la condition ci-dessus et soit un ouvert de V' dont nous désignons par A l'image inverse. Montrons que A contient une boule ouverte centrée en chacun de ses points.

Soit ω un point de A, son image appartient par hypothèse à V'. Il existe donc une boule $S_\epsilon(f(\omega))$ centrée en $f(\omega)$ incluse dans V', puisque cet ensemble est ouvert. On en déduit qu'il existe une boule $S_\delta(\omega)$ centrée sur ω telle que l'image de cette boule soit incluse dans $S_\epsilon(f(\omega))$, et donc aussi dans V'. Comme A est l'image inverse de V' tous les points de cette boule doivent être dans A. Comme ceci est vrai quelque soit $\omega \in A$, A est ouvert. □

Dans les espaces métriques continuité et convergence sont reliées de façon très forte comme suit :

Une fonction d'un espace métrique dans un autre espace métrique est continue si, et seulement si *l'image de toute suite convergente est une suite convergente. La limite de l'image est unique et égale à l'image de la limite.*

Théorème. *Un espace métrique est compact si et seulement si toute suite contient une sous-suite convergente*.

Diamètre et espaces bornés. On définit le diamètre d'un espace métrique par

$$d(\Omega) \stackrel{\triangle}{=} \sup\{d(x,y) \ : x,y \in \Omega\}, \tag{F.12}$$

Un espace métrique est borné si son diamètre est fini.

Compacité, bornation, et séparabilité. On a les relations suivantes :

1. Dans un espace métrique tout ensemble compact est nécessairement borné (mais la réciproque n'est pas vraie en général).
2. Tout espace métrique compact (donc borné) est séparable.

Espaces métriques complets. Un espace métrique est dit complet si et seulement si toute suite de Cauchy converge. Une suite ω_n d'un espace métrique est dite de Cauchy, si et seulement si

$$\forall \epsilon > 0, \exists M(\epsilon) > 0 | \forall n,m \geq M(\epsilon) \ : d(\omega_n, \omega_m) < \epsilon. \tag{F.13}$$

Les espaces métriques complets ont "beaucoup" de suites convergentes, et jouent pour cela un rôle important en analyse fonctionnelle.

Un sous-espace d'un espace métrique complet est complet si, et seulement s'il est fermé.

On montre en analyse réelle que $\mathbb{R}$ muni de la distance $|x-y|$ est complet. On montre également que l'espace $\mathbb{R}^p$ muni de la distance euclidienne est complet. Il en est de même pour $\boldsymbol{C}^p$.

L'ensemble $\mathbb{R}^\infty$ de suites de carré sommable sur $\mathbb{R}$ et l'ensemble $\boldsymbol{C}^\infty$ sont complets.

Enfin, tout ce qui vient d'être dit sur les normes euclidiennes reste vrai pour une norme d'ordre $p \geq 1$ quelconque.

Complétion. Puisque nous avons fait la promotion des espaces métriques complets, voyons comment produire à partir d'un espace métrique "incomplet" un espace complet.

Tout d'abord, si l'espace est un sous-espace d'un espace complet $\mathcal{L}$, il suffit pour le compléter de prendre sa fermeture dans l'espace parent $\mathcal{L}$, ce qui a pour effet d'y inclure toutes les limites des suites de Cauchy (qui existent dans $\mathcal{L}$) et donc de rendre le sous-espace complet.

On montre qu'il est possible d'effectuer cette opération de complétion dans le cas général.

Théorème du point fixe. Un problème souvent recontré est la résolution d'une équation du type

$$T(\omega) = \omega, \tag{F.14}$$

où $T(\cdot)$ est une fonction de Ω dans Ω. Le théorème du point fixe dit que si $T(\cdot)$ est une contraction, alors cette équation admet une solution unique si l'espace est complet.

On dit que $T(\cdot)$ de $\Omega \to \Omega$ (où Ω est un espace métrique) est une *contraction* si

$$\exists \alpha < 1 \ | \forall \omega, \omega' \in \Omega \ : d(T(\omega), T(\omega')) \leq \alpha d(\omega, \omega'). \tag{F.15}$$

Alors, si $T(\cdot)$ est une contraction, et si Ω est complet, $T(\cdot)$ admet un et un seul point fixe. Ce point peut être approché au moyen de la suite sur Ω définie par

$$\omega_0 \in \Omega, \omega_{i+1} = T(\omega_i), \forall i \tag{F.16}$$

dont le premier terme est un point de départ quelconque, et les points suivants sont obtenus par applications successives de $T(\cdot)$.

