


Pierre Brémaud

Initiation aux Probabilités et aux chaînes de Markov

Deuxième édition entièrement révisée

 Springer

Initiation aux Probabilités et aux chaînes de Markov

Pierre Brémaud

Initiation aux Probabilités

et aux chaînes de Markov

Deuxième édition entièrement révisée

 Springer

Pierre Brémaud
www.bluepinko.fr
pierre.bremaud@ens.fr

Précédemment paru sous le titre : Introduction aux Probabilités. Modélisation des
Phénomènes Aléatoires

ISBN: 978-3-540-31421-9

e-ISBN: 978-3-540-68402-2

Library of Congress Control Number: 2008934756

Mathematics Subject Classification (2000): 60-01

© 1984, 2009 Springer-Verlag Berlin Heidelberg

Tous droits de traduction, de reproduction et d'adaptation réservés pour tous pays. La loi du 11 mars 1957 interdit les copies ou les reproductions destinées à une utilisation collective. Toute représentation, reproduction intégrale ou partielle faite par quelque procédé que ce soit, sans le consentement de l'auteur ou de ses ayants cause, est illicite et constitue une contrefaçon sanctionnée par les articles 425 et suivants du Code pénal.

Cover design: WMX Design GmbH, Heidelberg

Imprimé sur papier non acide

9 8 7 6 5 4 3 2 1

springer.com

Introduction

Ce cours s'adresse aux étudiants des universités et des grandes écoles. Son objectif est de donner les éléments de la théorie des probabilités utiles à la compréhension des modèles probabilistes de leurs spécialités respectives, ainsi que la pratique du calcul des probabilités nécessaire à l'exploitation de ces modèles. Il a été plus spécialement conçu pour ceux qui ne souhaitent pas investir dans les aspects théoriques d'une discipline sans motivation préalable. Ainsi, bien que la théorie de la mesure et de l'intégration soit, à terme, l'outil indispensable pour quiconque veut atteindre le niveau où se situent d'importantes applications, j'ai préféré ne pas en parler dès le début, où elle risquerait de prendre l'aspect d'un barrage plutôt que d'un véritable secours. C'est seulement dans le Chapitre 6 que sont résumés les aspects de la théorie abstraite de l'intégration directement utiles au but fixé. Les premiers chapitres utilisent les intégrales de façon informelle, sans mentionner les conditions de mesurabilité des ensembles ou des intégrandes. Cela ne présente aucune gêne véritable pour le lecteur familier du calcul intégral au sens de Riemann. Après avoir assimilé le contenu du Chapitre 6, il pourra vérifier que les raisonnements et les calculs des premiers chapitres restent valides à condition de remplacer systématiquement “fonction” par “fonction mesurable” et “ensemble” par “ensemble mesurable”, et d'interpréter les intégrales au sens de Lebesgue.

Cette initiation comporte trois degrés : le calcul des probabilités, la théorie des probabilités, les chaînes de Markov.

La première partie du cours (les 5 premiers chapitres) introduit les notions essentielles : événements, probabilité, variable aléatoire, probabilité conditionnelle, indépendance. L'accent est mis sur les outils de base (fonction génératrice, fonction caractéristique) et le *calcul des probabilités* (règles de Bayes, changement de variables, calcul sur les matrices de covariance et les vecteurs gaussiens). Un court chapitre est consacré à la notion d'entropie et à sa signification en théorie des communications et en physique statistique. Le seul prérequis pour cette première étape est une connaissance pratique des séries, de l'intégrale de Riemann et de l'algèbre matricielle.

La deuxième partie (Chapitres 6 et 7) concerne la *théorie des probabilités* proprement dite. Elle débute par un résumé des résultats de la théorie de l'intégration de Lebesgue, qui fournit le cadre mathématique de la théorie axiomatique des probabilités et donne une hauteur de vue bien appréciable, y compris dans la pratique du calcul. On en profite pour préciser les points techniques laissés provisoirement dans l'ombre dans la première partie. Puis vient un chapitre où sont étudiées les différentes notions de convergence, et dans lequel sont présentés les deux sommets de la théorie, la loi forte des grands nombres et le théorème de la limite gaussienne.

Le chapitre final, qui constitue à lui seul la troisième étape de l'initiation, traite des *chaînes de Markov*, la plus importante classe de processus stochastiques pour les applications. Il donne une idée de la puissance et de la flexibilité de la modélisation probabiliste.

En fin de chaque chapitre se trouve une section d'exercices, la plupart corrigés, sauf ceux marqués d'un astérisque.

Remerciements

Mes collègues Augustin Chaintreau, Marc Lelarge et Justin Salez ont relu et corrigé certains chapitres du manuscrit. Je tiens à les remercier chaleureusement pour cette aide.

Table des matières

1	La notion de probabilité	1
1.1	Épreuves, événements et variables aléatoires	1
1.2	Événements probabilisables et probabilité	8
1.3	Probabilité conditionnelle et indépendance	18
1.4	Exercices	30
2	Variables aléatoires discrètes	35
2.1	Distributions de probabilité discrètes	35
2.2	Espérance	40
2.3	Fonctions génératrices	49
2.4	Exercices	57
3	Vecteurs aléatoires	61
3.1	Distribution des vecteurs aléatoires	61
3.2	Échantillonnage de distributions de probabilité	79
3.3	Corrélation et régression linéaire	83
3.4	Vecteurs aléatoires gaussiens	90
3.5	Exercices	98
4	Espérance conditionnelle	105
4.1	Définitions et propriétés élémentaires	105
4.2	Conditionnement des vecteurs gaussiens	113
4.3	Tests d'hypothèses bayésiennes	114
4.4	Exercices	122
5	Information et entropie	125
5.1	L'inégalité de Gibbs	125
5.2	Suites typiques et compression des données	127
5.3	Codage de source	130
5.4	Autres interprétations de l'entropie	138
5.5	Exercices	140

6	L'espérance comme intégrale	145
6.1	Résumé de la théorie de l'intégration	145
6.2	Les grands théorèmes	155
6.3	Probabilité	161
6.4	Exercices	173
7	Suites de variables aléatoires	177
7.1	Convergence presque-sûre	177
7.2	Convergence en distribution	190
7.3	Autres types de convergence	195
7.4	Exercices	198
8	Chaînes de Markov	203
8.1	La matrice de transition	203
8.2	Récurrence	222
8.3	Comportement asymptotique	235
8.4	Méthode de Monte Carlo	244
8.5	Exercices	247
	SOLUTIONS DES EXERCICES	255
	INDEX	307

Chapitre 1

La notion de probabilité

1.1 Épreuves, événements et variables aléatoires

Le hasard

Le hasard intervient dans la vie quotidienne sous forme d'événements dont on ne peut pas prévoir s'ils vont ou ne vont pas se produire : la sortie d'un numéro se terminant par 42 à la loterie, la naissance d'une fille plutôt que d'un garçon, ou encore la victoire de l'équipe de football d'Argentine sur celle d'Angleterre. Ces événements sont imprévisibles dans la mesure où l'on ne peut pas dire avec certitude s'ils auront lieu. Cependant, notre ignorance du résultat à venir n'est pas totale. A chacun des événements ci-dessus, on attribue un nombre qui mesure notre attente, notre espoir ou notre pronostic : on dit que le 42 sortira avec la probabilité 0.01, qu'une petite fille naîtra avec la probabilité 0.5, que l'Argentine l'emportera sur l'Angleterre avec la probabilité 0.5.

Ces chiffres sont certes discutables (en particulier le dernier est tout à fait subjectif) mais ils témoignent que la notion de probabilité est solidement ancrée dans notre mode de pensée sous une forme quantifiable et ils suggèrent qu'une théorie mathématique en est possible. Afin de rendre cette notion opérationnelle et de lui donner un intérêt pratique, il faut l'établir dans un formalisme clair et sans ambiguïtés. C'est une chose simple, comme on le verra bientôt, bien que la présentation axiomatique moderne ait mis quelques siècles à se dégager ! Pendant ce temps, Pascal, Fermat, Huyghens, Bernoulli, Laplace et bien d'autres faisaient des probabilités et poussaient des calculs qui étaient souvent justes. Est-ce à dire que la formalisation de la notion de probabilité est inutile ?

Il faut répondre à cela que les événements dont ces pionniers calculaient les probabilités n'étaient pas très compliqués : il s'agissait la plupart du temps de jeter des

dés, ou de tirer des cartes, et tout le monde voyait bien de quoi l'on parlait. Enfin ... presque tout le monde, car entre ces respectables savants s'installaient parfois des disputes mathématiques qui n'auraient sans doute pas eu lieu si un formalisme précis avait été à la disposition des antagonistes. Le fameux paradoxe de Bertrand est un exemple frappant du dialogue de sourds qui s'établit entre trois personnes de bonne foi qui emploient les mêmes mots pour désigner des choses différentes. Ces trois personnes, à qui l'on avait demandé de calculer la probabilité pour qu'une corde d'un cercle *choisie au hasard* soit plus grande que le côté du triangle équilatéral inscrit dans le même cercle, avaient fourni les réponses $\frac{1}{2}$, $\frac{1}{3}$ et $\frac{1}{4}$ respectivement, et aucune d'elles n'avait tort en suivant sa propre idée de "choix au hasard". Ce paradoxe (présenté en détail dans l'Exemple 1.2.7) montrera la nécessité d'une formalisation de la théorie des probabilités.

Description ensembliste des événements

On n'emploie jamais le mot "probabilité" tout seul, on parle de la "probabilité d'un événement". Notre première tâche consistera donc à donner à la notion d'événement une consistance mathématique, ce qui sera fait dans le cadre de la théorie élémentaire des ensembles.

On supposera connues les opérations de base : *union*, *intersection*, et *complémentation*. On rappelle la notation : si A et B sont des sous-ensembles d'un ensemble Ω , $A \cup B$ dénote leur union et $A \cap B$ leur intersection. Dans ce livre on notera \bar{A} le complémentaire de A dans Ω . La notation $A + B$ (la *somme* de A et B) implique par convention que A et B sont *disjoints*, auquel cas elle représente l'union $A \cup B$. Semblablement, la notation $\sum_{k=1}^{\infty} A_k$ est utilisée à la place de $\cup_{k=1}^{\infty} A_k$ seulement quand les A_k sont deux à deux disjoints. La notation $A - B$ est utilisée seulement quand $B \subseteq A$, et elle signifie $A \cap \bar{B}$. En particulier, si $B \subset A$, alors $A = B + (A - B)$.

On ne sait si un événement a lieu qu'après avoir fait une expérience et observé son résultat. Ainsi, à la loterie, il faudra tirer des boules numérotées et noter le résultat ; la maman devra accoucher et constater le sexe de son bébé ; les équipes d'Angleterre et d'Argentine se rencontreront et on lira le score dans les journaux. Chaque résultat d'expérience doit être décrit. On appellera Ω l'ensemble de toutes les descriptions possibles relatives à une expérience de type donné : l'élément générique ω de Ω est donc une description. Par exemple :

$$\begin{aligned}\Omega &= \{00,01, \dots, 99\} && \text{(la loterie)} \\ \Omega &= \{M, F\} && \text{(l'accouchement)} \\ \Omega &= \{0-0, 0-1, \dots, 7-5, 7-6, \dots, \dots\} && \text{(le match de foot)}\end{aligned}$$

On aurait aussi bien pu choisir, respectivement,

$$\begin{aligned}\Omega &= \{0, 1, 2, \dots, 99\} \\ \Omega &= \{0, 1\} \\ \Omega &= \{(m, n)\} = \text{ensemble des couples ordonnés d'entiers.}\end{aligned}$$

Définition 1.1.1 (provisoire) *Un sous-ensemble A de l'ensemble des descriptions est, par définition, un événement.*

Cette définition ensembliste correspond bien à notre intuition. Prenons par exemple l'espace Ω correspondant au match Angleterre-Argentine. Une description du résultat est une paire de nombres entiers $\omega = (m, n)$, où m est le nombre de buts de l'Angleterre et n celui de l'Argentine. Le sous-ensemble de Ω ,

$$A = \{(m, n) | m > n\},$$

est bien l'événement "Angleterre bat Argentine".

Une description ω qui appartient à l'ensemble A est appelée une *réalisation* de l'événement A . Ainsi dans l'exemple du match de foot, la description (3,2) (c'est-à-dire le score 3-2) est bien dans l'ensemble A . C'est une des façons possibles pour l'Angleterre de battre l'Argentine, et on dit que l'épreuve $\omega = (3, 2)$ *réalise l'événement* $A = \text{"Angleterre bat Argentine"}$.

Le choix d'un type de description, c'est-à-dire le choix de Ω , est à la discrétion de celui qui modélise tel ou tel phénomène aléatoire. Il doit cependant veiller à ce que l'espace Ω choisi soit suffisamment riche. Par exemple, un bébé peut être décrit autrement que par son genre, masculin ou féminin. Son poids, sa taille, la couleur de ses yeux, bien d'autres paramètres sont importants pour le statisticien, le médecin ou les parents. On peut donc choisir un espace Ω plus riche que $\{0,1\}$, par exemple :

$$\Omega = \{(u, x, y) | u \in \{0,1\}, x \geq 0, y \geq 0\}.$$

Ici une description ω est un triplet (u, x, y) avec l'interprétation suivante : si $u = 0$, c'est un garçon, si $u = 1$ c'est une fille, x est le poids en livres du bébé, y est sa taille en décimètres. Si l'on s'intéresse uniquement au genre du bébé, l'espace $\Omega = \{0, 1\}$ convient et il est préférable du point de vue de l'économie des notations. Par contre les médecins trouveront le "grand espace" plus proche de leurs préoccupations, car le rapport taille sur poids est un indice important de la bonne santé du nourrisson.

Si la description est détaillée, c'est-à-dire si l'ensemble Ω est riche, le modèle probabiliste construit va être apte à décrire le phénomène aléatoire étudié avec précision. Mais trop de détails risquent d'être encombrants. Supposons qu'on cherche à étudier le débit (aléatoire) d'un fleuve au cours de l'année, ceci afin de gérer et de faire la prévision des stocks hydroélectriques. On peut prendre un modèle continu :

$$\Omega = \{(x_t, t \in [0, T])\} \text{ où } x_t,$$

où la fonction $t \rightarrow x_t$ est une fonction continue à valeurs dans \mathbb{R}_+ définie sur l'intervalle $[0, T]$ représentant une année ; si on prend l'heure comme unité, $T = 365 \times 24$, et x_t est le débit horaire au temps t . Pour les spécialistes, ce modèle en temps continu est un peu trop précis car on ne mesure généralement pas le débit d'un fleuve à chaque instant. Pour ce qui est de l'établissement d'une politique de gestion des ressources en hydroélectricité,

des mesures journalières suffisent. On est donc conduit à choisir un modèle plus simple, à temps discret :

$$\Omega = \{(y_n, 1 \leq n \leq 365)\}, \text{ où } y_n \in \mathbb{R}_+.$$

Ici y_n est le débit horaire mesuré au milieu de la n -ème journée, ou bien le débit moyen au cours de la n -ème journée.

En résumé, nous avons un ensemble Ω qui est la collection des descriptions des résultats d'une expérience donnée où le hasard intervient. Un événement est un sous-ensemble A de Ω . La terminologie en usage chez les probabilistes est

$$\omega = \text{épreuve}, \quad \Omega = \text{espace des épreuves}.$$

Cette terminologie peut être gênante au début, si on assimile, comme il est naturel, "épreuve" à "expérience". En effet, quel est l'ensemble des expériences relatives au lancer d'une pièce de monnaie (le jeu de pile ou face) ? S'agit-il de toutes les expériences déjà effectuées au cours de l'histoire de l'humanité, faut-il y inclure les expériences à venir, ou bien est-ce l'ensemble très abstrait de tous les lancers imaginables ? Ces questions sont peu intéressantes, aussi bien sur le plan mathématique que sur le plan philosophique. On s'en tiendra à l'identification :

$$\text{EPREUVE} \equiv \text{DESCRIPTION}.$$

L'espace des épreuves est alors toujours clairement défini par l'utilisateur qui sait exactement ce qu'est une description appropriée des résultats qui l'intéressent. La plupart du temps c'est une suite finie ou infinie de nombres, ou bien une fonction, en général un objet mathématique qui possède une interprétation physique, comme dans l'exemple de la gestion des stocks électriques où une fonction continue représente l'évolution du débit horaire d'un fleuve au cours du temps. Le choix de l'espace Ω est dicté à l'utilisateur par la nécessité (il faut un modèle assez riche pour ses préoccupations) et par l'économie (il n'est pas souhaitable d'avoir un modèle trop riche). Ainsi, pour modéliser un pari portant sur deux lancers consécutifs d'une pièce de monnaie, l'espace

$$\Omega = \{0,1\}$$

est insuffisant. Le choix

$$\Omega = \{000,001,010,011,100,101,110,111\}$$

est du gaspillage ; le choix raisonnable est

$$\Omega = \{00,01,10,11\}.$$

Plus généralement, l'espace des épreuves adéquat pour n lancers consécutifs est

$$\Omega = \{0,1\}^n,$$

l'ensemble des n -uplets de 0 et de 1.

Terminologie probabiliste

L'interprétation des ensembles comme événements s'accompagne d'une terminologie spécifique à la théorie des probabilités que nous allons passer en revue. On dit que l'épreuve ω réalise l'événement A si $\omega \in A$. Évidemment, si l'épreuve ω ne réalise pas A , elle réalise \overline{A} . L'événement $A \cap B$ est réalisé par l'épreuve ω si et seulement si ω réalise à la fois A et B . Semblablement, $A \cup B$ est réalisé par ω si et seulement si *au moins* un événement parmi A and B est réalisé par ω . Deux événements A et B sont dits *incompatibles* quand $A \cap B = \emptyset$. En d'autres termes, l'événement $A \cap B$ est impossible : aucune épreuve ω ne peut réaliser à la fois A and B . Pour cette raison, on appelle \emptyset *l'événement impossible*. Naturellement, Ω est appelé *l'événement certain*.