Il suffit de prouver que la suite converge et que la limite satisfait en effet l'equation du point fixe. Pour prouver que la suite converge il suffit de montrer qu'elle est de Cauchy, puisque l'espace est complet. Comme la fonction est une contraction, on a

$$d(\omega_{i+1}, \omega_{i+2}) \leq \alpha d(\omega_i, \omega_{i+1}) \tag{F.17}$$

ce qui implique que

$$d(\omega_n, \omega_{n+1}) \leq \alpha^n d(\omega_0, \omega_1) \tag{F.18}$$

dont on tire par applications multiples de l'inégalité triangulaire

$$d(\omega_n, \omega_{n+k}) \leq (\alpha^n + \ldots + \alpha^{n+k-1}) d(\omega_0, \omega_1), \tag{F.19}$$

ou encore

$$d(\omega_n, \omega_{n+k}) \leq \frac{\alpha^n - \alpha^{n+k}}{1-\alpha} d(\omega_0, \omega_1) \leq \frac{\alpha^n}{1-\alpha} d(\omega_0, \omega_1). \tag{F.20}$$

La borne supérieure peut alors être rendue aussi faible que voulu en rendant n suffisamment grand, et la suite est bien de Cauchy. Soit, ω la limite. Observons tout d'abord que $T(\cdot)$ est continue. En effet, le fait que cette fonction soit une contraction implique que $\forall \delta > 0$ on a

$$d(\omega_0, \omega) < \delta \Rightarrow d(T(\omega_0), T(\omega)) < \epsilon \leq \alpha\delta, \tag{F.21}$$

et la fonction est bien continue en tout point ω_0. (En fait $\mathcal{T}(\cdot)$ est même absolument continue, car δ ne dépend que de ϵ et pas du point ω_0.)

Comme la fonction est continue, on a

$$\lim_{n\to\infty} T(\omega_n) = T(\omega). \tag{F.22}$$

D'autre part, par définition de la suite ω_n on a

$$\lim_{n\to\infty} T(\omega_n) = \lim_{n\to\infty} \omega_{n+1} = \omega \tag{F.23}$$

et comme la limite est unique on a $T(\omega) = \omega$. □

F.3 ESPACES LINEAIRES TOPOLOGIQUES

Nous allons maintenant combiner les notions de topologie avec celles issues d'une structure d'espace vectoriel. Un espace linéaire (ou vectoriel) introduit essentiellement une structure linéaire entre les éléments de l'ensemble (aussi appelés vecteurs ou points), qui permet de définir à partir de la notion de norme une distance. Cette distance est alors utilisée pour induire une topologie.

F.3.1 Espaces linéaires

Un espace vectoriel $\mathcal{L}$ sur un corps[2] (dit de scalaires) K (dans la suite nous nous intéressons au cas où $K = \mathbb{R}$ ou $K = C$) est un ensemble muni de deux lois : produit par un scalaire et addition interne. Rappelons que ces opérations doivent vérifier les propriétés suivantes (nous désignons par des lettres latines les vecteurs et par des lettres grecques les scalaires) :

1. L'addition vectorielle est commutative, associative et dispose d'un élément neutre, appelé le vecteur nul (qui est forcément unique). Celui-ci est noté $\mathbf{0}$.
2. Pour tout vecteur il existe un vecteur opposé (qui donne par addition le vecteur nul).

3. La multiplication par un scalaire de K est associative : $\alpha(\beta x) = (\alpha\beta)x$.

4. $1x = x$.

5. L'addition scalaire (resp. vectorielle) est distributive par rapport à la multiplication par un scalaire (resp. par rapport à l'addition vectorielle).

Exemples.

1. Le singleton contenant l'élément nul $\mathbf{0}$ est un espace linéaire (dit trivial).

2. $\mathbb{R}^p$ (sur $\mathbb{R}$) et C^p (sur C) munis de l'addition vectorielle usuelle et de la multiplication par un scalaire sont des espaces linéaires.

3. L'ensemble $\mathbb{R}^\infty$ des suites de nombres réels (ou complexes, C^∞) munis de somme terme à terme, et du produit des termes par un nombre réel (resp. complexe) est un espace linéaire.

4. L'ensemble des fonctions à valeurs réelles (ou complexes) définies sur un même ensemble X quelconque, munis de l'addition et de la multiplication "ponctuelles"

$$(f+g)(x) \stackrel{\triangle}{=} f(x) + g(x) \tag{F.24}$$
$$(\alpha f)(x) \stackrel{\triangle}{=} \alpha(f(x)) \tag{F.25}$$
$$\tag{F.26}$$

est un espace vectoriel dont l'élément neutre est la fonction identiquement nulle sur X.

5. L'ensemble des matrices $n \times n$ (réelles ou complexes) avec addition et multiplication par un scalaire définies comme à la section précédente est un espace linéaire.

6. Soit $\mathcal{L}$ un espace linéaire sur le corps K (notons que K est un espace linéaire par rapport à lui-même). Alors l'ensemble de toutes les transformations linéaires de $\mathcal{L}$ vers K est encore un espace linéaire. Il s'agit de l'espace dual de $\mathcal{L}$ et il est noté $\mathcal{L}^*$. [3] Dans cet espace l'addition vectorielle et la multiplication sont définies ponctuellement par

$$(t_1 + t_2)(x) = t_1(x) + t_2(x) \tag{F.27}$$
$$(\alpha t)(x) = \alpha(t(x)), \tag{F.28}$$

$\forall t_1, t_2, t \in \mathcal{L}^*, \forall x \in \mathcal{L}$.

7. L'ensemble des polynômes définis sur $\mathbb{R}$ est un espace linéaire.

F.3.2 Propriétés importantes

Les notions de dépendance linéaire, de dimension, de bases, et de sous-espaces vectoriels introduites dans l'appendice D sur $\mathbb{R}^p$ se généralisent à des espaces vectoriels quelconques, avec quelques précautions sur lesquelles nous allons insister.

Sous-espaces linéaires. Un sous-ensemble $\mathcal{M} \subset \mathcal{L}$ est un sous-espace linéaire (c'est-à-dire qu'il obéit aux axiomes ci-dessus), si, et seulement si

$$\forall x, y \in \mathcal{M}, \forall \alpha, \beta \in K \ : \alpha x + \beta y \in \mathcal{M}. \tag{F.29}$$

On peut se convaincre que cela est bien suffisant pour que $\mathcal{M}$ soit un espace linéaire.

Un espace $\mathcal{L}$ est dit somme directe de deux de ses sous-espaces $\mathcal{M}_1$ et $\mathcal{M}_2$, si tout point x de $\mathcal{L}$ peut être décomposé de façon unique en $x = x_1 + x_2$ avec $x_1 \in \mathcal{M}_1$ et $x_2 \in \mathcal{M}_2$. On écrit alors $\mathcal{L} = \mathcal{M}_1 \oplus \mathcal{M}_2$. Dans ce cas $\mathcal{M}_1 \cap \mathcal{M}_2 = \{\mathbf{0}\}$, où $\mathbf{0}$ désigne l'élément neutre de $\mathcal{L}$ (à ne pas confondre avec 0 (zéro) l'élément neutre de l'addition de K.)

On montre que si tout vecteur de $\mathcal{L}$ peut être exprimé comme une somme de deux vecteurs respectivement de $\mathcal{M}_1$ et de $\mathcal{M}_2$, alors $\mathcal{L} = \mathcal{M}_1 \oplus \mathcal{M}_2$ si, et seulement si, $\mathcal{M}_1 \cap \mathcal{M}_2 = \{\mathbf{0}\}$.

Indépendance linéaire. On dit qu'un ensemble fini de vecteurs $\{x_1, \ldots, x_k\}$ de $\mathcal{L}$ est linéairement indépendant si

$$\sum_i \alpha_i x_i = \mathbf{0} \Rightarrow \forall i \ : \alpha_i = 0. \tag{F.30}$$

On dit qu'un ensemble infini (pas nécessairement dénombrable) de vecteurs est linéairement indépendant, si toute partie finie de cet ensemble est linéairement indépendant.

L'intérêt principal de cette notion est de rendre les combinaisons linéaires uniques. En effet, si les x_k sont linéairement indépendants, alors

$$\sum_i \alpha_i x_i = \sum_i \beta_i x_i \Rightarrow \alpha_i = \beta_i, \forall i. \tag{F.31}$$

Notons que pour le moment nous ne sommes en mesure de parler que de combinaisons linéaires avec un nombre fini de termes. Nous serons en mesure de parler de combinaisons linéaires avec un nombre infini dénombrable de termes seulement lorsque nous aurons couplé la structure d'espace linéaire avec une topologie qui donnera un sens à la notion de convergence et de limite. Cela sera fait ci-dessous à la section F.3.4.

Dimension. Un espace $\mathcal{L}$ est dit de dimension égale à n s'il contient un ensemble de n vecteurs linéairement indépendants, et ne contient aucun ensemble de $n+1$ vecteurs linéairement indépendants.

Un espace est dit de dimension infinie s'il est possible de trouver un ensemble de n vecteurs linéairement indépendants, $\forall n \geq 0$.

Ces définitions s'appliquent évidemment aussi aux sous-espaces de $\mathcal{L}$.

L'ensemble $\mathbf{0}$ est un sous-espace contenant 0 vecteurs linéairement indépendants. Il est de dimension égale à zéro.

(On a coutume de définir la dimension comme la cardinalité d'un ensemble de taille maximale de vecteurs linéairement indépendants, ce qui permet en principe de distinguer entre les dimensions dénombrables et non-dénombrables.)

Enveloppes linéaires. Soit $\{x_1, \ldots, x_k\}$ une collection finie de vecteurs. Alors, l'ensemble formé par toutes les combinaisons linéaires de ces vecteurs est un sous-espace linéaire. On l'appelle enveloppe linéaire des points. On dit également que ce sous-espace est engendré par les vecteurs.

Notons que nous avons restreint notre définition à une collection finie de vecteurs. La raison en est que nous ne sommes pas en mesure, sans faire appel à la notion de convergence (voir ci-dessous), de dire ce que nous entendons par une somme infinie de termes.