Rappelons que la notation $\sum_{k=1}^{\infty} A_k$ est utilisée pour $\cup_{k=1}^{\infty} A_k$ si et seulement si les A_k sont deux à deux disjoints. Les sous-ensembles A_1, A_2, \dots sont dits former une *partition* de Ω si

$$\sum_{k=0}^{\infty} A_k = \Omega.$$

Les probabilistes disent alors que A_1, A_2, \dots sont *mutuellement exclusifs* et *exhaustifs*. (Ils sont exhaustifs en ce sens que toute épreuve ω réalise au moins l'un d'eux. Ils sont dits exclusifs parcequ'ils sont deux à deux incompatibles. Donc toute épreuve ω réalise *un et un seul* des événements A_1, \dots, A_n .)

Si $B \subseteq A$, l'événement B est dit *impliquer* l'événement A , parceque A est réalisé par ω dès que B est réalisé par ω .

Variables aléatoires

Tout de suite après les notions d'épreuve et d'événement, c'est la notion de *variable aléatoire* qui est la plus fondamentale. Elle formalise le concept de "nombre au hasard".

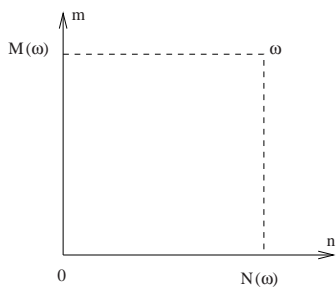
Définition 1.1.2 (provisoire.) *Une variable aléatoire est une fonction $X : \Omega \rightarrow \overline{\mathbb{R}}$. Si cette fonction ne prend que des valeurs finies, on dit que c'est une variable aléatoire réelle. L'abréviation "v.a." sera couramment utilisée.*

(La définition qui vient d'être donnée est tout-à-fait simple, même un peu trop d'ailleurs. Elle nous satisfera pleinement au début, mais, plus tard nous aurons besoin de la raffiner.)

Les variables aléatoires apparaissent très naturellement, elles sont en fait implicitement contenues dans l'espace des épreuves Ω . Elles permettent d'exprimer simplement certains événements.

EXEMPLE 1.1.1: LE MATCH DE FOOT. Dans l'exemple du match de foot, on avait défini Ω comme l'ensemble des épreuves $\omega = (m, n)$, où m et n sont deux entiers. Voici deux variables aléatoires, M et N , définies par

$$M(\omega) = m, N(\omega) = n.$$



On dit que M et N sont les *variables aléatoires coordonnées* de Ω . En effet

$$\omega = (M(\omega), N(\omega)).$$

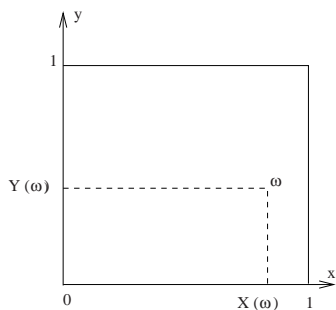
L'événement "Angleterre bat Argentine" s'exprime à l'aide de M et N par $\{\omega | M(\omega) > N(\omega)\}$ ou, en abrégé, $\{M > N\}$.

EXEMPLE 1.1.2: UN POINT AU HASARD DANS LE CARRÉ, TAKE 1. Supposons qu'on veuille modéliser le tirage au sort d'un point dans le carré unité. On prendra

$$\Omega = [0,1]^2,$$

l'ensemble des épreuves

$$\omega = (x,y) \text{ où } x \in [0,1], y \in [0,1].$$



D'où les deux *variables aléatoires coordonnées* de Ω :

$$X(\omega) = x, Y(\omega) = y.$$

EXEMPLE 1.1.3: PILE OU FACE, TAKE 1. On veut définir un espace des épreuves qui modélise une suite infinie de lancers de pièces. Avec la convention 0 = pile, 1 = face, on peut choisir

$$\Omega = \{0,1\}^\infty,$$

l'ensemble des suites infinies de 0 et de 1 :

$$\omega = (x_n, n \geq 1) \text{ où } x_n = 0 \text{ ou } 1.$$

Ici, les variables aléatoires coordonnées sont les X_n définies par

$$X_n(\omega) = x_n,$$

et la suite de variables aléatoires $(X_n, n \geq 1)$ est appelée la *suite coordonnée de Ω* . La variable aléatoire $\frac{1}{n}S_n$, où

$$S_n = X_1 + \cdots + X_n,$$

est la *fréquence empirique* de “face” sur les n premiers lancers. Les joueurs s'intéressent aussi à la variable aléatoire

$$\Sigma_n = Y_1 + \cdots + Y_n,$$

où

$$Y_n(\omega) = \begin{cases} +1 & \text{si } X_n(\omega) = 1 \\ -1 & \text{si } X_n(\omega) = 0 \end{cases}$$

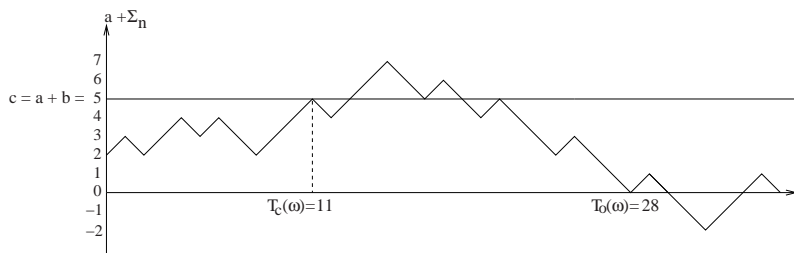
On interprètera Y_n comme le gain (qui peut être une perte!) de celui qui parie sur “face” au n -ème coup, Σ_n étant alors le gain total en n coups d'un joueur pariant systématiquement sur “face”. Appelons J_1 ce joueur et J_2 son adversaire (celui qui parie toujours sur “pile”). Soit a et b les fortunes initiales respectives de J_1 et J_2 , où a et b sont deux nombres entiers strictement positifs. Posons $c = a + b$. Si on définit les variables aléatoires T_c et T_0 par

$$T_c(\omega) = \inf \{ n \geq 1 \mid a + \Sigma_n(\omega) = c \}$$

(avec la convention $\inf \emptyset = +\infty$) et

$$T_0(\omega) = \inf \{ n \geq 1 \mid a + \Sigma_n(\omega) = 0 \}$$

(avec la même convention que ci-dessus), l'événement “ J_1 gagne par ruine de J_2 ” c'est-à-dire l'ensemble des parties ω pour lesquelles J_1 gagne admet la représentation $\{\omega \mid T_c(\omega) < T_0(\omega)\}$, ou encore, avec des notations plus sobres, $\{T_c < T_0\}$.



EXEMPLE 1.1.4: FONCTION ALÉATOIRE. On a déjà rencontré l'espace des épreuves

$$\Omega = \mathcal{C}([0, T]),$$

l'ensemble des fonctions continues à valeurs réelles définies sur $[0, T]$. Une épreuve ω est donc une fonction continue

$$\omega = (x_t, t \in [0, T]).$$

La variable aléatoire coordonnée de Ω au point $t \in [0, T]$, notée X_t , est définie par

$$X_t(\omega) = x_t,$$

si bien qu'avec ces nouvelles notations

$$\omega = (X_t(\omega), t \in [0, T]).$$

La famille de variables aléatoires $(X_t, t \in [0, T])$ s'appelle une *fonction aléatoire*. L'usage moderne est de dire *processus stochastique*.

1.2 Événements probabilisables et probabilité

Événements intéressants

Tous les sous-ensembles d'un espace d'épreuves Ω ne sont pas également intéressants pour telle ou telle application. Cette distinction entre les événements intéressants et les autres va nous conduire à la notion d'*algèbre*.

Soit A un événement, c'est-à-dire, dans notre définition provisoire, un sous-ensemble de Ω . Cet événement correspond à la réalisation d'une propriété \mathcal{P}_A . L'épreuve ω réalise l'événement A si et seulement si ω satisfait à la propriété caractéristique \mathcal{P}_A de A , ce qu'on écrit : $\mathcal{P}_A(\omega)$. Mais qu'est-ce que la propriété caractéristique de A ? En fait, c'est une notion assez peu précise, et il peut y avoir plusieurs descriptions de la même propriété caractéristique. On peut toujours définir \mathcal{P}_A par la tautologie " $\mathcal{P}_A(\omega) \Leftrightarrow \omega \in A$ ". Toutefois, le concepteur du modèle a en général une interprétation concrète des événements. Ainsi, dans l'exemple des deux joueurs, l'événement $\{T_c < T_0\}$ est l'ensemble des suites $\omega = (x_n, n \geq 1)$, où $x_n \in \{0, 1\}$, qui correspondent à une partie se soldant par la ruine du joueur J_2 . Cette dernière façon de décrire l'événement $\{T_c < T_0\}$ est concrète et imagée.

Les événements "intéressants" sont ceux qui correspondent à une propriété jugée intéressante. Chaque utilisateur a défini un ensemble de propriétés $\mathcal{U} = \{\mathcal{P}_i, i \in I\}$ qu'il qualifie de telles. Il est naturel de supposer que cette famille de propriétés satisfait aux axiomes suivants : si $\mathcal{P}_i \in \mathcal{U}$, alors non- \mathcal{P}_i (notée $\overline{\mathcal{P}_i}$) est aussi dans \mathcal{U} (car savoir si une propriété a lieu équivaut à savoir si son contraire n'a pas lieu, et vice versa). De même la conjonction de deux propriétés intéressantes est une propriété intéressante : si \mathcal{P}_i et \mathcal{P}_j sont dans \mathcal{U} alors $\mathcal{P}_i \wedge \mathcal{P}_j$ est aussi dans \mathcal{U} , où \wedge est la conjonction logique.

Soit $\mathcal{A} = \{A_i, i \in I\}$ la famille d'événements correspondant à la famille de propriétés \mathcal{U} , où A_i est définie par $\omega \in A_i \Leftrightarrow \mathcal{P}_i(\omega)$ ou encore $\mathcal{P}_i = \mathcal{P}_{A_i}$. D'après ce que nous venons de voir, \mathcal{A} satisfait aux propriétés

$$\begin{aligned} A \in \mathcal{A} &\Rightarrow \overline{A} \in \mathcal{A}, \\ A \in \mathcal{A}, B \in \mathcal{A} &\Rightarrow A \cap B \in \mathcal{A}, \end{aligned}$$

car, logiquement, $\overline{\mathcal{P}_A} = \mathcal{P}_{\overline{A}}$ et $\mathcal{P}_A \wedge \mathcal{P}_B = \mathcal{P}_{A \cap B}$. Ceci nous amène à la définition suivante :

Définition 1.2.1 Soit Ω un espace d'épreuves et \mathcal{A} une famille de sous-ensembles de Ω qui satisfait aux propriétés suivantes :

- (α) \emptyset et Ω sont dans \mathcal{A} ,
- (β) si A est dans \mathcal{A} , alors \overline{A} est dans \mathcal{A} , et
- (γ) si A et B sont dans \mathcal{A} , alors $A \cap B$ est dans \mathcal{A} .

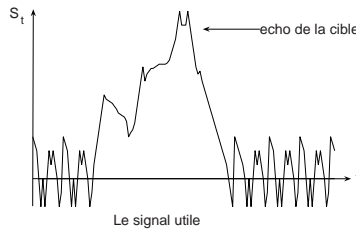
On dit alors que \mathcal{A} est une algèbre sur Ω .

Événements observables

On aurait trouvé la même structure si on avait considéré les événements *observables*. Par observable, on entend la chose suivante. On suppose que les tirages au sort sont effectués par un opérateur qui nous cache le résultat ω et n'accepte de donner que quelques indications. Par exemple pour certains événements A , cet intermédiaire acceptera de répondre à la question "est-ce que ω réalise A ?". Ce sont les événements observables. Pour d'autres événements il refusera de répondre.

Cet intermédiaire entre le résultat et nous peut paraître artificiel, il existe pourtant dans de nombreux cas pratiques. Ce peut être l'oscilloscope devant lequel veille un sonariste et sur l'écran duquel ce dernier voit s'inscrire une courbe qui est la somme du signal réfléchi et du bruit (bruit de fond électronique, réflexions parasites dues aux trajets multiples du signal acoustique, etc) :

$$Y_t(\omega) = S_t(\omega) + B_t(\omega) \quad (\text{observation} = \text{signal utile} + \text{bruit})$$



Le sonariste aimerait bien avoir la réponse à toutes les questions concernant les variables aléatoires S_t et B_t , et pourtant l'écran ne lui permet de répondre qu'aux questions qui concernent les sommes $S_t + B_t$.

Il est clair que si A est un événement observable, \overline{A} est également un événement observable puisque répondre à la question “est-ce que ω est dans A ?” c'est aussi répondre à la question “est-ce que ω n'est pas dans A ?” De même si A et B sont observables, $A \cap B$ est observable : en effet, puisqu'on obtient une réponse à toutes les questions concernant les événements observables, il suffit pour obtenir la réponse à “est-ce que ω est dans $A \cap B$?” de décomposer cette question en deux questions : “est-ce que ω est dans A ?” et “est-ce que ω est dans B ?” On aboutit donc naturellement pour les ensembles observables à une structure d'algèbre.

Un mathématicien n'a pas besoin de cette heuristique pour accepter la définition d'une algèbre. Dans les sciences appliquées, cette notion d'observable est fondamentale et a un sens physique : en télécommunications, en reconnaissance des formes (parole, écriture, etc. . .), en téléguidage, l'information utile est toujours bruitée ou brouillée.

EXEMPLE 1.2.1: ALGÈBRE TRIVIALE ET ALGÈBRE GROSSIÈRE. Sur un espace d'épreuves Ω quelconque, on trouve toujours les deux algèbres extrémales suivantes :

$$\begin{aligned}\mathcal{P}(\Omega) &= \text{la famille de tous les sous-ensembles de } \Omega, \text{ et} \\ \mathcal{G}(\Omega) &= \{\Omega, \emptyset\},\end{aligned}$$

respectivement : l'algèbre *triviale* et l'algèbre *grossière* sur Ω .

EXEMPLE 1.2.2: UNIONS D'INTERVALLES. Sur $\Omega = \mathbb{R}$, on définit $\mathcal{A}(\mathbb{R})$ comme la collection des unions finies d'intervalles de tous types de \mathbb{R} (avec en plus l'ensemble vide). On vérifie facilement que c'est une algèbre.

Tribus

Plus tard, en particulier pour la loi des grands nombres, nous aurons besoin d'une notion plus évoluée que celle d'algèbre. A titre d'exemple, considérons la situation suivante. Ω est l'ensemble des suites infinies de 0 et de 1, $\Omega = \{0,1\}^\infty$, muni des applications coordonnées $(X_n, n \geq 1)$ (voir l'Exemple 1.1.3). Une algèbre qu'il est tout à fait naturel de considérer consiste en tous les sous-ensembles de Ω de la forme

$$\{\omega; X_1(\omega) \in A_1, \dots, X_k(\omega) \in A_k\},$$

où k est un entier positif arbitraire, et les A_i sont pris parmi les quatre sous-ensembles de $\{0,1\}$. Il est clair que la famille des sous-ensembles de ce type contient le sous-ensemble vide et Ω (prendre dans un cas tous les $A_i = \emptyset$, et dans l'autre tous les $A_i = \{0,1\}$) et que c'est une algèbre, que l'on notera \mathcal{A} . Cette algèbre contient bien des

événements intéressants, mais elle ne les contient pas tous. En particulier elle ne contient pas l'événement

$$\left\{ \omega ; \lim_{n \uparrow \infty} \frac{1}{n} \sum_{i=1}^n X_i(\omega) \text{ existe} \right\}.$$

Le fait que cet événement ne soit pas contenu dans \mathcal{A} est intuitif, car la convergence de $\sum_{i=1}^n X_i(\omega)/n$ ne dépend pas des premiers termes $X_1(\omega), \dots, X_k(\omega)$ quel que soit $k \geq 1$ fixé. Or un événement de \mathcal{A} est, par définition, un événement exprimable en fonction du début de la suite infinie $(X_n, n \geq 1)$. Il nous faudra donc étendre notre champ d'action et introduire la notion de *tribu*.

Définition 1.2.2 Une tribu sur Ω est une famille \mathcal{F} de sous-ensembles de Ω telle que :

- (α) Ω et \emptyset sont dans \mathcal{F} ,
- (β) si A est dans \mathcal{F} alors \overline{A} est dans \mathcal{F} , et
- (γ) si A_n ($n \geq 1$) est dans \mathcal{F} , alors $\bigcup_{n=1}^{\infty} A_n$ est dans \mathcal{F} .

Nous verrons dans l'Exercice 1.4.1 que si \mathcal{A} est une algèbre, pour toute famille *finie* $(A_n, 1 \leq n \leq k)$ de sous-ensembles dans \mathcal{A} , alors l'union $\bigcup_{n=1}^k A_n$ est dans \mathcal{F} . On ne peut pas, si \mathcal{A} n'est qu'une algèbre, étendre ce résultat au cas d'une famille dénombrable infinie (voir l'Exercice 1.4.2).

EXEMPLE 1.2.3: UN ÉVÉNEMENT COMPLEXE. Dans l'espace des épreuves $\Omega = \{0,1\}^{\infty}$, considérons l'événement

$$A = \{ \omega ; \lim_{n \uparrow \infty} X_n(\omega) = 0 \}.$$

Pour les raisons invoquées quelques lignes plus haut, cet événement n'est pas dans l'algèbre \mathcal{A} définie plus haut. Par contre A est exprimable en termes de tels ensembles. En effet

$$A = \bigcup_{n \geq 1} \bigcap_{m \geq n} \{ \omega ; X_m(\omega) = 0 \},$$

puisque l'ensemble du second membre de cette identité n'est autre que l'ensemble des ω tels que $X_n(\omega) = 0$ pour tous les $n \geq 1$ sauf un nombre fini (voir l'Exercice 1.4.10). Donc si \mathcal{F} est une tribu qui contient \mathcal{A} , alors \mathcal{F} contient l'ensemble A puisque celui-ci est une union *dénombrable* d'intersections *dénombrables* d'événements de \mathcal{A} . Il est donc naturel de choisir pour tribu \mathcal{F} sur $\Omega = \{0,1\}^{\infty}$ la plus petite tribu qui contienne \mathcal{A} . Ou encore : \mathcal{F} est la plus petite tribu qui contient les événements $\{X_n = 0\}$, $n \geq 1$ (ces deux définitions de \mathcal{F} sont évidemment équivalentes).

La manière de définir une tribu sur un ensemble Ω donnée dans l'Exemple 1.2.3 est très courante. On commence par sélectionner une famille \mathcal{C} de sous-ensembles de Ω que l'on juge, pour une raison ou pour une autre, intéressants, et on définit la tribu \mathcal{F} comme étant la plus petite tribu contenant tous les ensembles dans \mathcal{C} . On note cette tribu $\sigma(\mathcal{C})$, et on l'appelle la *tribu engendrée par \mathcal{C}* . Par exemple, sur $\Omega = \mathbb{R}$, on

définit la *tribu de Borel* $\mathcal{B}(\mathbb{R})$ comme étant la plus petite tribu qui contient en plus de l'ensemble vide et de \mathbb{R} , tous les intervalles de type $[a, b]$. Le problème avec une telle définition, c'est que les ensembles de ces tribus (la tribu de Borel par exemple) sont souvent difficilement représentables en fonction des événements de \mathcal{C} . Cela crée des difficultés (surmontables) sur le plan théorique, mais nous ne les approfondirons que dans le Chapitre 6, car nous ne nous intéressons pour l'instant qu'aux *concepts* probabilistes, et au *calcul* des probabilités.

Voici maintenant la définition précise d'une variable aléatoire.

Définition 1.2.3 Soit Ω espace d'épreuves et soit \mathcal{F} une tribu définie sur Ω . On appelle variable aléatoire toute application $X : \Omega \rightarrow \overline{\mathbb{R}}$ telle que pour tout $a \in \mathbb{R}$, $\{X \leq a\} := \{\omega \in \Omega; X(\omega) \leq a\} \in \mathcal{F}$.

Cette définition est naturelle si l'on souhaite assigner une probabilité aux événements liés à X .

Les axiomes de probabilité

Définition 1.2.4 Soit Ω espace d'épreuves et \mathcal{F} une tribu définie sur Ω . Une probabilité sur (Ω, \mathcal{F}) est, par définition, une application de \mathcal{F} dans $[0, 1]$ telle que :

(α) $P(\Omega) = 1$, et :

(β) si $(A_n, n \geq 1)$ est une suite d'événement de \mathcal{F} deux à deux disjoints, alors

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) . \quad (1.1)$$

Le triplet (Ω, \mathcal{F}, P) s'appelle alors un *espace de probabilité*. La propriété (1.1) est la *sigma-additivité*. Si $P(A) = 0$ on dit que A est un événement *presque impossible*, tandis que si $P(A) = 1$ on dit que A est *presque certain*, ou encore qu'il a lieu *presque sûrement*.

EXEMPLE 1.2.4: UN POINT AU HASARD DANS LE CARRÉ, TAKE 2. Considérons l'espace des épreuves de l'Exemple 1.1.2 modélisant un point au hasard dans le carré unité $[0, 1] \times [0, 1]$. On prendra pour tribu \mathcal{F} la tribu engendrée par les rectangles $[a, b] \times [c, d] \in [0, 1] \times [0, 1]$, qu'on appelle la *tribu de Borel* sur $[0, 1]^2$, et qu'on note $\mathcal{B}([0, 1] \times [0, 1])$. Un résultat de la théorie de la mesure (voir le chapitre 6) dit qu'il existe une et une seule probabilité P sur (Ω, \mathcal{F}) telle que

$$P([a, b] \times [c, d]) = (b - a)(d - c).$$

La probabilité ainsi définie étend la notion de surface au sens usuel. Ainsi, la probabilité pour que le point ω "tombe" dans la partie du carré au dessus de la bissectrice du carré est la surface du triangle $\{(x, y); 0 \leq x \leq y \leq 1\}$, c'est-à-dire $\frac{1}{2}$.

EXEMPLE 1.2.5: PILE OU FACE, TAKE 2. On considère l'espace Ω des suites infinies de 0 et de 1 avec ses applications coordonnées X_n (voir l'Exemple 1.1.3). On cherche à construire une probabilité telle que

$$P(X_n = 1) = p, \quad P(X_n = 0) = 1 - p \quad (n \geq 1),$$

et

$$P(X_1 = x_1, \dots, X_k = x_k) = P(X_1 = a_1) \dots P(X_k = a_k)$$

pour tous $a_1, \dots, a_k \in \{0, 1\}$. La première condition dit que les probabilités de pile et de face ne dépendent pas de n (on garde la même pièce inaltérable pour tous les lancers). La deuxième dit que les lancers sont indépendants (cette notion "intuitive" sera formalisée dans le prochain chapitre). Considérons la tribu \mathcal{F} définie dans l'Exemple 1.2.3. Cette tribu est la plus petite tribu contenant les événements de la forme

$$\{\omega; X_1(\omega) = a_1, \dots, X_k(\omega) = a_k\}$$

pour tout $k \geq 1$ et tous $a_1, \dots, a_k \in \{0, 1\}$. La probabilité de tels événements est bien déterminée par les données précédentes. En fait, il est facile d'en déduire la formule

$$P(\{\omega \mid X_1(\omega) = a_1, \dots, X_k(\omega) = a_k\}) = p^{\sum_{i=1}^k a_i} \times (1 - p)^{k - \sum_{i=1}^k a_i}.$$

Mais la tribu \mathcal{F} contient beaucoup plus d'événements, comme on l'a vu lors de la discussion sur la notion de tribu. Il existe cependant un résultat de la théorie de la mesure qui dit qu'il existe sur l'espace de probabilité (Ω, \mathcal{F}) une et une seule probabilité satisfaisant la dernière égalité.

Quelques conséquences immédiates des axiomes

Tout d'abord

$$P(\overline{A}) = 1 - P(A). \quad (1.2)$$

(En effet, comme $\Omega = A + \overline{A}$, il vient, d'après la partie (β) de la définition, $P(A) + P(\overline{A}) = P(\Omega)$ et donc $P(\overline{A}) = 1 - P(A)$ puisque $P(\Omega) = 1$.) En faisant $A = \Omega$ dans (1.2) on obtient

$$P(\emptyset) = 0. \quad (1.3)$$

On a la propriété de *monotonie* :

$$\text{si } B \subseteq A \text{ alors } P(B) \leq P(A). \quad (1.4)$$

(Pour prouver (1.4) on écrit simplement $A = B + (A - B)$ et on applique (β) : puisque B et $A - B$ sont disjoints, $P(A) = P(B) + P(A - B)$, donc $P(A) \geq P(B)$ puisque $P(A - B) \geq 0$.) On a enfin la propriété de *sous-sigma-additivité* : si $(A_n, n \geq 1)$ est une suite d'événement de \mathcal{F}

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n). \quad (1.5)$$

Pour démontrer cette inégalité, on utilise le résultat de l'Exercice 1.4.6 : on peut trouver des A'_n disjoints tels que pour tout $n \geq 1$ $A'_n \subset A_n$ et $\bigcup_{n=1}^{\infty} A'_n = \bigcup_{n=1}^{\infty} A_n$. Donc $P(\bigcup_{n=1}^{\infty} A_n) = P(\bigcup_{n=1}^{\infty} A'_n) = \sum_{n=1}^{\infty} P(A'_n)$. Comme par monotonie on a $P(A'_n) \leq P(A_n)$, on obtient finalement (1.5).

Continuité séquentielle de la probabilité

Théorème 1.2.1 Soit (Ω, \mathcal{F}, P) un espace de probabilité et soit $(B_n, n \geq 1)$ une suite non décroissante $(B_n \subseteq B_{n+1})$ d'événements de \mathcal{F} . On a :

$$P(\bigcup_{n=1}^{\infty} B_n) = \lim_{n \uparrow \infty} P(B_n) \quad (1.6)$$

Démonstration. On utilise les décompositions suivantes :

$$\bigcup_{n=1}^{\infty} B_n = \sum_{n=1}^{\infty} (B_n - B_{n-1}),$$

et

$$B_p = \sum_{n=1}^p (B_n - B_{n-1})$$

(avec la convention $B_0 = \emptyset$), d'où l'on tire (sigma-additivité de P) :

$$\begin{aligned} P(\bigcup_{n=1}^{\infty} B_n) &= \sum_{n=1}^{\infty} P(B_n - B_{n-1}) \\ &= \lim_{p \uparrow \infty} \sum_{n=1}^p P(B_n - B_{n-1}) = \lim_{p \uparrow \infty} P(B_p). \end{aligned}$$

□

Statistique et probabilités

La probabilité d'un événement a une interprétation fréquentielle entièrement contenue dans la "loi des grands nombres" : si on réalise n expériences identiques et indépendantes se traduisant par les résultats $\omega_1, \dots, \omega_n$ et si n_A est le nombre des expériences ω_i qui réalisent l'événement A (c'est-à-dire $\omega_i \in A$), alors

$$\lim_{n \uparrow \infty} \frac{n_A}{n} = P(A), \text{ presque-sûrement.}$$

Ce qui vient d'être écrit manque de précision. Qu'entend-on par "expériences identiques et indépendantes" ? Que veut dire "presque-sûrement" ? Nous répondrons bientôt à ces questions. Pour l'instant, contentons-nous d'observer que la fonction probabilité P satisfait aux mêmes propriétés que la fonction fréquence F définie, pour n épreuves fixées $\omega_1, \dots, \omega_n$, par

$$F(A) = \frac{n_A}{n}.$$

L'inconvénient de la fonction fréquence, c'est qu'elle dépend des épreuves $\omega_1, \dots, \omega_n$. D'autre part, lorsqu'on cherche à introduire la notion de probabilité à l'aide de la notion de fréquence, on risque de faire passer la théorie des probabilités pour une science expérimentale. En réalité, c'est une *théorie mathématique*, avec ses axiomes et ses théorèmes, qui se contente de fournir un cadre formel pour la *modélisation des phénomènes aléatoires*. Un modèle étant choisi, les axiomes de la théorie permettent d'en dérouler les conséquences théoriques qui sont ensuite confrontées à l'expérience. Si un écart est observé, on change le modèle mathématique, soit quantitativement en ajustant certains paramètres, soit qualitativement en changeant la "forme" du modèle.

Considérons par exemple le jeu de pile ou face de l'Exemple 1.2.5. Le nombre p est un paramètre du modèle probabiliste choisi et nous ne nous interrogeons pas sur la provenance de la valeur adoptée pour ce paramètre, du moins pour l'instant. La théorie des probabilités fournit un théorème (qu'on démontrera plus tard), la *loi forte des grands nombres* d'Émile Borel :

Théorème 1.2.2 *La probabilité pour que la fréquence empirique de "face" tende vers p lorsque n tend vers l'infini est égale à 1. Plus précisément :*

$$P(\{\omega ; \lim_{n \uparrow \infty} S_n(\omega)/n = p\}) = 1.$$

C'est grâce à ce *théorème* que le paramètre p peut maintenant être interprété comme la fréquence empirique asymptotique de "face". La loi des grands nombres (qui est, répétons-le, un théorème et non pas une loi physique), va nous permettre d'ajuster notre modèle à la réalité. On va effectuer une infinité (c'est-à-dire, pratiquement, un nombre très grand) de lancers et on va voir si la valeur choisie pour p s'écarte notablement de la valeur de la fréquence des faces dans l'expérience réalisée. La théorie des probabilités a justement des théorèmes d'un autre type que la loi forte des grands nombres qui permettent d'évaluer l'écart acceptable entre les conséquences logiques du modèle probabiliste et l'expérience (ici un nombre *fini* d'expériences).

La *statistique* est la discipline qui s'occupe de tester l'adéquation des modèles à la réalité. Cette adéquation est jugée en fonction de certains critères arbitraires (mais physiquement satisfaisants), et elle est mesurée à l'aide de théorèmes de la théorie des probabilités. Un exemple simple va servir à illustrer la démarche intellectuelle des statisticiens. On va décrire un *test statistique* qui permet de se faire une idée de la validité d'une hypothèse relative au biais d'une pièce de monnaie. Ce test s'appuie sur l'autre grand classique de la théorie des probabilités (avec la loi des grands nombres), à savoir le *théorème central limite*. La forme primitive de ce résultat que nous allons énoncer est due à Laplace et de Moivre :

Théorème 1.2.3

$$\lim_{n \uparrow \infty} P\left(\left|\frac{S_n - np}{\sqrt{npq}}\right| \geq x\right) = \frac{2}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy.$$

En pratique on utilise ce résultat en faisant pour les grandes valeurs de n l'approximation

$$P\left(\left|\frac{S_n - np}{\sqrt{npq}}\right| \geq x\right) \simeq \frac{2}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy.$$

EXEMPLE 1.2.6: TESTONS CETTE PIÈCE POUR VOIR... Pour $x = 5$, la valeur du second membre de la dernière égalité est inférieure à 6×10^{-8} . C'est une valeur extrêmement faible, et une valeur de $\frac{S_n - np}{\sqrt{npq}}$ supérieure à 5 sera donc hautement improbable. Supposons maintenant que pour une épreuve ω et une série de 10 000 lancers, on ait trouvé $S_{10\,000}(\omega) = 4\,400$ et que le propriétaire de la pièce affirme qu'elle n'est pas truquée. Si le litige est soumis à un statisticien, celui-ci calculera la déviation standardisée effective (pour ce ω) correspondant à l'hypothèse contestée $p = 1/2$, soit

$$\frac{|S_{10\,000}(\omega) - \frac{1}{2} 10\,000|}{\sqrt{10\,000 \times \frac{1}{2} \times \frac{1}{2}}} = 12.$$

Or on vient de voir qu'un écart supérieur à 5 est très improbable. On aura tendance à se méfier du propriétaire de la pièce.

(Nous en dirons plus sur les tests statistiques dans le Chapitre 7.)

EXEMPLE 1.2.7: RÉCRÉATION : LA CORDE DE BERTRAND. On choisit "au hasard" une corde d'un cercle. Quelle est la probabilité p pour que la longueur de la corde soit supérieure au côté du triangle équilatéral inscrit dans le cercle ? (cf. Fig. 1.1)

Première solution : $p = \frac{3}{4}$ (Fig. 1.2). On prend pour Ω le disque de rayon 1 centré à l'origine O , et on pose $P(A)$ = surface de A divisée par la surface du disque. Dans ce modèle, ω est un point du disque. Dans la Figure 1.2, CD est perpendiculaire à $O\omega$ (cette corde est donc aléatoire car fonction de ω). On veut calculer la probabilité pour que la longueur de CD soit supérieure à $\sqrt{3}$, c'est à dire, la probabilité pour que $O\omega$ soit plus grand que $\frac{1}{2}$, ou encore $P(\{\omega; O\omega \geq \frac{1}{2}\})$. L'événement $\{\omega; O\omega \geq \frac{1}{2}\}$ correspond au domaine hachuré de la figure. Sa surface divisée par celle du disque entier est $\frac{3}{4}$.

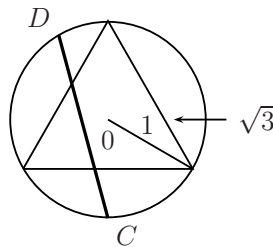


FIG. 1.1 – La corde aléatoire.

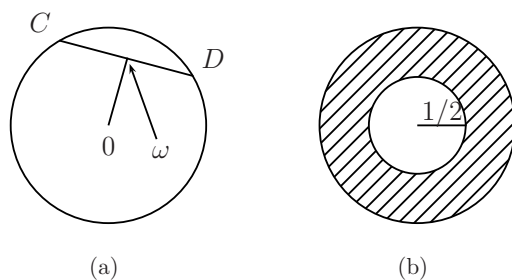


FIG. 1.2 – Première solution.

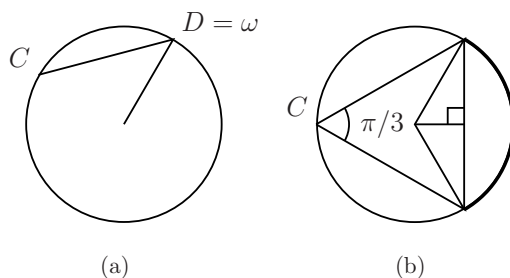


FIG. 1.3 – Deuxième solution.

Deuxième solution : $p = \frac{1}{3}$ (Fig. 1.3). On prend pour Ω la *circonférence* du disque de rayon 1 centré à l'origine O , et pour $A \subset \Omega$, on pose $P(A) = \text{longueur de } A \text{ divisée par la longueur de la circonférence}$. Dans ce modèle, ω est un point de la circonférence. Le point C étant fixé sur la circonférence, on prend $D = \omega$. Pour que la longueur de CD soit supérieure à $\sqrt{3}$, il faut et il suffit que ω tombe sur la partie de circonférence en trait gras sur la figure correspondante. La probabilité cherchée est donc $\frac{1}{3}$.

Troisième solution : $p = \frac{1}{2}$ (Fig. 1.4). On prend pour $\Omega = [0, 1]$ et pour $A \subset \Omega$, on pose $P(A) = \text{longueur de } A$. Dans ce modèle, ω est un point d'un rayon particulier (fixe) du cercle de rayon 1 centré à l'origine O . On choisit la corde CD perpendiculaire à $O\omega$. Pour que la longueur de cette corde aléatoire (fonction de ω) soit supérieure à $\sqrt{3}$, il faut et il suffit que $\omega \in [\frac{1}{2}, 1]$. La probabilité cherchée est donc $\frac{1}{2}$.