Nous pouvons cependant étendre la notion d'espace engendré par une collection infinie de vecteurs de la manière suivante : nous dirons qu'une combinaison linéaire de vecteurs $\{\ldots x_\nu \ldots\}$ est une combinaison linéaire quelconque d'un sous-ensemble quelconque de taille finie de $\{\ldots x_\nu \ldots\}$.

F.3.3 Bases vectorielles

Un ensemble linéairement indépendant de vecteurs B qui engendre un sous-espace $\mathcal{M}$ est appelé base de $\mathcal{M}$. (En particulier s'il engendre tout l'espace $\mathcal{L}$, il s'agit d'une base de $\mathcal{L}$.)

Tout vecteur de $\mathcal{M}$ peut alors s'écrire sous la forme d'une combinaison linéaire unique de vecteurs de la base, ce qui permet de ramener l'étude des vecteurs à l'étude des coefficients des combinaisons linéaires (que nous appellerons les composantes).

Voici quelques théorèmes relatifs aux bases.

Théorème d'extension. Si S est un ensemble de vecteurs linéairement indépendants de $\mathcal{L}$ alors il existe une base de $\mathcal{L}$ qui contient S.

En conséquence, puisque $\{x\}$ est un ensemble linéairement indépendant de $\mathcal{L}$, si $x \neq \mathbf{0}$, on a le théorème suivant.

Théorème d'existence. Tout espace linéaire admet au moins une base.

Dimension finie. Toute base d'un espace de dimension finie n comporte exactement n vecteurs.

On peut donc dire que la dimension d'un espace de dimension finie est égale au nombre de vecteurs de ses bases.

Tout ensemble de n vecteurs linéairement indépendants est une base.

Cardinalité. Toutes les bases ont même cardinalité.

Dans un espace de dimension infinie il n'est pas suffisant pour un ensemble linéairement indépendant de comporter suffisamment d'éléments pour former une base.

Notons que les bases ne sont pas nécessairement de dimension finie.

Exemple.

Dans $\mathbb{R}^\infty$ l'ensemble de suites e_i dont tous les termes sont nuls à l'exception du i-ème qui vaut 1, forme un ensemble linéairement indépendant de taille infinie (cardinalité dénombrable). Mais la dimension de cet espace n'est pas "dénombrable". Par exemple, soit $\alpha \in \mathbb{R}$ et non nul alors la suite dont le k-ème terme vaut α^k fait évidemment partie de $\mathbb{R}^\infty$. En faisant varier α dans $\mathbb{R}$ on obtient un nombre infini et non-dénombrable de suites, et cet ensemble est linéairement indépendant. En effet, supposons qu'il existe un sous-ensemble de taille k linéairement dépendant. Cela voudrait dire qu'il existe $\lambda_1, \ldots, \lambda_k$ non tous nuls et $\alpha_1, \ldots, \alpha_k$ non nuls et distincts tels que

$$\forall j \geq 1 \ : \sum_{i=1}^{k} \lambda_i \alpha_i^j = 0. \tag{F.32}$$

Mais comme les α_j sont non-nuls, cela implique que

$$\forall j \ : \sum_{i}^{k} \lambda_i \alpha_i^{j-1} = 0. \tag{F.33}$$

ce qui veut dire qu'il existe une combinaison linéaire entre les lignes du tableau $k \times k$

$$\begin{bmatrix} 1 & \alpha_1 & \cdots & \alpha_1^{k-1} \\ 1 & \alpha_2 & \cdots & \alpha_2^{k-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \alpha_k & \cdots & \alpha_k^{k-1} \end{bmatrix} \tag{F.34}$$

donc aussi entre ses colonnes. On en déduit qu'il existe $\mu_1, \ldots, \mu_k$ non tous nuls tels que

$$\forall i = 1, k \ : \sum_{j=1}^{k} \mu_j \alpha_i^{j-1} = 0, \tag{F.35}$$

ce qui veut dire que le polynôme $\mu_1 + \mu_2 x + \ldots + \mu_k x^{k-1}$ au plus de degré $k-1$ possède k zéros distincts, ce qui est impossible puisque les μ_i sont non tous nuls. □

De cet exemple nous pouvons tirer quelques enseignements. Nous avons vu que $\mathbb{R}^\infty$ possède un ensemble linéairement indépendant ayant une infinité non-dénombrable de vecteurs. Cet ensemble peut être étendu pour former une base. Donc il existe une base ayant au moins une infinité non-dénombrable d'éléments. Par conséquent, l'ensemble de suites e_i (qui est dénombrable) ne peut pas constituer une base de $\mathbb{R}^\infty$, car toutes les bases doivent avoir même cardinalité. Enfin, il est facile de se convaincre que l'ensemble de vecteurs ci-dessus n'est pas une base, car les termes pairs de toute combinaison linéaire de ces suites sont strictement positifs, et il n'est donc pas possible couvrir $\mathbb{R}^\infty$ avec ces vecteurs.