Dans chacune des solutions, l'expression "au hasard" a été interprétée comme signifiant "uniformément distribué". Dans la première solution, le point ω est uniformément distribué *sur le disque*; dans la deuxième, ω est uniformément distribué *sur la circonférence*; dans la troisième, ω est uniformément distribué *sur le rayon*. Chacun de

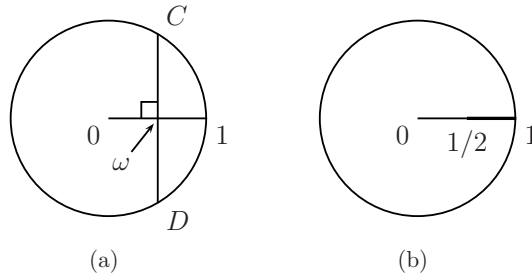


FIG. 1.4 – Troisième solution.

ces choix est extrêmement “naturel” et *a priori* irréfutable. Le probabiliste est libre de construire son modèle et d’en tirer les conclusions selon les axiomes de la théorie des probabilités. Les trois modèles ci-dessus sont différents et conduisent a priori (et a posteriori dans ce cas précis) à des résultats différents.

Il n’y a pas lieu de départager ces modèles, dans lesquels la construction de la corde CD est purement intellectuelle. Mais si la sélection de la corde aléatoire était réalisée par un mécanisme physique, alors un seul modèle correspondrait à la réalité physique (sans doute un modèle différent des 3 modèles ci-dessus!). La théorie des probabilités ne dit rien sur l’adéquation des modèles à la réalité. C’est la statistique qui permet de valider un modèle et de tester son accord avec les données expérimentales.

1.3 Probabilité conditionnelle et indépendance

Fréquence relative et probabilité conditionnelle

Soit n épreuves $\omega_1, \dots, \omega_n$. Pour tout événement A , on note n_A le nombre de fois où ces épreuves réalisent A . On rappelle la définition de la fréquence de A pour ces n épreuves :

$$F(A) = \frac{n_A}{n}.$$

Soit B un événement tel que $n_B > 0$. La *fréquence relative* de A par rapport à B est définie par

$$F(A|B) = \frac{n_{A \cap B}}{n_B},$$

où $n_{A \cap B}$ est le nombre de fois où l’événement A et l’événement B ont simultanément eu lieu. On a évidemment

$$F(A|B) = \frac{F(A \cap B)}{F(B)}.$$

Il est naturel d’étendre la notion de fréquence relative aux probabilités comme suit :

Définition 1.3.1 Soit P une probabilité définie sur (Ω, \mathcal{F}) et soit A et B deux événements de \mathcal{F} . La probabilité conditionnelle de A étant donné B , notée $P(A|B)$, est définie par

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (1.7)$$

Si $P(B) = 0$, alors $P(A \cap B) = 0$ et le second membre de (1.7) est une forme indéterminée $0/0$ que l'on prendra égale à 0. On lit $P(A|B)$ "probabilité de A si B ".

Si l'on considère que $P(A|B)$ est une idéalisation de la fréquence relative $F(A|B)$, on peut dire que $P(A|B)$ est une évaluation de $P(A)$ lorsqu'on nous permet d'observer seulement B . En effet, dans une telle situation, on ne peut pas connaître n_A mais seulement $n_{A \cap B}$ et n_B , car toutes les expériences ω qui ne réalisent pas B ne sont pas enregistrées, et le statisticien ne peut se faire une idée de la fréquence des occurrences de A qu'en calculant la fréquence relative $F(A|B)$. Le nombre $P(A|B)$ mesure l'espoir qu'on a de voir l'événement A se réaliser pour l'expérience ω lorsqu'on sait que l'événement B s'est réalisé. Ainsi la probabilité qu'on ait le cancer (événement A) sachant qu'on fume (événement B) est différente de (et dans ce cas, plus forte que) la probabilité *a priori* d'avoir le cancer.

Nous allons donner trois formules faisant intervenir les probabilités conditionnelles, que l'on appelle *formules de Bayes*.

La règle de rétrodiction

Théorème 1.3.1 Soit P une probabilité sur (Ω, \mathcal{F}) et soient A et B deux événements de probabilité strictement positive. On a

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \quad (1.8)$$

Démonstration. La définition (1.7) de la probabilité conditionnelle s'écrit

$$P(A \cap B) = P(B)P(A|B),$$

ou encore, en échangeant les rôles de A et B ,

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

L'égalité des seconds membres des identités précédentes permet de conclure. \square

La règle des causes totales

Théorème 1.3.2 Soit $(B_n, n \geq 1)$ une partition de Ω . Alors, pour tout événement A ,

$$P(A) = \sum_{n=1}^{\infty} P(A|B_n) P(B_n). \quad (1.9)$$

Démonstration. On a

$$A = A \cap \Omega = A \cap \left(\sum_{n=1}^{\infty} B_n \right) = \sum_{n=1}^{\infty} (A \cap B_n).$$

D'où

$$\begin{aligned} P(A) &= P\left(\sum_{n=1}^{\infty} A \cap B_n\right) \\ &= \sum_{n=1}^{\infty} P(A \cap B_n) \\ &= \sum_{n=1}^{\infty} P(A|B_n)P(B_n). \end{aligned}$$

□

La formule de Bayes séquentielle

Dans la suite, on remplacera dans la notation les signes d'intersection par des virgules. Ainsi :

$$\begin{aligned} P(A_1, A_2, \dots, A_n) &= P(A_1 \cap A_2 \cap \dots \cap A_n) \\ P(A_n|A_1, A_2, \dots, A_{n-1}) &= P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}), \text{ etc.} \end{aligned}$$

Théorème 1.3.3 Soit $(A_n, n \geq 1)$ une suite d'événements. On a

$$P(A_1, A_2, \dots, A_n) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_1, A_2, \dots, A_{n-1}). \quad (1.10)$$

Démonstration. Supposons la formule vraie à l'ordre $n - 1$. On a, d'après la définition de la probabilité conditionnelle,

$$P(A_1, \dots, A_n) = P(A_1, A_2, \dots, A_{n-1})P(A_n|A_1, \dots, A_{n-1}).$$

Il suffit alors d'appliquer l'hypothèse de récurrence à $P(A_1, \dots, A_{n-1})$. □

Evénements indépendants

Le nombre $P(A|B)$ mesure notre espérance de voir l'événement A se réaliser, sachant que B est réalisé. C'est la probabilité *a posteriori* de A sachant B . En général $P(A|B)$ diffère de la probabilité *a priori* $P(A)$, mais une interprétation intéressante peut être donnée lorsque les probabilités *a posteriori* et *a priori* sont égales :

$$P(A|B) = P(A),$$

ou encore, au vu de la définition de $P(A|B)$,

$$P(A \cap B) = P(A) P(B) . \quad (1.11)$$

On dit alors que A et B sont *indépendants*. Comme le montre (1.11) l'indépendance est une notion symétrique entre les événements. Lorsque A est indépendant de B , l'information " B a eu lieu" ne change pas notre prévision de voir A se réaliser.

La notion d'indépendance se généralise à plusieurs événements comme suit.

Définition 1.3.2 Soit $(A_i, i \in I)$ une famille quelconque d'événements. On dit que cette famille, ou ces événements, sont *indépendants* si et seulement si pour toute sous-famille finie $(A_{i_1}, \dots, A_{i_n})$,

$$P(A_{i_1} \cap \dots \cap A_{i_n}) = P(A_{i_1}) \dots P(A_{i_n}) . \quad (1.12)$$

Interprétation fréquentielle de l'indépendance

Soit n épreuves $\omega_1, \dots, \omega_n$ et soit F la fonction fréquence associée. Si l'on remplace dans l'égalité $P(A \cap B) = P(B)P(A|B)$ qui exprime l'indépendance de A et B , la probabilité P par la fréquence F , on obtient

$$\frac{n_A}{n} \simeq \frac{n_{B \cap A}}{n_B} .$$

Effectuons par exemple des tirages au sort parmi la population à l'aide d'un appareil de loterie qui sélectionne au hasard les numéros de sécurité sociale $\omega_1, \omega_2, \dots, \omega_n$ où $n = 100,000$, pour fixer les idées. L'événement A est défini comme suit : $\omega \in A$ si le numéro ω correspond à une femme ; quant à l'événement B , on dira qu'il est réalisé si ω est un numéro pair. On a le sentiment que A est "indépendant" de B , et donc le tableau de données expérimentales

n	n_A	n_B	$n_{A \cap B}$
100,000	49,797	51,001	24,952

nous paraîtra plus vraisemblable que

n	n_A	n_B	$n_{A \cap B}$
100,000	49,797	51,001	37,002

ce dernier laissant penser qu'on a beaucoup plus de chances de trouver une femme parmi les numéros pairs de sécurité sociale, ce qui serait étonnant.

Variables aléatoires indépendantes

Définition 1.3.3 On dit que les variables aléatoires $(X_n, n \geq 1)$ sont *indépendantes* si pour toute suite $([a_n, b_n], n \geq 1)$ d'intervalles réels, les événements $(\{X_n \in [a_n, b_n]\}, n \geq 1)$ sont *indépendants*.

EXEMPLE 1.3.1: UN POINT AU HASARD DANS LE CARRÉ, TAKE 3. (suite des Exemples 1.1.2 et 1.2.4) On a, pour tous intervalles $[a_1, b_1]$, $[a_2, b_2]$,

$$P(X \in [a_1, b_1], Y \in [a_2, b_2]) = (b_1 - a_1)(b_2 - a_2).$$

Si on fait $[a_1, b_1] = [0, 1]$

$$P(X \in [a_1, b_1]) = (b_1 - a_1).$$

De même

$$P(Y \in [a_2, b_2]) = (b_2 - a_2),$$

et donc

$$P(X \in [a_1, b_1], Y \in [a_2, b_2]) = P(X \in [a_1, b_1])P(Y \in [a_2, b_2]).$$

Les variables aléatoires X et Y sont donc indépendantes.

EXEMPLE 1.3.2: LES TROIS PRISONNIERS. Trois prisonniers, A , B et C , croupissent dans une prison, quelque part dans un pays lointain. Deux d'entre eux ont été condamnés à mort et passeront par les armes le lendemain matin, à l'aube, comme le veut l'usage. Dans le pays en question, ce sont des tribunaux exceptionnels qui décident du sort des gens en leur absence. Aussi nos prisonniers ne savent même pas lesquels parmi eux seront fusillés. Par contre le geôlier qui leur a appris la fâcheuse nouvelle connaît le nom des deux victimes. Comme il a l'habitude de monnayer toute information et que les trois prisonniers sont sans le sou, il décide de ne rien leur dire. Cependant, un des prisonniers, A , s'approche de la porte de sa cellule et s'adresse à lui à travers le judas :

- Ecoute, garde, je sais que tu ne veux donner aucun renseignement, mais peut-être pourras-tu répondre à une seule de mes questions ?
- Pas d'argent, pas de réponse, répond le géolier.
- Attends un peu, insiste le prisonnier, il s'agit d'une question telle que ta réponse ne peut me donner *aucune information*...
- Ah bon ? fait le gardien, incrédule.
- Tu nous a déjà dit que deux d'entre nous allaient mourir. Donc, parmi mes deux collègues, au moins un va être fusillé demain.
- Beau raisonnement, approuve le gardien, un sourire sadique sur les lèvres.

Le prisonnier poursuit :

- Au moins un, mais peut-être deux. Ce que tu ne veux pas me dire, c'est si c'est un seul ou les deux, car cela me renseignerait immédiatement sur mon sort.
- Juste, je ne te le dirai pas !
- Par contre si tu me dis le nom d'un parmi eux qui va mourir, je n'ai aucune information, car je ne saurais toujours pas si c'est celui-là tout seul qui va mourir, ou si tous les deux vont être fusillés.

Le gardien réfléchit un instant et dit :

- Oui, en effet, ça ne changerait rien pour toi.
- Alors, dis-moi le nom !

Le geôlier, que l'insistance du prisonnier à obtenir la réponse d'une question inutile amuse, décide de satisfaire à sa requête.

- Je t'annonce que B va mourir. Mais attention, je ne te dis pas s'il est le seul parmi les deux autres à avoir été condamné, ou si l'autre aussi va mourir.
- Je sais, je sais, dit le prisonnier, mais merci quand même. Je sens que je vais passer une bien meilleure nuit maintenant !

Le geôlier est surpris :

- Une meilleure nuit ?
- Oh oui ! Avant ton renseignement, j'avais deux chances sur trois de mourir. Mais maintenant tout se passe entre C et moi puisque le sort de B est réglé. J'ai donc une chance sur deux d'être fusillé.

Et le prisonnier ajoute pour être sûr de son effet :

- Et comme tu sais, $1/2$ c'est moins que $2/3$.

Le geôlier est furieux d'avoir été roulé et le prisonnier rejoint son grabat, l'air narquois.

Voici comment traiter (et dissoudre) le paradoxe. D'abord il faut revenir sur un petit point qui a pu passer inaperçu dans le récit. Le prisonnier a demandé un nom de condamné, mais quel nom le garde va-t-il choisir si les *deux* autres prisonniers, B et C sont condamnés ? Nommera-t-il B ou nommera-t-il C ? Si l'on veut décrire complètement la situation en termes probabilistes, il faut dire comment s'effectue le choix du nom. En l'absence de toute autre information le prisonnier doit faire une hypothèse, la plus plausible étant celle qui tient compte de son ignorance totale : Si B et C sont condamnés, le geôlier choisira de nommer B avec la probabilité $1/2$, et par conséquent il nommera C avec la probabilité $1/2$.

La formalisation du problème est maintenant possible. Il faut d'abord choisir l'espace des épreuves Ω . Une épreuve ω est une description du résultat de tous les tirages au sort, celui de la paire de prisonniers condamnés (AB , BC ou CA) et celui du prisonnier nommé par le gardien (B ou C). On prendra donc pour Ω l'ensemble des ω de la forme

$$\omega = (x, y) ,$$

où x prend ses valeurs dans l'ensemble $\{AB, BC, CA\}$ et y dans $\{B, C\}$. On définit les variables aléatoires X et Y comme les *applications coordonnées* de Ω :

$$\begin{cases} X(\omega) = x \\ Y(\omega) = y . \end{cases}$$

L'événement " A est condamné" s'écrit

$$\{\omega ; X(\omega) = AB \text{ ou } AC\} ,$$

ou encore, au choix des notations,

$$\begin{aligned} & \{\omega; X(\omega) = AB\} \cup \{\omega|X(\omega) = AC\} \\ & \{X = AB\} \cup \{X = AC\} \\ & \{X = AB \text{ ou } X = AC\} \end{aligned}$$

On veut obtenir la probabilité (notée a) que A a été condamné sachant que B a été nommé par le geôlier :

$$a = P(\{X = AB\} \cup \{X = AC\} | Y = B) .$$

Quelles sont les *données* du problème ? Ne sachant rien a priori sur la décision du tribunal, l'hypothèse suivante est raisonnable :

$$P(X = AB) = P(X = AC) = P(X = BC) = \frac{1}{3} .$$

Si $X = AB$, le nom du prisonnier nommé par le gardien sera forcément B , donc

$$P(Y = B | X = AB) = 1 .$$

De même

$$P(Y = C | X = AC) = 1 .$$

Par contre, si $X = BC$, le geôlier tire au sort B ou C , avec la même probabilité $1/2$ pour B et C , d'où

$$P(Y = B | X = BC) = P(Y = C | X = BC) = \frac{1}{2} .$$

Nous sommes maintenant prêts à effectuer les *calculs*. Puisque $\{X = AB\}$ et $\{X = AC\}$ sont des événements disjoints,

$$\begin{aligned} a &= P(\{X = AB\} \cup \{X = AC\} | Y = B) \\ &= P(X = AB | Y = B) + P(X = AC | Y = B) . \end{aligned}$$

On a :

$$P(X = AC | Y = B) = \frac{P(X = AC, Y = B)}{P(Y = B)} = 0 ,$$

puisque $\{X = AC, Y = B\} = \emptyset$ (le geôlier ne ment pas : si A et C seuls sont condamnés, il ne dira pas que B est condamné). Il reste

$$a = P(X = AB | Y = B) = \frac{P(X = AB, Y = B)}{P(Y = B)} .$$

Mais si $X = AB$, alors nécessairement $Y = B$, ou encore, en termes ensemblistes, $\{X = AB\} \subset \{Y = B\}$ et donc $\{X = AB, Y = B\} = \{X = AB\}$. L'égalité précédente se réduit donc à

$$a = \frac{P(X = AB)}{P(Y = B)} .$$

On connaît $P(X = AB)$, c'est $1/3$. Il reste donc à calculer $P(Y = B)$. On applique la règle des causes totales : puisque les événements $\{X = AB\}$, $\{X = AC\}$ et $\{X = BC\}$ forment une partition de Ω ,

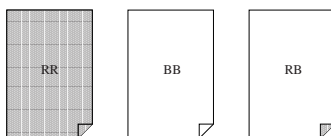
$$\begin{aligned} P(Y = B) &= P(Y = B|X = AB)P(X = AB) \\ &\quad + P(Y = B|X = AC)P(X = AC) \\ &\quad + P(Y = B|X = BC)P(X = BC) \\ &= 1 \times \frac{1}{3} + 0 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} = \frac{1}{2}. \end{aligned}$$

Finalement :

$$x = \frac{1}{3} / \frac{1}{2} = \frac{2}{3}.$$

Donc la probabilité que A soit condamné sachant que $Y = B$ est la même que la probabilité *a priori* que A soit condamné. Les événements : “ A est condamné” et “le geôlier a nommé B ” sont *indépendants*.

EXEMPLE 1.3.3: COMMENT PARIER ? Voici un jeu de cartes et d'argent qui pourra vous enrichir aux dépens des profanes. Il faut avoir à sa disposition 3 “cartes” identiques en tous points sauf en ce qui concerne la couleur. La première carte, RR , est rouge sur les deux faces. La seconde, BB , est blanche sur les deux faces. Enfin, la troisième, RB , est rouge d'un côté et blanche de l'autre.



Un distributeur neutre tire une carte au hasard et la présente une face exposée. La face exposée est elle aussi choisie au hasard. Les parieurs n'ont rien vu des opérations effectuées par le distributeur, ils n'observent que la couleur de la face exposée, disons, pour fixer les idées, rouge. On demande de parier sur la couleur de la face cachée : rouge ou blanc ? Quel serait votre choix ?

La formalisation du problème est tout-à-fait analogue à celle du paradoxe des prisonniers. Nous sauterons l'étape du choix de l'espace des épreuves pour arriver aux deux variables aléatoires intéressantes : X représente la carte tirée au hasard et prend donc ses valeurs dans l'ensemble $\{RR, BB, RB\}$, Y représente la couleur de la face exposée et a ses valeurs dans $\{R, B\}$.

Le parieur cherche à évaluer la probabilité que la face non exposée soit, disons, rouge, sachant que la face exposée est rouge. L'événement “la face non exposée est rouge” s'exprime en fonction de X et de Y de la façon suivante :

$$\{X = RR\} + \{X = RB\} \cap \{Y = B\}.$$

En effet, les deux façons *distinctes* (d'où le signe + au lieu de \cup) que l'événement en question a de se réaliser sont : “la carte est rouge des deux côtés” ou bien “la carte est rouge-blanche et la face exposée est blanche”. Reste à calculer la probabilité conditionnelle

$$a = P(\{X = RR\} + \{X = RB\} \cap \{Y = B\} | \{Y = R\}) .$$

La probabilité conditionnelle se conduit comme une probabilité (Exercice 1.4.17). En particulier elle est additive, donc

$$a = P(\{X = RR\} | \{Y = R\}) + P(\{X = RB\} \cap \{Y = B\} | \{Y = R\}) .$$

La dernière probabilité conditionnelle est nulle, et donc

$$a = P(X = RR | Y = R) .$$

Il est temps maintenant de faire l'*inventaire des données* qui serviront à calculer a . Nous les écrivons sans les justifier pour ne pas répéter ce qui a déjà été dit dans le paradoxe des prisonniers :

$$\begin{cases} P(X = RR) = P(X = RB) = P(X = BB) = \frac{1}{3} \\ P(Y = R | X = BB) = 0 \\ P(Y = R | X = RR) = 1 \\ P(Y = R | X = RB) = \frac{1}{2} . \end{cases}$$

Appliquons maintenant la formule de Bayes :

$$a = P(X = RR | Y = R) = \frac{P(X = RR, Y = R)}{P(Y = R)} .$$

Comme $\{X = RR\} \cap \{Y = R\} = \{X = RR\}$ (si c'est la carte rouge-rouge qui a été tirée, alors la face exposée est nécessairement rouge),

$$a = \frac{P(X = RR)}{P(Y = R)} .$$

La quantité $P(X = RR)$ est connue, c'est $\frac{1}{3}$. Reste à calculer $P(Y = R)$. Pour cela, on dispose de la *règle des causes totales* :

$$\begin{aligned} P(Y = R) &= P(Y = R | X = RR)P(X = RR) \\ &\quad + P(Y = R | X = RB)P(X = RB) \\ &\quad + P(Y = R | X = BB)P(X = BB) \\ &= 1 \times \frac{1}{3} + \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} = \frac{1}{2} , \end{aligned}$$

d'où

$$a = \frac{1}{3} / \frac{1}{2} = \frac{2}{3} .$$

La probabilité que la face non exposée soit rouge si la face exposée est rouge est donc $2/3$. Il faut donc parier rouge. Comme les couleurs rouge et blanche jouent un rôle symétrique dans ce problème, la stratégie optimale consiste à *parier sur la même couleur que la couleur exposée*.

EXEMPLE 1.3.4: LE THÉORÈME DE HARDY–WEINBERG. Dans les organismes diploïdes, certains *caractères* héréditaires sont portés par une paire de *gènes*, un gène pouvant prendre deux formes (ou *allèles*). Dans les expériences du moine tchèque Mendel en 1865, le caractère étudié était la nature de la peau de petits pois, lisse ou ridée. Les deux allèles correspondant au gène “nature de la peau” étaient donc

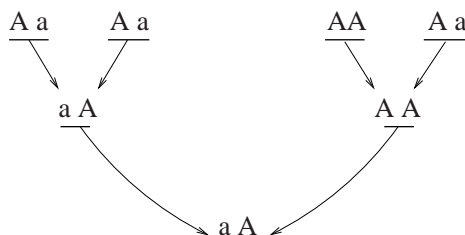
$$a = \text{peau lisse}, A = \text{peau ridée}.$$

Les gènes étant groupés en paires et les allèles étant au nombre de deux, trois *génotypes* sont donc possibles pour le caractère étudié :

$$AA, Aa, aa$$

(on ne distingue pas Aa et aA). Bien que cela soit hors de notre propos, mentionnons que le génotype est à distinguer du *phénotype* : à deux génotypes différents, ne correspondent pas nécessairement deux phénotypes différents. Il se peut en effet qu’un des allèles, disons A , soit *dominant*, auquel cas on dit que l’autre allèle, a , est *récessif*, et les deux génotypes Aa et AA ont alors le même phénotype. L’allèle a est invisible extérieurement lorsqu’il est uni à l’allèle dominant A et l’individu porteur de la paire aA ressemble du point de vue du caractère étudié à un individu de type AA . Dans la plupart des cas, le phénotype de Aa est intermédiaire entre celui de AA et celui de aa .

Chacun des deux parents contribue au patrimoine génétique du descendant commun en fournissant un gène de sa paire par l’intermédiaire des cellules reproductives, les *gamètes* (chez les animaux supérieurs, ovule et spermatozoïde). Le gène transmis par chaque parent est sélectionné au hasard dans la paire.



Ainsi, dans le deuxième couple de la première génération (voir la figure ci-dessus) le deuxième parent avait autant de chances de transmettre A que de transmettre a (dans l’exemple, il a effectivement transmis A). Évidemment, le premier parent, porteur de la paire AA , ne peut que transmettre le gène A .

On part d'une population infinie dans laquelle les proportions des génotypes AA , Aa et aa sont respectivement x , $2z$ et y . Comme il s'agit de proportions, on a

$$x + 2z + y = 1.$$

Chaque parent est sélectionné au hasard et indépendamment de l'autre. Le problème est de trouver les probabilités p , $2r$ et q d'obtenir chez le descendant les génotypes AA , Aa et aa respectivement. Le triplet $(p, 2r, q)$ peut être également interprété comme la répartition génotypique de la deuxième génération, $(x, 2z, y)$ étant celle de la première génération.

Nous allons démontrer la répartition génotypique de toutes les générations à partir de la deuxième est constante (*théorème de Hardy-Weinberg*).

Dans ce but, on définit quatre variables aléatoires : X_1 et X_2 , qui prennent les valeurs AA , Aa et aa , représentent les génotypes des deux parents ; Y_1 et Y_2 , qui prennent les valeurs A et a , représentent le gène légué par chacun des 2 parents au descendant. L'événement "premier parent de génotype Aa " s'écrit donc $\{X_1 = Aa\}$. Le problème est celui de calculer la distribution du couple (Y_1, Y_2) , c'est-à-dire les quantités du type $P(Y_1 = A, Y_2 = a)$.

Faisons l'inventaire des *données*. Tout d'abord, on connaît la probabilité d'obtenir un génotype donné chez un parent donné :

$$P(X_i = AA) = x, \quad P(X_i = Aa) = 2z, \quad P(X_i = aa) = y,$$

où i prend les valeurs 1 et 2. On connaît également la probabilité pour que le premier parent fournisse le gène A si son génotype est Aa et les quantités analogues :

$$\begin{cases} P(Y_i = A | X_i = AA) = 1, & P(Y_i = a | X_i = AA) = 0 \\ P(Y_i = a | X_i = aa) = 0, & P(Y_i = A | X_i = aa) = 1 \\ P(Y_i = A | X_i = Aa) = \frac{1}{2}, & P(Y_i = a | X_i = Aa) = \frac{1}{2}. \end{cases}$$

Il faut aussi exprimer que le tirage au sort d'un parent et le tirage au sort parmi les gènes de ce parent du gène qui contribuera au génotype du descendant sont indépendants des tirages au sort correspondants pour le deuxième parent. Par exemple $P(X_1 = Aa, Y_1 = a, X_2 = AA, Y_2 = A) = P(X_1 = Aa, Y_1 = a) \times P(X_2 = AA, Y_2 = A)$. Seule une conséquence de cette propriété sera utilisée, à savoir que Y_1 et Y_2 sont indépendantes :

$$P(Y_1 = y_1, Y_2 = y_2) = P(Y_1 = y_1) P(Y_2 = y_2).$$

Nous pouvons maintenant développer les calculs. On commence par calculer

$$p = P(Y_1 = A, Y_2 = A).$$

D'après l'indépendance de Y_1 et Y_2 ,

$$p = P(Y_1 = A)P(Y_2 = A).$$

La règle des causes totales donne

$$\begin{aligned}
 P(Y_1 = A) &= P(Y_1 = A|X_1 = AA)P(X_1 = AA) \\
 &\quad + P(Y_1 = A|X_1 = Aa)P(X_1 = Aa) \\
 &\quad + P(Y_1 = A|X_1 = aa)P(X_1 = aa) \\
 &= 1 \times x + \frac{1}{2} \times 2z + 0 \times y \\
 &= x + z,
 \end{aligned}$$

et par symétrie $P(Y_2 = A) = x + z$. On a donc

$$p = (x + z)^2.$$

En échangeant les rôles des allèles A et a , on obtient

$$q = (y + z)^2.$$

Calculons maintenant $2r$, la probabilité pour que le descendant ait le génotype Aa . Il y a deux façons distinctes d'obtenir ce génotype : " $Y_1 = A$ et $Y_2 = a$ " ou " $Y_1 = a$ et $Y_2 = A$ ". Donc $2r = P(Y_1 = A, Y_2 = a) + P(Y_1 = a, Y_2 = A)$. Par symétrie $P(Y_1 = A, Y_2 = a) = P(Y_1 = a, Y_2 = A)$, donc $r = P(Y_1 = A, Y_2 = a)$. D'après l'hypothèse d'indépendance de Y_1 et Y_2 , $r = P(Y_1 = A) \times P(Y_2 = a)$. Donc, d'après les calculs qui viennent d'être faits,

$$r = (x + z)(y + z).$$

(On aurait pu obtenir r simplement en écrivant que $p + 2r + q = 1$.) La répartition génotypique de la deuxième génération est donc :

$$p = (x + z)^2, \quad 2r = 2(x + z)(y + z), \quad q = (y + z)^2.$$

Définissons les fonctions f_1, f_2, f_3 par

$$\begin{cases} f_1(x, y, z) &= (x + z)^2 \\ f_2(x, y, z) &= (y + z)^2 \\ f_3(x, y, z) &= 2(x + z)(y + z) \end{cases}$$

On aura prouvé le théorème de Hardy-Weinberg si on vérifie que pour $i = 1, 2, 3$,

$$f_i(f_1(x, y, z), f_2(x, y, z), f_3(x, y, z)) = f_i(x, y, z).$$

On utilisera la relation $x + y + 2z = 1$ pour vérifier les 3 égalités

$$\begin{aligned}
 f_1(f_1(x, y, z), f_2(x, y, z), f_3(x, y, z)) &= ((x + z)^2 + (x + z)(y + z))^2 \\
 &= ((x + z)(x + z + y + z))^2 = (x + z)^2
 \end{aligned}$$

$$\begin{aligned}
f_2(f_1(x,y,z), f_2(x,y,z), f_3(x,y,z)) &= 2((x+z)^2 + (x+z)(y+z)) \\
&\quad ((y+z)^2 + (x+z)(y+z)) \\
&= 2((x+z)(x+z+y+z)) \\
&\quad ((y+z)(y+z+x+z)) \\
&= 2(x+z)(y+z)
\end{aligned}$$

$$\begin{aligned}
f_3(f_1(x,y,z), f_2(x,y,z), f_3(x,y,z)) &= ((y+z)^2 + (x+z)(y+z))^2 \\
&= ((y+z)(y+z+x+z))^2 = (y+z)^2.
\end{aligned}$$

1.4 Exercices

Exercice 1.4.1.

Soit \mathcal{A} une algèbre (resp. tribu) sur Ω et soit $A_i \in \mathcal{A}$, $1 \leq i \leq n$ (resp. $i \geq 1$). Montrez que $\cap_{i=1}^n A_i$ (resp. $\cap_{i=1}^\infty A_i$) est dans \mathcal{A} .

Exercice 1.4.2.

Montrez que si \mathcal{A} est une algèbre et si $A_i \in \mathcal{A}$ pour tout $i \geq 1$, alors $\cup_{i=1}^\infty A_i$ et $\cap_{i=1}^\infty A_i$ ne sont pas nécessairement dans \mathcal{A} . (Suggestion : Considérez $\Omega = \mathbb{R}$, $\mathcal{A} = \{\text{sommes finies d'intervalles de tous types}\}$, et l'ensemble \mathbf{Q} des rationnels.)

Exercice 1.4.3.

Soit $(\mathcal{A}_i, i \in I)$ une famille d'algèbres (resp. de tribus) sur Ω où I est un index quelconque. Vérifiez que $\cap_{i \in I} \mathcal{A}_i$ est aussi une algèbre (resp. une tribu) sur Ω (par définition $A \in \cap_{i \in I} \mathcal{A}_i$ si et seulement si $A \in \mathcal{A}_i$ pour tout $i \in I$).

Exercice 1.4.4.

Trouvez un contre-exemple à l'affirmation suivante : si \mathcal{A} et \mathcal{B} sont deux algèbres sur Ω , alors $\mathcal{A} \cup \mathcal{B}$ est une algèbre sur Ω (par définition $A \in \mathcal{A} \cup \mathcal{B}$ si et seulement si A est dans l'une au moins des familles \mathcal{A} et \mathcal{B}).

Exercice 1.4.5.

Soit $(\mathcal{A}_n, n \geq 1)$ une suite d'algèbres sur Ω non décroissante, c'est-à-dire : $\mathcal{A}_{n+1} \supset \mathcal{A}_n$. Montrez que $\cup_{n=1}^\infty \mathcal{A}_n$ est une algèbre sur Ω . Montrez que l'énoncé peut-être faux si on remplace "algèbre" par "tribu".

Exercice 1.4.6.

Soit $(A_n, n \geq 1)$ une suite de sous-ensembles de Ω appartenant tous à une même algèbre \mathcal{A} . Trouver une partition $(A'_n, n \geq 1)$ de Ω telle que

$$\bigcup_{i=1}^{\infty} A'_i = \bigcup_{i=1}^{\infty} A_i.$$

Exercice 1.4.7.

Soit $(A_i, 1 \leq i \leq k)$ une *partition* finie de Ω . Décrire explicitement la plus petite algèbre sur Ω contenant tous les A_i .

Exercice 1.4.8.

Existe-t-il une algèbre à 6 éléments (y compris Ω et \emptyset) ?

Exercice 1.4.9.

Soit $(A_n, n \geq 1)$ une suite quelconque d'événements. Montrez que ω appartient à l'événement

$$B = \bigcup_{n \geq 1} \bigcap_{m \geq n} A_m$$

si et seulement si $\omega \in A_n$ pour tous les n à part un nombre fini.

Exercice 1.4.10.

On se place dans la situation décrite dans l'exemple 1.1.3 dont on adopte également les notations. Montrez l'identité d'événements suivante :

$$A = \{\omega; \lim_{n \uparrow \infty} X_n(\omega) = 0\} = \bigcup_{n \geq 1} \bigcap_{m \geq n} \{\omega | X_m(\omega) = 0\}.$$

Exercice 1.4.11.

Démontrez la *formule de Poincaré* :

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) + \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^n A_i\right),$$

le terme général étant

$$(-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k}).$$

Exercice 1.4.12.

Soit (Ω, \mathcal{F}, P) un espace de probabilité et soit $(C_n, n \geq 1)$ une suite non croissante $(C_{n+1} \subseteq C_n)$ d'événements de \mathcal{F} . Montrez que

$$P(\lim_n \downarrow C_n) = \lim_n \downarrow P(C_n) .$$

Exercice 1.4.13.

Deux événements A et B de probabilités non nulles sont disjoints. Peuvent-ils être indépendants ?

Exercice 1.4.14.

Soit A, B, C , trois événements. Prouvez que

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

Exercice 1.4.15.

Si A_1, A_2, \dots, A_n sont mutuellement indépendants, alors $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n$ sont mutuellement indépendants, où $\tilde{A}_i = A_i$ ou \overline{A}_i au choix (ce choix n'étant pas forcément le même d'un i à l'autre).

Exercice 1.4.16.

Quelqu'un affirme qu'il a découvert n événements ($n \geq 2$) mutuellement indépendants et de même probabilité strictement inférieure à 1. Le croiriez-vous s'il prétend en plus que l'union de tous ces événements est l'événement certain ?

Exercice 1.4.17.

Soit P une probabilité sur (Ω, \mathcal{F}) et B un événement de probabilité strictement positive. Vérifiez que la fonction P_B de \mathcal{F} dans $[0, 1]$ définie par :

$$P_B(A) = P(A|B)$$

est une probabilité sur (Ω, \mathcal{F}) .

Exercice 1.4.18.

D'après le tableau suivant, dans quels cas les événements A et B sont-ils indépendants ?

	$P(A)$	$P(B)$	$P(A \cup B)$
cas I	0.1	0.9	0.91
cas II	0.4	0.6	0.76
cas III	0.5	0.3	0.73

Exercice 1.4.19.