L'ensemble $\{e_1, \ldots e_i, \ldots\}$ ne permet de générer par combinaison linéaire que des suites comportant un nombre fini (mais quelconque) de termes non-nuls. Notons que l'ensemble de ces suites est bien un sous-espace linéaire de $\mathbb{R}^\infty$ mais est beaucoup plus petit que ce dernier. Les $\{e_1, \ldots e_i, \ldots\}$ forment évidemment une base de ce sous-espace.

F.3.4 Espaces linéaires topologiques

Soit $\mathcal{L}$ une espace linéaire (un ensemble de vecteurs munis d'une structure d'espace linéaire comme définie ci-dessus). Soit K le corps sur lequel $\mathcal{L}$ est défini. Soit $\mathcal{T}$ une topologie définie sur l'ensemble des vecteurs, et $\mathcal{T}_K$ une topologie définie sur K. Ces topologies induisent des topologies sur $\mathcal{L} \times \mathcal{L}$ et sur $K \times \mathcal{L}$.

On dit que l'ensemble forme un espace topologique linéaire si

1. la somme vectorielle est continue (par rapport aux topologies de $\mathcal{L} \times \mathcal{L}$ et de $\mathcal{L}$),
2. le produit par un scalaire est continu (par rapport aux topologies de $K \times \mathcal{L}$ et de $\mathcal{L}$).

Nous excluons donc les espaces linéaires munis de topologies qui seraient incompatibles avec les opérations linéaires (au sens de la continuité). Pour motiver ce choix, disons que si les deux structures (espace linéaire et espace topologique) ne sont pas compatibles (au moins sous une certaine forme) alors il n'y a pas de raison que leur étude conjointe donne lieu à des résultats nouveaux intéressants. Or le but de cette section est justement de mettre en évidence le résultat d'un couplage et non d'un juxtaposition.

Exemple. Si nous munissons $\mathcal{L}$ de la topologie indiscrète alors toute fonction est continue. Nous avons donc un espace linéaire topologique (non métrisable).

Contre-exemple. Si nous munissons $\mathcal{L}$ de la topologie discrète alors on peut montrer que le produit par un scalaire ne peut pas être continu. Cela ne donne donc pas un espace topologique linéaire. Toutes les distances n'induisent donc certainement pas une topologie compatible, puisque la distance discrète métrise cet espace.

F.3.5 Espaces linéaires normés

Un espace vectoriel est dit normé s'il y a été défini une fonction $n(\cdot)$ de cet espace dans $\mathbb{R}$ qui vérifie les axiomes de la norme, c'est-à-dire

1. $n(x) > 0, \forall x \neq \mathbf{0}$.
2. $n(\mathbf{0}) = 0$.
3. $n(\alpha x) = |\alpha| n(x)$.
4. $n(x + y) \leq n(x) + n(y)$.

On voit que les deux dernières propriétés n'ont de sens que si la multiplication par un scalaire et l'addition vectorielle ont été définies et si le corps K est normé lui-même.

Toute norme induit une distance, définie par $d(x, y) \stackrel{\triangle}{=} n(x - y)$, et par conséquent induit aussi une topologie métrique (elle est invariante par translation). On montre qu'il y a compatibilité entre la topologie et la structure d'espace linéaire, c'est-à-dire que les opérations de somme vectorielle et de produit par un scalaire sont bien continues par rapport à une topologie induite par une norme quelconque. Donc toute norme induit une topologie compatible.

Nous appelerons dorénavant espace topologique linéaire un espace linéaire normé muni de sa topologie naturelle.

F.3.6 Sous-espace linéaire topologique

Nous appelerons sous-espace linéaire topologique tout sous-espace linéaire fermé. Donc, si nous prenons la fermeture d'un sous-espace linéaire, nous obtenons un sous-espace linéaire topologique.

Esquissons que ce type d'espace est bien un espace linéaire. Nous devons montrer que si x et y appartiennent à $\overline{\mathcal{M}}$ (où $\mathcal{M}$ est un sous-espace linéaire de $\mathcal{L}$), alors toute combinaison linéaire $\alpha x + \beta y$ y appartient aussi. Comme l'espace est fermé sur l'espace linéaire de départ, on peut construire deux suites de x_i et y_i de $\mathcal{M}$ qui convergent vers x et y. Par continuité, on en déduit que $\alpha x_i + \beta y_i$, suite de $\mathcal{M}$ converge vers $\alpha x + \beta y$, et cette limite est forcément dans $\overline{\mathcal{M}}$. □

Pour éclairer la subtile distinction entre ces deux notions (sous-espace linéaire et sous-espace linéaire topologique) prenons un exemple. Soit $\mathbb{R}$ muni de sa topologie habituelle; il s'agit d'un espace topologique linéaire. Prenons le sous-ensemble des nombres rationnels : cet ensemble contient toutes ses combinaisons linéaires (sous-entendu, d'un nombre fini de rationnels). Il s'agit donc d'un sous-espace linéaire qui n'est pas fermé (ni d'ailleurs ouvert). Comme il est dense dans $\mathbb{R}$, le sous-espace topologique linéaire correspondant est $\mathbb{R}$.