Soit un espace d'épreuves à 4 points $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ et soit P la probabilité sur $(\Omega, \mathcal{P}(\Omega))$ définie par

$$P(\omega_1) = P(\omega_2) = P(\omega_3) = P(\omega_4) = \frac{1}{4}.$$

On considère les événements suivants :

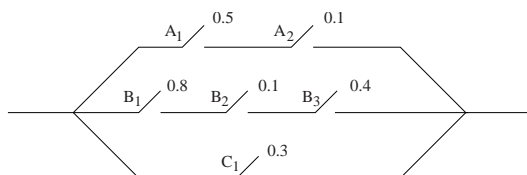
$$A = \{\omega_1, \omega_2\}, \quad B = \{\omega_2, \omega_3\}, \quad C = \{\omega_1, \omega_3\}.$$

Montrez que A est indépendant de B , B est indépendant de C , C est indépendant de A , mais que A , B et C ne forment pas une famille indépendante.

Avertissement. Les exercices qui suivent sont parfois énoncés dans un style non mathématique. C'est au lecteur de bien interpréter les questions, et au besoin d'introduire les hypothèses d'indépendance là où elles font défaut (et où elles sont plausibles).

Exercice 1.4.20.

On considère le circuit de la figure ci-dessous.



Les A_i , B_j , C_k sont des relais, en position ouverte ou fermée. Les nombres indiquent la probabilité que le relais correspondant soit ouvert. Les relais sont "indépendants" (formalisez cette notion). Calculez la probabilité pour que le circuit total "passe" (c'est-à-dire qu'il existe au moins une branche sur laquelle tous les relais sont fermés).

Exercice 1.4.21. LE PROFESSEUR NÉBULUS PERD SA VALISE.

Le professeur Nébulus voyage par avion de Los Angeles à Paris avec deux escales, la première à New York, la seconde à Londres. La probabilité de perdre un bagage est la même, p , à Los Angeles, New York et Londres. Arrivé à Paris, le professeur Nébulus constate l'absence de sa valise. Avec quelle probabilité celle-ci est-elle restée à Los Angeles ? à New York ? à Londres ?

Exercice 1.4.22. IMITATIONS ROLEX.

Un lot de montres identiques est reçu par un détaillant parisien. Ce lot provient soit d'une usine située à Cheaptown, soit d'une usine à Junkcity. L'usine de Junkcity produit un article défectueux sur 200 en moyenne, celle de Cheaptown un sur 1000. Le détaillant inspecte une première montre. Elle fonctionne correctement. Quelle est la probabilité pour que la deuxième montre inspectée fonctionne correctement ?

Exercice 1.4.23. LE SAFARI DES TROIS BEAUX-FRÈRES.

Trois touristes, grands buveurs de bière et amateurs d'émotions fortes, tirent en même temps sur un éléphant au cours d'un safari. La bête meurt frappée par deux balles. On estime la valeur de chacun des chasseurs par sa probabilité d'atteindre sa cible en un coup. Ces probabilités sont respectivement $\frac{1}{4}$, $\frac{1}{2}$ et $\frac{3}{4}$. Trouver pour chacun des chasseurs la probabilité que ce soit lui qui ait raté l'éléphant.

Exercice 1.4.24.

On tire deux points au hasard sur le segment $[0, 1]$ indépendamment l'un de l'autre. Le plus petit des deux nombres obtenus est supérieur à $\frac{1}{3}$. Quelle est la probabilité pour que le plus grand soit supérieur à $\frac{3}{4}$?

Exercice 1.4.25. CHEZ LE DOCTEUR BADNEWS.

Pour dépister une certaine maladie, on applique un test. Si le patient est effectivement atteint, le test donne un résultat positif dans 99% des cas. Mais il se peut aussi que le résultat du test soit positif alors que le consultant est en bonne santé, et ceci se produit dans 2% des cas. On sait qu'en moyenne un consultant sur 1000 est atteint de la maladie à dépister. Le docteur Badnews vous apprend que votre test est positif. Quelle probabilité avez-vous d'être vraiment atteint ?

Chapitre 2

Variables aléatoires discrètes

2.1 Distributions de probabilité discrètes

Soit (Ω, \mathcal{F}, P) un espace de probabilité et \mathcal{X} un ensemble *dénombrable*.

Définition 2.1.1 Une application $X : \Omega \rightarrow \mathcal{X}$ est appelée variable aléatoire (à valeurs dans \mathcal{X}) si $\{X = x\} \in \mathcal{F}$ pour tout $x \in \mathcal{X}$.

La variable aléatoire ci-dessus est dite *discrète* car elle prend ses valeurs dans un ensemble *dénombrable*. Cette définition précise la définition provisoire du chapitre 1. Elle demande simplement que l'on sache parler de la probabilité des événements tels que $\{X = x\}$, et donc que ces événements appartiennent à \mathcal{F} , la famille des événements probabilisables.

Définition 2.1.2 Une suite de nombres $(p(x), x \in \mathcal{X})$ telle que

$$0 \leq p(x) \leq 1 \text{ et } \sum_{x \in \mathcal{X}} p(x) = 1$$

est appelée distribution de probabilité (discrète) sur \mathcal{X} . Si la variable aléatoire X à valeurs dans \mathcal{X} est telle que

$$P(X \in A) = \sum_{x \in A} p(x) ,$$

pour tout $A \subseteq \mathcal{X}$, on dit que $(p(x), x \in \mathcal{X})$ est la distribution de probabilité de X .

Considérons l'espace produit $\mathcal{X} = \prod_{i=1}^n \mathcal{X}_i$, où les \mathcal{X}_i sont des ensembles dénombrables. Toute variable aléatoire X à valeurs dans \mathcal{X} a la forme

$$X = (X_1, \dots, X_n) ,$$

où X_i est une variable aléatoire à valeurs dans \mathcal{X}_i . C'est un *vecteur aléatoire* discret. Soit $(p(x), x \in \mathcal{X})$ sa distribution de probabilité.

Sélectionnons un sous-ensemble $\{i_1, \dots, i_k\}$ de $\{1, 2, \dots, n\}$, disons $(i_1, \dots, i_k) = (1, \dots, k)$ pour fixer les idées et simplifier les notations. La distribution de probabilité du vecteur (X_1, \dots, X_k) est, par définition

$$p_{1, \dots, k}(x_1, \dots, x_k) = P(X_1 = x_1, \dots, X_k = x_k) .$$

On a donc

$$p_{1, \dots, k}(x_1, \dots, x_k) = \sum_{x \in A} p(x)$$

où $A = \{y \in \mathcal{X}_1 \times \dots \times \mathcal{X}_n ; y_1 = x_1, \dots, y_k = x_k\}$. Par exemple, avec $n = 2$ et $k = 1$,

$$p_1(x_1) = \sum_{x_2 \in \mathcal{X}_2} p_{1,2}(x_1, x_2) .$$

La fonction $p_{1, \dots, k}(x_1, \dots, x_k)$ est appelée la *distribution marginale* de X sur $(1, \dots, k)$.

Distribution binomiale

On appellera *poids de Hamming* de $x = (x_1, \dots, x_n) \in \{0, 1\}^n$ la quantité

$$h(x) = \sum_{i=1}^n x_i .$$

Définition 2.1.3 Soit $\mathcal{X} = \{0, 1\}^n$. La distribution sur \mathcal{X} définie par

$$p(x) = p^{h(x)}(1-p)^{n-h(x)} , \tag{2.1}$$

où $p \in [0, 1]$, est appelée la distribution de Bernoulli d'ordre n et de paramètre p . Si le vecteur aléatoire X à valeurs dans $\mathcal{X} = \{0, 1\}^n$ a la distribution (2.1), on dit que X est un vecteur aléatoire de Bernoulli d'ordre n et de paramètre p .

On vérifie que $\sum_{x \in \mathcal{X}} p(x) = 1$. En effet cette somme est égale à $\sum_{k=0}^n C_n^k p^k (1-p)^{n-k}$ puisqu'il y a C_n^k vecteurs de \mathcal{X} de poids de Hamming k , et que (formule du binôme)

$$(x+y)^n = \sum_{k=0}^n C_n^k x^k y^{n-k} .$$

EXEMPLE 2.1.1: PILE OU FACE, TAKE 5. On lance une même pièce, biaisée ou non, n fois de suite. On a donc une suite de variables aléatoires X_1, X_2, \dots, X_n , où X_j est le résultat du j -ème lancer, 0 si “pile”, 1 si “face”. On note p la probabilité de tomber sur face :

$$P(X_j = 1) = p, \quad P(X_j = 0) = 1 - p.$$

On supposera de plus que les lancers sont indépendants, c’est-à-dire que les variables aléatoires X_1, X_2, \dots, X_n sont indépendantes. En particulier

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i),$$

pour tout $(x_1, \dots, x_n) \in \{0, 1\}^n$. On retrouve donc l’expression (2.1) pour la distribution de probabilité du vecteur $X = (X_1, \dots, X_n)$.

Définition 2.1.4 Soit $\mathcal{X} = \{0, 1, 2, \dots, n\}$. La distribution de probabilité

$$p(k) = C_n^k p^k (1-p)^{n-k} \quad (0 \leq k \leq n), \quad (2.2)$$

où $p \in [0, 1]$, est appelée la distribution binomiale d’ordre n et de paramètre p . Si la variable aléatoire X à valeurs dans $\{0, 1, \dots, n\}$ admet la distribution de probabilité (2.2), on dit que c’est une variable binomiale d’ordre n et de paramètre p . On note ceci $\mathcal{X} \sim \mathcal{B}(n, p)$.

La formule (2.2) définit bien une distribution de probabilité puisque $\sum_{k=0}^n p(k) = 1$ d’après la formule du binôme.

EXEMPLE 2.1.2: PILE OU FACE, TAKE 5. Si $X = (X_1, \dots, X_n)$ est un vecteur aléatoire de Bernoulli, la somme

$$S_n = X_1 + \dots + X_n$$

est une variable aléatoire binomiale, car il y a C_n^k façons disjointes d’obtenir $S_n = X_1 + \dots + X_n = k$. Chacune correspond à une suite $1 \leq i_1 < i_2 < \dots < i_k \leq n$ et à l’événement

$$\left(\bigcap_{1 \leq \ell \leq k} \{X_{i_\ell} = 1\} \right) \cap \left(\bigcap_{r \neq i_1, \dots, i_k} \{X_r = 0\} \right)$$

dont la probabilité est $p^k (1-p)^{n-k}$.

Distribution multinomiale

Définition 2.1.5 Soit n un entier strictement positif et $\mathbf{p} = (p_1, \dots, p_k)$ une distribution de probabilité sur $\{1, \dots, k\}$. Soit \mathcal{X} l'ensemble des k -uplets de nombres entiers non négatifs (n_1, n_2, \dots, n_k) tels que

$$\sum_{i=1}^k n_i = n .$$

La distribution de probabilité sur \mathcal{X}

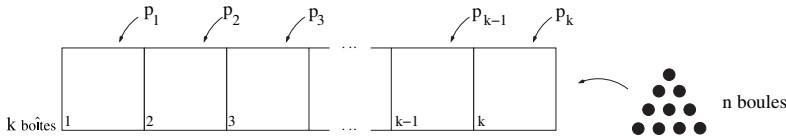
$$p(n_1, \dots, n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} , \quad (2.3)$$

s'appelle la distribution multinomiale de paramètres n, k, \mathbf{p} . Tout vecteur aléatoire $X = (X_1, \dots, X_k)$ à valeurs dans \mathcal{X} et de distribution de probabilité (2.3) s'appelle vecteur aléatoire multinomial de paramètres n, k, \mathbf{p} . On note ceci $X \sim \mathcal{M}(n, k, \mathbf{p})$.

Dans le cas où $k = 2$, on retrouve la distribution binomiale. Plus précisément : si $X = (X_1, X_2)$, alors $X_1 \sim \mathcal{B}(n, p_1)$. En effet, comme $p_1 + p_2 = 1$,

$$\begin{aligned} P(X_1 = k) &= P(X_1 = k, X_2 = n - k) \\ &= \frac{n!}{k!(n-k)!} p_1^k (1-p_1)^{n-k} . \end{aligned}$$

EXEMPLE 2.1.3: LES BOULES DANS LES BOÎTES. On considère k boîtes dans lesquelles on place n boules indépendamment les unes des autres. On note p_i la probabilité pour qu'une boule donnée se retrouve dans la boîte i et X_i le nombre de boules dans cette boîte.



On a

$$P(X_1 = n_1, \dots, X_k = n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} ,$$

où $\sum_{i=1}^k n_i = n$. En effet, il y a $\frac{n!}{n_1! \dots n_k!}$ façons distinctes d'obtenir la configuration (n_1, \dots, n_k) et chaque configuration a la probabilité $p_1^{n_1} \dots p_k^{n_k}$.

Distribution de Poisson

EXEMPLE 2.1.4: LES ÉVÉNEMENTS RARES DE POISSON. Considérons une suite X_1, \dots, X_n de variables indépendantes de distribution de probabilité commune

$$P(X_i = 1) = 1 - P(X_i = 0) = p_n .$$

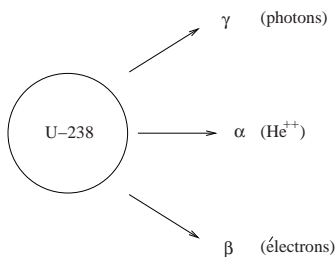
La somme S_n suit une loi de Bernoulli d'ordre n et de paramètre p_n :

$$P(S_n = k) = C_n^k p_n^k q_n^{1-k} \quad (0 \leq k \leq n) ,$$

de moyenne

$$E[S_n] = np_n .$$

Considérons par exemple le phénomène de désintégration spontanée de l'uranium 238 par émission de rayonnement α (entre autres), c'est-à-dire de noyaux d'hélium. Pour une masse donnée d'uranium 238 comprenant des milliards d'atomes, on cherche à calculer la loi du nombre Y de noyaux d'hélium émis pendant un certain laps de temps. On peut considérer que chaque noyau $i \in \{1, 2, \dots, n\}$ émet indépendamment des autres et que, du moins pendant un intervalle de temps fixé et pas trop long, il émet $X_i = 0$ ou 1 noyau d'hélium. Le nombre n des atomes d'uranium est inaccessible à l'expérience, mais on sait mesurer le nombre moyen d'atomes d'hélium émis dans un intervalle de temps de longueur donnée.



On est donc conduit à se poser la question suivante : soit Y une variable aléatoire dont on sait qu'elle est la somme d'un nombre n "très grand" de variables de Bernoulli indépendantes et identiquement distribuées. La seule donnée expérimentale étant la moyenne λ de Y , quel n choisir ? Pour résoudre le problème du choix de n dans $Y = S_n$, ou plutôt pour éviter de faire ce choix, on fait " $n = \infty$ ". Plus exactement, on passe à la limite :

$$P(Y = k) = \lim_{n \uparrow \infty} P(S_n = k) .$$

Comme la moyenne λ de Y est connue, on fait tendre n vers l'infini de façon à maintenir la moyenne de S_n égale à celle de Y :

$$np_n = \lambda .$$

L'intérêt d'une telle approximation réside dans la simplicité de la distribution limite. En effet, $P(S_n = 0) = (1 - p_n)^n = (1 - \frac{\lambda}{n})^n$ et donc

$$\lim_{n \uparrow \infty} P(S_n = 0) = e^{-\lambda},$$

et d'autre part pour $n \geq k + 1$,

$$\begin{aligned} \frac{P(S_n = k + 1)}{P(S_n = k)} &= \frac{n - k}{k + 1} \frac{p_n}{1 - p_n} \\ &= \frac{n - k}{k + 1} \frac{\lambda}{n} \left(1 - \frac{\lambda}{n}\right)^{-1} \end{aligned}$$

et donc

$$\lim_{n \uparrow \infty} \frac{P(S_n = k + 1)}{P(S_n = k)} = \frac{\lambda}{k + 1}.$$

En combinant les résultats ci-dessus on trouve pour tout $k \geq 0$:

$$\lim_{n \uparrow \infty} P(S_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

L'exemple précédent nous conduit à la définition suivante :

Définition 2.1.6 Soit $\mathcal{X} = \mathbb{N} = \{0, 1, 2, \dots\}$. La distribution

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (k \geq 0) \tag{2.4}$$

(avec la convention $0! = 1$), où λ est un nombre réel positif, est la distribution de Poisson de paramètre λ . Si X admet la distribution (2.4) on dit que c'est une variable aléatoire de Poisson de paramètre λ . On note ceci $X \sim \mathcal{P}(\lambda)$.

Il faut évidemment montrer que (2.4) définit une distribution de probabilité. On vérifie qu'on a bien $\sum_{k=0}^{\infty} p(k) = 1$, car $e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$.

2.2 Espérance

Soit X une variable aléatoire à valeurs dans l'ensemble dénombrable \mathcal{X} et de distribution $(p(x), x \in \mathcal{X})$. Soit maintenant une fonction $f : \mathcal{X} \rightarrow \overline{\mathbb{R}}$ (f est donc susceptible de prendre des valeurs infinies, positives ou négatives). Nous allons définir l'*espérance* de la variable aléatoire $f(X)$, notée $E[f(X)]$. Différents cas de figure se présentent.

(A) La fonction f ne prend que des valeurs non négatives, y compris $+\infty$. Alors, *par définition*,

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x) .$$

On notera que la quantité du second membre est toujours bien définie, mais qu'elle peut être infinie.

Il est clair que si f et g sont des applications de \mathcal{X} dans $\overline{\mathbb{R}}_+$ telles que $f \leq g$, alors $E[f(X)] \leq E[g(X)]$. C'est la propriété de *monotonie* de l'espérance.