Dans la suite nous utiliserons le terme de sous-espace (les termes linéaire et topologique étant sous-entendus).

Nous poursuivons notre raisonnement en voyant comment la topologie permet de faire le pont entre combinaisons linéaires (nombre fini de termes) et séries (nombre infini de termes).

Exemples.

Nous donnons ci-dessous les exemples les plus importants d'espaces linéaires normés rencontrés en pratique, en laissant le soin au lecteur de vérifier que les espaces sont bien des espaces topologiques linéaires. Nous reviendrons sur ces exemples tout au long de la suite.

1. Espaces ℓ_n^p : il s'agit de l'espace des vecteurs (réels ou complexes) équipés de la norme

$$\|\boldsymbol{x}\| = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}. \tag{F.36}$$

Le cas particulier pour $p = 2$ donne l'espace euclidien (dont la norme est induite par le produit scalaire).

On note ℓ_n^∞ l'espace des vecteurs équipé de la norme infinie

$$\|\boldsymbol{x}\| = \max_{i=1,\ldots,p} |x_i|. \tag{F.37}$$

2. Espaces ℓ^p : il s'agit de l'espace des suites x_i telles que

$$\sum_{i=1}^{\infty} |x_i|^p < \infty, \tag{F.38}$$

c'est-à-dire convergentes en norme d'ordre p. Le cas particulier pour $p = 2$ donne l'espace euclidien ℓ^2 (dont la norme est induite par le produit scalaire).

On note ℓ^∞ l'espace des suites équipé de la norme infinie

$$\|\boldsymbol{x}\| = \sup_i |x_i|. \tag{F.39}$$

3. Espaces L^p : il s'agit d'une généralisation de ℓ^p aux espaces de fonctions (ℓ^p peut être vu comme un espace de fonctions définies sur $\mathbb{N}$). Soit $(X, \mathcal{B}, \mu)$ un espace mesuré[4] Considérons les fonctions mesurables $f(\cdot)$ définies sur X et telles que $|f(\cdot)|^p$ soit intégrable par rapport à μ. Cet espace est noté $L^p(X, \mu)$ et est un espace topologique linéaire lorsqu'il est équipé de la norme

$$\|f\| = \left(\int_X |f|^p d\mu \right)^{\frac{1}{p}}. \tag{F.40}$$

En réalité nous devons considérer comme éléments de cet ensemble de fonctions des *classes d'équivalences* de fonctions égales presque partout (au sens de la mesure μ).

Le cas particulier le plus fréquemment rencontré est celui où $(X, \mathcal{B}, \mu)$ est la droite réelle munie de la mesure de Lebesgue ou d'un distribution de probabilité.

Généralisation de la notion de combinaison linéaire. Grâce à la topologie nous pouvons maintenant généraliser au cas infini ce que nous allons dans la suite entendre par combinaison linéaire, espace engendré et base.

Soit V un ensemble de vecteurs (fini, dénombrable, ou non-dénombrable). Nous disons que V engendre l'espace $\mathcal{L}$ (ou un sous-espace $\mathcal{M}$) si, et seulement si $\mathcal{L}$ (ou $\mathcal{M}$) est la fermeture du sous-espace linéaire engendré par V (c'est-à-dire contenant toutes les combinaisons linéaires finies de vecteurs de V).

Clairement, si $\mathcal{L}$ est engendré par un ensemble fini V, l'ancienne et la nouvelle définition coïncident. Si maintenant V est dénombrable, nous voyons que la notion signifie que $\mathcal{L}$ est engendré par V si tout vecteur de $\mathcal{L}$ est soit une combinaison linéaire finie, soit la limite (au sens de la topologie) d'une suite telle que

$$\sum_{k=1}^{\infty} \alpha_k x_k. \tag{F.41}$$

On peut montrer que si $\mathcal{L}$ est engendré par V alors il est possible d'exprimer $x \in \mathcal{L}^\circ$ (un point intérieur de $\mathcal{L}$) sous forme d'un combinaison linéaire finie (mais le nombre de termes peut être arbitrairement grand, et est généralement différent d'un point à un autre). Ce ne sont donc que les points à la frontière de $\mathcal{L}$ qui nécessitent une combinaison linéaire infinie, même si V est non-dénombrable.

Généralisation de la notion de base vectorielle. Nous dirons qu'un ensemble B (fini, dénombrable, ou non-dénombrable) est une base vectorielle de $\mathcal{L}$ (ou de $\mathcal{M}$) s'il est linéairement indépendant et engendre l'espace $\mathcal{L}$ (ou $\mathcal{M}$) .

Rappelons qu'une collection infinie de vecteurs est linéairement indépendante si toutes ses parties finies le sont. Nous gardons cette définition.

F.3.7 Espaces de Banach

Un espace vectoriel normé est dit de Banach, si la topologie induite est complète. C'est donc un cas particulier d'espace complet.

Tout espace normé de dimension finie est un espace de Banach. En effet, on montre que tout espace normé de dimension finie n est essentiellement identique (topologiquement isomorphe) à ℓ_n^p, quel que soit p. On en déduit que tous ces espaces sont complets.

Dans tout espace normé de dimension finie, tout ensemble fermé borné est compact.

Tout sous-espace topologique linéaire d'un espace de Banach est encore un espace de Banach. C'est une conséquence directe du fait que tout sous-espace topologique fermé d'un espace complet est complet.

Rappelons ici que tout espace métrique incomplet peut être complété. La question est de savoir si cette opération est compatible avec la structure d'espace linéaire. La réponse est affirmative, et cela veut dire qu'un espace linéaire topologique normé peut être complété pour donner un espace de Banach.

Théorème. Si $\mathcal{B}$ est un espace de Banach, et (x_k) un suite de vecteurs de $\mathcal{B}$ tels que

$$\sum_{i=1}^{\infty} ||x_k|| < \infty \tag{F.42}$$

alors la suite de vecteurs

$$v_n = \sum_{i=1}^{n} x_k \tag{F.43}$$

converge vers un vecteur de $\mathcal{B}$. On peut donc écrire la limite sous la forme

$$v = \sum_{i=1}^{\infty} x_k. \tag{F.44}$$

Exemples.

1. Tous les espaces ℓ_n^p sont des espaces de Banach, puisqu'ils sont normés et complets.

2. Les espaces ℓ^p, pour $p < \infty$ fixé sont tous des espaces de Banach. (Nous savons déjà que c'est le cas pour $p = 2$.)

3. Les espaces $L^p(X, \mathcal{B}, \mu)$, où $(X, \mathcal{B}, \mu)$ désigne un espace mesuré (par exemple un espace probabilisé, ou $\mathbb{R}^p$ équipé de la mesure de Lebesgue), formés par les ensembles de fonctions $f(\cdot)$ définies sur X et telles que $|f|^p$ soit μ-intégrable. Ce type d'espace est un espace de Banach.

4. L'espace $B(X)$ des fonctions bornées définies sur un ensemble quelconque X, et l'espace $C(X)$ des fonctions continues sur X, quelles que soient les normes utilisées. De cette propriété générale on déduit que les espaces ℓ_n^∞ et ℓ^∞ sont des espaces de Banach. Un élément de ℓ_n^∞ est en effet une fonction bornée de $X = \{1, \ldots, n\}$ dans $\boldsymbol{C}$; de même un élément de ℓ^∞ est une fonction bornée de $\mathbb{N}$ dans $\boldsymbol{C}$.

F.3.8 Produit scalaire et espaces de Hilbert

Etant donné un espace linéaire $\mathcal{L}$, un produit scalaire est une fonction notée $\langle \cdot, \cdot \rangle$ de $\mathcal{L} \times \mathcal{L}$ dans C (nombres complexes) qui vérifie les propriétés suivantes :

1. $\langle x, x \rangle \in \mathbb{R}$ et $\langle x, x \rangle > 0, \forall x \neq 0$.
2. $\langle \mathbf{0}, \mathbf{0} \rangle = 0$
3. $\langle x, y \rangle = \langle y, x \rangle^*$ (complexe conjugué).
4. $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$ et $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle$ (linéarité: $\Rightarrow \langle \mathbf{0}, y \rangle = 0$).

On voit qu'il s'agit bien d'une généralisation du produit scalaire dans $\mathbb{R}^p$.

Inégalité de Schwarz. Dans tout espace linéaire $\mathcal{L}$ muni d'un produit scalaire on a

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle \tag{F.45}$$

$\forall x, y \in \mathcal{L}$, et l'égalité à lieu si, et seulement si les vecteurs x et y sont linéairement dépendants.

Norme induite. Un produit scalaire induit une norme de la manière suivante :

$$n(x) \stackrel{\triangle}{=} \sqrt{\langle x, x \rangle}. \tag{F.46}$$

Il est assez aisé de vérifier qu'il s'agit bien d'une norme.

Par conséquent, un espace muni d'un produit scalaire est par la même occasion muni d'une norme, d'une distance et d'une topologie. Cet espace devient donc un espace linéaire topologique spécial.

Orthogonalité. Au-delà de la topologie qu'il induit, l'intérêt principal du produit scalaire est de permettre la définition de la notion d'orthogonalité. Deux vecteurs sont dits orthogonaux si leur produit scalaire est nul. Un ensemble de vecteurs est dit orthogonal si les vecteurs de cet ensemble pris deux à deux sont orthogonaux. L'ensemble est orthonormé, si la norme des vecteurs vaut 1. Il est évident qu'un ensemble orthonormé ne peut pas contenir le vecteur nul et qu'à partir d'un ensemble orthogonal on peut construire un ensemble orthonormé si, et seulement si le vecteur nul ne fait pas partie de l'ensemble.