(B) Cas général. On décompose la fonction f en ses parties non négative et non-positive

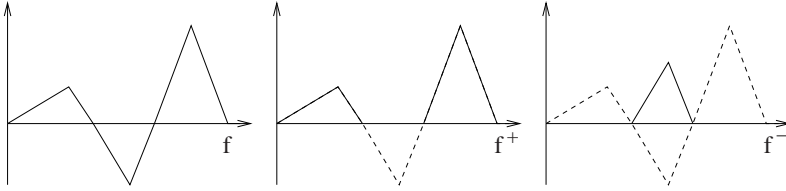
$$f = f^+ - f^- ,$$

où

$$f^+(x) = \max(f(x), 0) \quad , \quad f^-(x) = \max(-f(x), 0) .$$

On a bien entendu

$$|f| = f^+ + f^- \quad \text{et} \quad f^\pm \leq |f| .$$



On a déjà donné un sens à $E[f^+(X)]$ et $E[f^-(X)]$, mais comme ces quantités non négatives peuvent être infinies, il faut prendre quelques précautions avant de définir l'espérance $E[f(X)]$ par la formule

$$E[f(X)] = E[f^+(X)] - E[f^-(X)] . \quad (2.5)$$

à cause de l'éventualité de la forme indéterminée $+\infty - \infty$. On distingue 3 sous-cas :

(B1) Si $E[|f(X)|] < \infty$, c'est-à-dire :

$$\sum_{x \in \mathcal{X}} |f(x)|p(x) < \infty ,$$

on dit alors que $f(X)$ est *intégrable*. Comme $f^\pm \leq |f|$, on a $E[f^\pm(X)] \leq E[|f(X)|] < \infty$, on peut donc définir $E[f(X)]$ par (2.5). Dans ce cas $E[f(X)]$ est *finie* et

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x) .$$

(B2) Si $E[|f(X)|] = \infty$ et *une seule* des deux quantités $E[f^+(X)]$ et $E[f^-(X)]$ est infinie, on peut toujours définir $E[f(X)]$ par (2.5). Dans les deux cas $|E[f(X)]| = +\infty$. On dit que $f(X)$ est *sommable*, mais non intégrable.

(B3) Si $E[f^+(X)]$ et $E[f^-(X)]$ sont en même temps infinis, on *renonce* à définir l'espérance $E[f(X)]$.

EXEMPLE 2.2.1: L'INDICATRICE COMME VARIABLE ALÉATOIRE. Soit $A \in \mathcal{F}$ un événement. On définit la variable aléatoire $X = 1_A$, où

$$1_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{si } \omega \notin A \end{cases}$$

On a en particulier

$$P(X = 1) = P(A) \text{ et } P(X = 0) = 1 - P(A),$$

si bien que

$$E[X] = 1 \times P(A) + 0 \times (1 - P(A)) = P(A).$$

D'où la formule triviale mais utile

$$E[1_A] = P(A). \tag{2.6}$$

Presque partout et presque sûrement

Soit $(p(x), x \in \mathcal{X})$ une distribution de probabilité sur \mathcal{X} .

Définition 2.2.1 Soit \mathcal{P} une propriété relative aux éléments x de \mathcal{X} . On dit que \mathcal{P} est vraie $p(x)$ - presque partout (abréviation : “p.p.”) si pour tout $x \in \mathcal{X}$ on a l'une ou l'autre (non exclusivement) des 2 occurrences suivantes :

- (a) x vérifie \mathcal{P} ,
- (b) $p(x) = 0$.

Soit X une variable aléatoire à valeurs dans \mathcal{X} , de distribution de probabilité $(p(x), x \in \mathcal{X})$. Il est clair que si f et g sont deux fonctions de \mathcal{X} dans $\overline{\mathbb{R}}$ telles que

$$f(x) = g(x), \text{ } p(x) - \text{presque partout,}$$

alors $E[f(X)]$ est bien définie dès que $E[g(X)]$ l'est définie, et

$$E[f(X)] = E[g(X)].$$

Définition 2.2.2 Soit X une variable aléatoire valeurs dans \mathcal{X} et \mathcal{P} une propriété relative aux éléments x de \mathcal{X} . Si

$$P(X \text{ vérifie } \mathcal{P}) = 1,$$

on dit que \mathcal{P} est vérifiée par X P -presque sûrement (abréviation : “p.s.”).

Lorsqu’aucune confusion n’est à craindre, on omet la mention de la probabilité et on abrège “ P -presque sûrement” en “presque sûrement” ou “p.s.” Il est clair que X vérifie \mathcal{P} P -presque sûrement si et seulement si \mathcal{P} est vraie $p(x)$ -presque partout.

Propriétés de l’espérance

A partir de la définition de l’espérance, les propriétés qui suivent sont évidentes.

Théorème 2.2.1 Linéarité. Soit f et g deux fonctions de \mathcal{X} dans \mathbb{R} telles que $f(X)$ et $g(X)$ soient intégrables. Soit a et b deux nombres réels. Alors $af(X) + bg(X)$ est intégrable et

$$E[af(X) + bg(X)] = aE[f(X)] + bE[g(X)] .$$

Cette égalité vaut également lorsque f et g sont des fonctions non négatives et a et b sont des réels non négatifs.

Monotonie. Soit f et g deux fonctions de \mathcal{X} dans $\overline{\mathbb{R}}$ telles que $f(X)$ et $g(X)$ admettent une espérance. Alors

$$f(X) \leq g(X), P\text{-p.s.} \Rightarrow E[f(X)] \leq E[g(X)] .$$

Mentionnons également l’inégalité

$$|E[f(X)]| \leq E[|f(X)|],$$

vraie dès que l’espérance $E[f(X)]$ est bien définie, et qui n’est autre que l’inégalité

$$|E[f^+(X)] - E[f^-(X)]| \leq E[f^+(X)] + E[f^-(X)] .$$

Moyenne et variance

Soit X une variable aléatoire à valeurs dans l’ensemble dénombrable \mathcal{X} de $\overline{\mathbb{R}}$ et admettant la distribution $(p(x), x \in \mathcal{X})$. Si X est intégrable, on définit la *moyenne* de X

$$m_X = E[X] .$$

Si X^2 est intégrable, alors X est intégrable (Exercice 2.4.1), et on définit la *variance* de X , notée σ_X^2 , par

$$\sigma_X^2 = E[X^2] - m_X^2 .$$

La variance mesure la dispersion autour de la valeur moyenne m_X , ainsi que le montre l'égalité (voir Exercice 2.4.3)

$$\sigma_X^2 = E[(X - m_X)^2] . \quad (2.7)$$

On appelle σ_X l'*écart-type* de X (on prend $\sigma_X \geq 0$).

EXEMPLE 2.2.2: POISSON. Soit $X \sim \mathcal{P}(\lambda)$. Des calculs élémentaires (voir Exercice 2.4.4) donnent

$$m_X = \lambda \text{ et } \sigma_X^2 = \lambda$$

EXEMPLE 2.2.3: BINOMIALE. Soit $X \sim \mathcal{B}(n, p)$. On trouve (voir Exercice 2.4.5) :

$$m_X = np \text{ et } \sigma_X^2 = np(1 - p) .$$

Inégalité de Markov

Voici maintenant un des outils fondamentaux du calcul des probabilités dont l'importance est proportionnelle à sa simplicité.

Théorème 2.2.2 *Soit Z une variable aléatoire X à valeurs dans \mathbb{R} et non négative et a un nombre réel positif. On a l'inégalité de Markov*

$$P(Z \geq a) \leq \frac{E[Z]}{a} . \quad (2.8)$$

En particulier, si X une variable aléatoire réelle, pour tout $\varepsilon > 0$, on a l'inégalité de Chebyshev

$$P(|X - m_X| \geq \varepsilon) \leq \frac{\sigma_X^2}{\varepsilon^2} . \quad (2.9)$$

(La preuve du théorème utilise des propriétés très générales de l'espérance, valables pour toutes les variables réelles, discrètes ou non.)

Démonstration. De l'inégalité

$$Z \geq a 1_{\{Z \geq a\}}$$

on tire (monotonie de l'espérance)

$$E[Z] \geq a E[1_{\{Z \geq a\}}] = P(Z \geq a) .$$

Pour obtenir (2.9), on applique (2.8) avec $f(x) = |x - m_X|^2$ et $a = \varepsilon^2$. □

Inégalité de Jensen

Théorème 2.2.3 Soit φ une fonction convexe définie sur un intervalle $I \subseteq \mathbb{R}$ contenant toutes les valeurs possibles d'une variable aléatoire X à valeurs dans \mathbb{R} . Alors si X et $\varphi(X)$ sont intégrables,

$$\varphi(E[X]) \leq E[\varphi(X)] . \quad (2.10)$$

(Voir la remarque qui suit l'énoncé du théorème précédent.)

Démonstration. On utilise la propriété suivante des fonctions convexes. Pour tout x_0 à l'intérieur (au sens topologique) de I il existe un réel $\alpha = \alpha(x_0)$ tel que pour tout $x \in I$,

$$\varphi(x) \leq \varphi(x_0) + \alpha(x - x_0) .$$

Si X est une constante (déterministe) alors l'inégalité (en fait une égalité dans ce cas) est triviale. Dans le cas contraire, $E[X]$ est bien à l'intérieur de I et donc, pour un certain réel α ,

$$\varphi(X) \leq \varphi(E[X]) + \alpha(X - E[X]) .$$

Si on prend les espérances de chacun des membres de cette inégalité, on trouve (2.10). \square

Somme de variables indépendantes

Soit X_1 et X_2 deux variables aléatoires discrètes à valeurs dans $\mathcal{X}_1 \subset \mathbb{R}$ et $\mathcal{X}_2 \subset \mathbb{R}$ respectivement. Soit $p(x_1, x_2)$, $p_1(x_1)$ et $p_2(x_2)$ les distributions de (X_1, X_2) , X_1 et X_2 respectivement. Posons $Z = X_1 + X_2$. On suppose X_1 et X_2 indépendantes, et donc

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) .$$

De l'identité ensembliste

$$\{X_1 + X_2 = z\} = \sum_{x_1 \in \mathcal{X}_1} \{X_2 + x_1 = z\} \cap \{X_1 = x_1\},$$

on tire

$$\begin{aligned} P(Z = z) &= P(X_1 + X_2 = z) \\ &= \sum_{x_1 \in \mathcal{X}_1} P(X_2 + x_1 = z, X_1 = x_1) \\ &= \sum_{x_1 \in \mathcal{X}_1} P(X_2 = z - x_1, X_1 = x_1), \end{aligned}$$

et donc, au vu de l'indépendance de X_1 et X_2 ,

$$p_Z(z) = \sum_{x_1 \in \mathcal{X}_1} p_1(x_1)p_2(z - x_1) . \quad (2.11)$$

Lorsque X_1 et X_2 (et donc Z) prennent leurs valeurs dans \mathbb{N} , on voit que p_Z est la *convolution* de p_1 et de p_2 :

$$p_Z(n) = \sum_{k=0}^{\infty} p_1(n-k) p_2(k) ,$$

ce qu'on note aussi $p_Z = p_1 * p_2$.

Théorème du produit des espérances

Théorème 2.2.4 *Soit Y et Z deux variables aléatoires discrètes indépendantes à valeurs dans \mathcal{Y} et \mathcal{Z} respectivement, et soit $f : \mathcal{Y} \rightarrow \mathbb{R}$ et $g : \mathcal{Z} \rightarrow \mathbb{R}$ deux fonctions qui sont soit non négatives, soit telles que $f(Y)$ et $g(Z)$ sont intégrables. Alors, dans le cas “intégrable”, le produit $f(Y)g(Z)$ est aussi intégrable, et dans tous les cas, on a la formule du produit des espérances :*

$$E[f(Y)g(Z)] = E[f(Y)] E[g(Z)] . \quad (2.12)$$

Démonstration. On traitera le cas “intégrable”, l'autre étant implicitement traité dans la démonstration qui suit. Posons $X = (Y, Z)$. C'est une variable aléatoire à valeurs dans $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$. On cherche à démontrer que $h(X) = f(Y)g(Z)$ est intégrable. D'après l'hypothèse d'indépendance $P(Y = y, Z = z) = P(Y = y)P(Z = z)$ et donc

$$E[|h(X)|] = \sum_x |h(x)|P(X = x) = \sum_{y,z} |f(y)||g(z)|P(Y = y)P(Z = z).$$

On peut séparer les y des z dans le deuxième membre de la dernière égalité (il s'agit en effet de séries dont le terme général est non négatif), pour obtenir

$$\begin{aligned} E[|h(X)|] &= \left(\sum_{y \in \mathcal{Y}} |f(y)|P(Y = y) \right) \left(\sum_{z \in \mathcal{Z}} |g(z)|P(Z = z) \right) \\ &= E[|f(Y)|] E[|g(Z)|] , \end{aligned}$$

et donc $E[|h(X)|] < \infty$. L'égalité (2.12) découle des mêmes calculs, sans les valeurs absolues, les interversions de l'ordre des intégrations étant licites dans le cas intégrable aussi bien que dans le cas non négatif. \square

Variance d'une somme de variables indépendantes

Théorème 2.2.5 *Soit X_1, X_2, \dots, X_n des variables aléatoires (discrètes dans ce chapitre, mais cette hypothèse n'est pas nécessaire) indépendantes à valeurs réelles. Alors :*

$$\sigma_{X_1 + \dots + X_n}^2 = \sigma_{X_1}^2 + \dots + \sigma_{X_n}^2 \quad (2.13)$$

(La variance d'une somme de variables indépendantes est la somme de leurs variances).

Démonstration. Il suffit de démontrer le résultat dans le cas $n = 2$.

$$\sigma_{X_1+X_2}^2 = E[((X_1 + X_2) - m_{X_1+X_2})^2]$$

Or $m_{X_1+X_2} = m_{X_1} + m_{X_2}$ et donc

$$\begin{aligned}\sigma_{X_1+X_2}^2 &= E[((X_1 - m_{X_1}) + (X_2 - m_{X_2}))^2] \\ &= E[(X_1 - m_{X_1})^2] + E[(X_2 - m_{X_2})^2] + 2E[(X_1 - m_{X_1})(X_2 - m_{X_2})] .\end{aligned}$$

Les variables X_1 et X_2 étant indépendantes, on a donc, d'après le Théorème 2.2.4,

$$E[(X_1 - m_{X_1})(X_2 - m_{X_2})] = E[X_1 - m_{X_1}]E[X_2 - m_{X_2}] = 0 ,$$

d'où

$$\sigma_{X_1+X_2}^2 = E[(X_1 - m_{X_1})^2] + E[(X_2 - m_{X_2})^2] .$$

□

Loi faible des grands nombres

La loi des grands nombres établit un lien entre probabilité et fréquence empirique. Elle dit par exemple que la fréquence empirique asymptotique des “face” dans le jeu de pile ou face avec une pièce non truquée est $\frac{1}{2}$ dans le sens suivant. Pour tout $\varepsilon > 0$,

$$\lim_{n \uparrow \infty} P \left(\left| \frac{S_n}{n} - \frac{1}{2} \right| \geq \varepsilon \right) = 0 .$$

Plus généralement, soit X_1, X_2, \dots une suite IID de variables aléatoires (discrètes dans ce chapitre, mais cette hypothèse n'est pas nécessaire), de moyenne $E[X_1] = m$ et de variance $\sigma^2 < \infty$. La variable aléatoire $S_n = X_1 + \dots + X_n$ est une variable de moyenne $E[S_n] = m_n = n \times m$ et de variance $\sigma_n^2 = n \times \sigma^2$. L'inégalité de Tchebychev donne

$$P \left(\left| \frac{S_n}{n} - m \right| \geq \varepsilon \right) \leq \frac{\sigma_n^2}{n^2 \varepsilon^2} = \frac{1}{n} \frac{\sigma^2}{\varepsilon^2} .$$

et donc, pour tout $\varepsilon > 0$,

$$\lim_{n \uparrow \infty} P \left(\left| \frac{X_1 + \dots + X_n}{n} - E[X_1] \right| \geq \varepsilon \right) = 0 . \quad (2.14)$$

C'est en ce sens que la moyenne empirique $\frac{X_1 + \dots + X_n}{n}$ tend vers la moyenne probabiliste $E[X_1]$ lorsque n tend vers l'infini. Le résultat (2.14) s'appelle la *loi faible des grands nombres*. (On verra plus tard la *loi forte* des grands nombres.)

EXEMPLE 2.2.4: L'APPROXIMATION POLYNOMIALE DE BERNSTEIN. On rencontre assez souvent des résultats d'analyse qui se démontrent aussi à l'aide d'arguments probabilistes. En voici un exemple parmi les plus simples.

Nous allons démontrer que pour toute fonction f continue de $[0, 1]$ dans \mathbb{R} , le polynôme (dit polynôme de Bernstein associé à f)

$$Q_n(x) = \sum_{k=0}^n f\left(\frac{k}{n}\right) C_n^k x^k (1-x)^{n-k}$$

converge vers f uniformément sur $[0, 1]$.