On a la propriété générale suivante : tout ensemble orthonormé de vecteurs est linéairement indépendant, qu'il soit fini, dénombrable ou non.

Réciproquement, tout ensemble linéairement indépendant et dénombrable peut être orthonormé.

Théorème. Si x_k est un ensemble dénombrable de vecteurs orthonormés, et α_k un ensemble correspondant de *coordonnées* de K, tel que $\sum_i \alpha_i x_i$ admette une limite, alors

$$\sum_i |\alpha_i|^2 = |\sum_i \alpha_i x_i|. \tag{F.47}$$

Tout espace induit par un produit scalaire et séparable contient une base orthonormée dénombrable. En fait tout ensemble orthonormé y est dénombrable. Tout vecteur peut dès lors être représenté sous la forme d'une série.

Par ailleurs, un espace à produit scalaire est séparable si, et seulement s'il possède une base orthonormée dénombrable.

Espaces de Hilbert. Un espace linéaire muni d'un produit scalaire dont la topologie est complète est appelé espace de Hilbert. Donc, tout espace de Hilbert est aussi un espace de Banach. L'inverse n'est évidemment pas vrai en général.

Exemples.

1. L'espace ℓ_n^2 (c'est-à-dire $\mathbb{R}^n$ muni de la norme euclidienne induite par le produit scalaire) est un espace de Hilbert.

2. L'espace ℓ^2 est un espace de Hilbert (suites de carré sommable).

3. L'espace $L^2(X, \mathcal{B}, \mu)$ de fonctions de carré intégrable est un espace de Hilbert.

4. Désignons par $\ell^2(X)$ l'ensemble de fonctions $f(\cdot)$ définies sur l'ensemble quelconque X et telles que

- $N(f(\cdot)) = \{x \in X | f(x) \neq 0\}$ soit dénombrable (noyau dénombrable),
- $\sum_{x \in X} |f(x)|^2 < \infty$ (carré sommable)

Cet ensemble est un espace de Hilbert.

F.4 ANALYSE FONCTIONNELLE EN ESPACES DE HILBERT

Les sections précédentes on montré comment en combinant trois types de notions on arrive à la structure d'espace de Hilbert : (i) une structure algébrique d'espace linéaire; (ii) un produit scalaire munissant celui-ci de la notion d'orthogonalité et de norme; (iii) un structure topologique héritée du produit scalaire (induisant une norme, donc une distance, donc une topologie métrique). Ce type de structure porte le nom d'espace pre-hilbertien, et si celui-ci est complet, d'espace de Hilbert.

Dans ce qui suit nous allons résumer les propriétés les plus importantes de ce type d'espace, en utilisant comme exemples les espaces en dimension finie (p.ex. $\mathbb{R}^p$) et les espaces de fonctions de carré intégrable utilisées en théorie des systèmes et en traitement du signal.

Notes

1. Nous utilisons indifférement les termes partie et sous-ensemble. Une partie ou un sous-ensemble peut être égal à l'ensemble universel; si nous voulons insister sur le fait qu'il ne l'est pas, nous dirons qu'il s'agit d'un sous-ensemble propre.

2. Un corps est un ensemble muni de deux opérations internes (usuellement appelées addition et multiplication) tel que les deux opérations conduisent à un groupe commutatif, que la multiplication soit distributive par rapport à l'addition, et que chaque élément non-nul admette un inverse par rapport à la multiplication. Le corps est muni de la topologie induite par la distance usuelle.

3. Une transformation linéaire de $\mathcal{L}_1$ vers $\mathcal{L}_2$ (espaces linéaires sur un même corps) est une transformation linéaire si l'image d'une combinaison linéaire est la combinaison linéaire des images.

4. X est un ensemble, $\mathcal{B}$ un ensemble borélien de parties de X (cela veut dire que $\mathcal{B}$ est fermé par union dénombrable, et complémentation), et μ est une fonction réelle et positive définie sur $\mathcal{B}$, telle que $\mu(\phi) = 0$ et $\mu\left(\bigcup B_i\right) = \sum \mu(B_i)$ pour toute suite dénombrable d'ensembles $B_i \in \mathcal{B}$ disjoints deux à deux.

Bibliographie

[Bil79] P. Billingsley, *Probability and measure*, John Wiley and Sons, 1979. B.2, B.21

[Gan66] F. R. Gantmacher, *Théorie des matrices. Tomes 1 et 2*, Dunod, 1966. D.6

[Rom75] P. Roman, *Some modern mathematics for physicists and other outsiders*, Pergamon Press, 1975. F.1

[Sap90] G. Saporta, *Probabilités, analyse des données et statistique*, Technip, 1990. B.2, B.21, C.5, C.17

[Vap95] V. N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, 1995. C.12