Pour ce faire, introduisons pour chaque $x \in [0, 1]$ une suite $X_n, n \geq 0$, de variables aléatoires indépendantes, de même loi : $P(X_n = 1) = x, P(X_n = 0) = 1 - x$. La variable aléatoire $S_n = X_1 + \dots + X_n$ suit donc une loi binomiale d'ordre n et de paramètre x :

$$P(S_n = k) = C_n^k x^k (1-x)^{n-k} \quad (0 \leq k \leq n).$$

On a donc

$$Q_n(x) = E \left[f\left(\frac{S_n}{n}\right) \right].$$

La fonction $f(x)$ étant continue sur $[0, 1]$, et donc uniformément continue sur cet intervalle, on peut associer à tout $\varepsilon > 0$ un nombre $\delta(\varepsilon) > 0$ tel que si $|y - x| < \delta(\varepsilon)$, alors $|f(x) - f(y)| \leq \varepsilon$. De plus $|f(x)|$ est bornée, disons par M , sur $[0, 1]$. Définissons pour un x fixé l'ensemble $A_\varepsilon = \{y \in \mathbb{R}; |y - x| < \delta(\varepsilon)\}$. On a

$$\begin{aligned} \left| f\left(\frac{S_n}{n}\right) - f(x) \right| &\leq \left| f\left(\frac{S_n}{n}\right) - f(x) \right| 1_{A_\varepsilon}\left(\frac{S_n}{n}\right) + \left| f\left(\frac{S_n}{n}\right) - f(x) \right| 1_{\bar{A}_\varepsilon}\left(\frac{S_n}{n}\right) \\ &\leq \varepsilon 1_{A_\varepsilon}\left(\frac{S_n}{n}\right) + 2M 1_{\bar{A}_\varepsilon}\left(\frac{S_n}{n}\right), \end{aligned}$$

d'où

$$E \left[\left| f\left(\frac{S_n}{n}\right) - f(x) \right| \right] \leq \varepsilon P \left(\left| \frac{S_n}{n} - x \right| < \delta(\varepsilon) \right) + 2M P \left(\left| \frac{S_n}{n} - x \right| \geq \delta(\varepsilon) \right)$$

On a donc la majoration

$$|Q_n(x) - f(x)| \leq \varepsilon + 2MP \left(\left| \frac{S_n}{n} - x \right| \geq \delta(\varepsilon) \right).$$

L'inégalité de Tchebychev donne :

$$P \left(\left| \frac{S_n}{n} - x \right| \geq \delta(\varepsilon) \right) \leq \frac{x(1-x)}{\delta^2(\varepsilon)} \frac{1}{n}$$

d'où

$$|Q_n(x) - f(x)| \leq \varepsilon + 2M \frac{x(1-x)}{\delta^2(\varepsilon)} \frac{1}{n} .$$

Donc

$$\lim_{n \uparrow \infty} \sup_{x \in [0,1]} |Q_n(x) - f(x)| \leq \varepsilon .$$

Ceci étant vrai pour tout $\varepsilon > 0$, on a le résultat annoncé. □

2.3 Fonctions génératrices

Définitions et exemples

Les fonctions génératrices sont un puissant outil de calcul. Pour définir cet objet mathématique, nous aurons besoin d'une extension immédiate de la définition de l'espérance, pour pouvoir considérer l'espérance de $f(X)$ dans le cas où f prend des valeurs complexes :

$$f(x) = f_1(x) + if_2(x) ,$$

où f_1 et f_2 sont des fonctions définies sur \mathcal{X} et à valeurs dans \mathbb{R} . On dit que $f(x)$ est intégrable si $f_1(X)$ et $f_2(X)$ sont intégrables et on définit alors

$$E[f(X)] = E[f_1(X)] + iE[f_2(X)] .$$

La propriété de linéarité de l'espérance est évidemment préservée.

Définition 2.3.1 Soit X une variable aléatoire discrète à valeurs dans \mathbb{N} . On appelle fonction génératrice de X l'application $g_X : \{s \in \mathbb{C}; |s| \leq 1\} \rightarrow \mathbb{C}$ définie par

$$g_X(s) = E[s^X] = \sum_{n=0}^{\infty} s^n P(X = n) . \quad (2.15)$$

EXEMPLE 2.3.1: POISSON. Soit $X \sim \mathcal{P}(\lambda)$. On a

$$g_X(s) = \sum_{n=0}^{\infty} e^{-\lambda} \frac{(\lambda s)^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda s)^n}{n!} ,$$

et donc

$$X \sim \mathcal{P}(\lambda) \Rightarrow g_X(s) = e^{\lambda(s-1)} .$$

EXEMPLE 2.3.2: BINOMIALE. Soit $X \sim \mathcal{B}(n, p)$. On a :

$$g_X(s) = \sum_{k=0}^{\infty} C_n^k p^k (1-p)^{n-k} s^k = \sum_{k=0}^{\infty} C_n^k (ps)^k (1-p)^{n-k},$$

et donc d'après la formule du binôme,

$$X \sim \mathcal{B}(n, p) \Rightarrow g_X(s) = (ps + q)^n.$$

Théorème 2.3.1 Soient X et Y deux variables aléatoires indépendantes à valeurs dans \mathbb{N} et de fonction génératrice g_X et g_Y . Alors la somme $X+Y$ a pour fonction génératrice $g_{X+Y}(s) = g_X(s)g_Y(s)$.

Démonstration. Le terme général de la série produit des séries $\sum P(X = n)s^n$ et $\sum P(Y = n)s^n$ est $(\sum_{k=0}^n P(X = k)P(Y = n-k))s^n$, c'est-à-dire, au vu de l'indépendance de X et Y et de la formule de convolution, $P(X + Y = n)s^n$. \square

EXEMPLE 2.3.3: La loi binomiale d'ordre n et de paramètre p est la distribution de la somme $X = Y_1 + \dots + Y_n$ de n variables indépendantes de distribution $P(Y_i = 0) = q = 1 - p$, $P(Y_i = 1) = p$. On a donc

$$g_X(s) = \prod_{i=1}^n g_{Y_i}(s) = (ps + q)^n$$

puisque $g_{Y_i}(s) = ps + q$. On retrouve donc le résultat de l'Exercice 2.3.2.

Somme aléatoire

Théorème 2.3.2 Soit N, X_1, X_2, \dots une famille de variables aléatoires indépendantes à valeurs dans \mathbb{N} . Les X_i , $i \geq 1$, sont identiquement distribuées de fonction génératrice commune g_X . Alors la fonction génératrice g_S de la somme aléatoire $S = X_1 + \dots + X_N$ ($S = 0$ si $N = 0$) est donnée par la formule

$$g_S(s) = g_N(g_X(s)), \tag{2.16}$$

où g_N est la fonction génératrice de N .

Démonstration. On a $g_S(s) = E[s^S] = E[s^{X_1 + \dots + X_N}]$. Or :

$$s^{X_1 + \dots + X_N} = \sum_{n=0}^{\infty} (1_{\{N=n\}} s^{X_1 + \dots + X_N}) = \sum_{n=0}^{\infty} (1_{\{N=n\}} s^{X_1 + \dots + X_n})$$

D'après l'additivité de l'espérance

$$E[s^{X_1+\dots+X_N}] = \sum_{n=0}^{\infty} E[1_{\{N=n\}} s^{X_1+\dots+X_n}] .$$

Comme N et (X_1, \dots, X_n) sont indépendantes, cette dernière quantité est égale à

$$\sum_{n=1}^{\infty} E[1_{\{N=n\}}] E[s^{X_1+\dots+X_n}] .$$

Mais $E[1_{\{N=n\}}] = P(N = n)$ (d'après (2.6)) et d'autre part, comme X_1, \dots, X_n sont indépendantes,

$$\begin{aligned} E[s^{X_1+\dots+X_n}] &= E\left[\prod_{i=1}^n s^{X_i}\right] \\ &= \prod_{i=1}^n E[s^{X_i}] = g_X(s)^n . \end{aligned}$$

On a donc

$$g_S(s) = \sum_{n=1}^{\infty} P(N = n) g_X(s)^n = g_N(g_X(s)) .$$

□

Dérivées successives et moments

La série $\sum_{n=0}^{\infty} P(X = n) s^n$ a un rayon de convergence R au moins égal à 1 puisque, pour $s = 1$, la série des valeurs absolues converge : $\sum_{n=0}^{\infty} P(X = n) = 1$. Ceci va nous permettre toutes les manipulations classiques sur les séries entières (en particulier la dérivation à tout ordre) à l'intérieur du disque *ouvert* $|s| < 1$. Ainsi la série dérivée

$$g'_X(s) = \sum_{n=1}^{\infty} n P(X = n) s^{n-1}$$

a un rayon de convergence supérieur ou égal à 1. De même pour la série dérivée seconde

$$g''_X(s) = \sum_{n=2}^{\infty} n(n-1) P(X = n) s^{n-2} .$$

Supposons que $R > 1$. Si X est intégrable, alors $E[X]$ est la valeur de la série dérivée première au point $s = 1$:

$$m_X = g'_X(1) . \quad (2.17)$$

En effet $g'_X(1) = \sum_{n=1}^{\infty} n P(X = n)$. Aussi :

$$E[X^2] = g''_X(1) g'_X(1) - (g'_X(1))^2 . \quad (2.18)$$

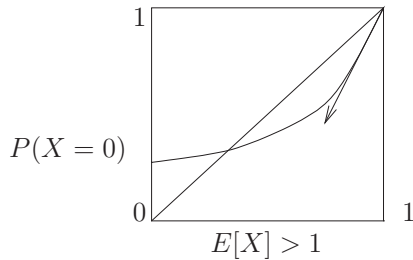
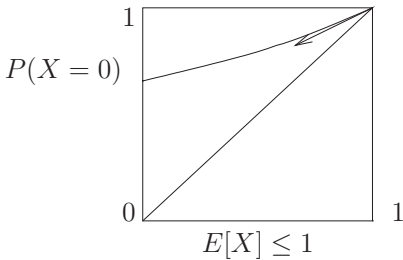
En effet

$$\begin{aligned} g_X''(1) &= \sum_{n=2}^{\infty} n^2 P(X = n) - \sum_{n=2}^{\infty} n P(X = n) \\ &= E[X^2] - P(X = 1) - (E[X] - P(X = 1)) \\ &= E[X^2] - E[X]. \end{aligned}$$

Si $R = 1$, on peut montrer (en utilisant le lemme d'Abel) que les formules obtenues restent vraies pourvu qu'on interprète $g'(1)$ et $g''(1)$ comme les limites, lorsque s tend vers 1 par valeurs réelles inférieures, de $g'(s)$ et $g''(s)$ respectivement.

Théorème 2.3.3 *(α) Soit $g : [0, 1] \rightarrow \mathbb{R}$ la fonction définie par $g(x) = E[x^X]$, où X est une variable aléatoire à valeurs entières. Alors g est non décroissante et convexe. De plus, si $P(X = 0) < 1$, alors g est strictement croissante, et si $P(X \leq 1) < 1$, elle est strictement convexe.*

(β) Supposons $P(X \leq 1) < 1$. Si $E[X] \leq 1$, l'équation $x = g(x)$ a une unique solution $x \in [0, 1]$, qui est $x = 1$. Si $E[X] > 1$, elle a 2 solutions dans $[0, 1]$, $x = 1$ et $x = \beta \in (0, 1)$.



Deux aspects de la fonction génératrice

Démonstration. Il suffit d'observer que pour tout $x \in [0, 1]$,

$$g'(x) = \sum_{n=1}^{\infty} n P(X = n) x^{n-1} \geq 0,$$

et donc g est non décroissante, et que

$$g''(x) = \sum_{n=2}^{\infty} n(n-1) P(X = n) x^{n-2} \geq 0,$$

et donc g est convexe. Pour que $g'(x)$ s'annule pour un $x \in (0, 1)$, il faut que $P(X = n) = 0$ pour tout $n \geq 1$, et donc $P(X = 0) = 1$. Donc, si $P(X = 0) < 1$, g' ne s'annule

pas, et g est strictement croissante. Pour que $g''(x)$ s'annule pour un $x \in (0, 1)$, il faut et il suffit que $P(X = n) = 0$ pour tout $n \geq 2$, et donc que $P(X = 0) + P(X = 1) = 1$. Donc, si $P(X = 0) + P(X = 1) < 1$, g' et g'' ne s'annulent pas, et g est strictement croissante et strictement convexe.

Le graphe de $g : [0, 1] \rightarrow \mathbb{R}$ a, dans le cas strictement croissant strictement convexe $P(X = 0) + P(X = 1) < 1$, la forme générale donnée dans la figure ci-dessous où l'on distingue deux cas : $E[X] = g'(1) \leq 1$, and $E[X] = g'(1) > 1$. Le reste de la preuve suit facilement. \square

Caractérisation de la loi par la fonction génératrice

Théorème 2.3.4 *La fonction génératrice caractérise la distribution d'une variable aléatoire discrète à valeurs dans \mathbb{N} .*

Plus précisément : si l'on sait que $g(s)$ est une fonction caractéristique, alors il n'existe qu'une seule distribution de probabilité $(p(n), n \geq 0)$ sur \mathbb{N} telle que

$$g(s) = \sum_{n=0}^{\infty} p(n) s^n \quad (|s| \leq 1) . \quad (2.19)$$

En effet, on sait que $g(s)$ est une fonction caractéristique et qu'elle admet donc la représentation (2.19) pour une distribution $(p(n), n \geq 0)$. L'unicité de $(p(n), n \geq 0)$ découle de l'unicité du développement en série entière autour de l'origine.

Le Théorème 2.3.4 permet une démonstration alternative du fait que la somme de deux variables de Poisson indépendantes est une variable de Poisson :

Théorème 2.3.5 *Soit X_1 et X_2 , deux variables de Poisson indépendantes, de paramètres λ_1 et λ_2 respectivement. Alors la somme $X_1 + X_2$ est aussi une variable de Poisson, de paramètre $\lambda_1 + \lambda_2$.*

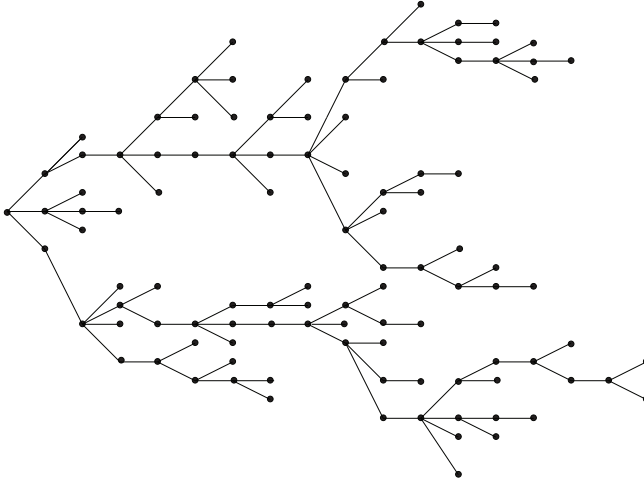
Démonstration. On a :

$$g_{X_1+X_2}(s) = g_{X_1}(s)g_{X_2}(s) = e^{\lambda_1(s-1)}e^{\lambda_2(s-1)} = e^{(\lambda_1+\lambda_2)(s-1)} . \quad (2.20)$$

La fonction génératrice de $X_1 + X_2$ est donc la fonction génératrice d'une variable de Poisson de paramètre $\lambda_1 + \lambda_2$, d'où le résultat, d'après Théorème 2.3.4. \square

EXEMPLE 2.3.4: LE PROCESSUS DE BRANCHEMENT. L'Anglais Galton qui s'intéressait à la survivance des noms de famille dans la haute société nobiliaire d'Angleterre est le fondateur de la théorie des processus de branchement, dont voici le modèle le plus simple.

$$\begin{array}{cccccccc}
x_1=3 & x_3=6 & x_5=6 & x_7=7 & x_9=7 & x_{11}=8 & x_{13}=9 & x_{15}=3 & x_{17}=2 \\
x_0=1 & x_2=6 & x_4=6 & x_6=9 & x_8=6 & x_{10}=10 & x_{12}=10 & x_{14}=6 & x_{16}=1 & x_{18}=0
\end{array}$$



On suppose que tous les individus d'une colonie (par exemple les mâles, porteurs et transmetteurs du nom, dans une société patriarcale donnée) donnent naissance au cours de leur vie à un certain nombre, aléatoire, d'enfants (dans le cas étudié par Galton, les enfants mâles). Chaque individu procrée indépendamment des autres individus de sa génération. Si on note X_n le nombre d'individus de la n -ème génération, on a

$$X_{n+1} = \begin{cases} \sum_{i=1}^{X_n} Z_i^{(n)} & \text{si } X_n \geq 1 \\ 0 & \text{si } X_n = 0 \end{cases}, \quad (2.21)$$

où les $Z_i^{(n)}$, $n \geq 0, i \geq 1$, sont des variables aléatoires indépendantes et identiquement distribuées de fonction génératrice commune

$$g_Z(s) = \sum_{n=0}^{\infty} P(Z = n) s^n,$$

de moyenne m_Z finie. La suite X_n , $n \geq 0$, s'appelle un *processus de branchement*.

Notons φ_n la fonction caractéristique de X_n :

$$\varphi_n(s) = \sum_{k=0}^{\infty} P(X_n = k) s^k.$$

De l'équation d'évolution (2.21) et du Théorème 2.3.2, on tire l'identité

$$\varphi_{n+1}(s) = \varphi_n(g_Z(s)).$$

En itérant cette relation, on obtient

$$\varphi_{n+1}(s) = \varphi_0(\underbrace{g_Z \circ \cdots \circ g_Z}_{n+1})(s) ,$$

où φ_0 est la fonction caractéristique de X_0 . En particulier, s'il n'y a qu'un seul ancêtre ($X_0 = 1$), $\varphi_0(s) = s$, et donc

$$\varphi_{n+1}(s) = g_Z(\varphi_n(s)) . \quad (2.22)$$

La *probabilité extinction* P_e du processus de branchement est la probabilité pour qu'il existe une génération vide (auquel cas toutes les suivantes sont vides), c'est-à-dire

$$P_e = P\left(\bigcup_{n=0}^{\infty} \{X_n = 0\}\right) . \quad (2.23)$$

On observe que $\{X_n = 0\} \subseteq \{X_{n+1} = 0\}$ puisque $X_n = 0$ entraîne que $X_{n+1} = 0$. D'après le théorème de continuité séquentielle de la probabilité, on a donc

$$P_e = \lim_{n \uparrow \infty} P(X_n = 0) .$$

Observons que $P(X_n = 0) = \varphi_n(0)$. De la relation de récurrence (2.22) on tire

$$P(X_{n+1} = 0) = g_Z(P(X_n = 0)) ,$$

et donc, en passant à la limite :

$$P_e = g_Z(P_e) . \quad (2.24)$$

Le Théorème 2.3.3 permet de conclure :

Si $m_Z \leq 1$, alors il n'y a qu'une solution à l'équation (2.24), $P_e = 1$, il y a presque sûrement extinction.

Si $m_Z > 1$, l'équation (2.24) a 2 solutions, 1 et $\beta \in (0, 1)$. En fait $P_e = \beta$ car $P_e = \lim_{n \uparrow \infty} x_n$, où $x_n = P(X_n = 0)$, et en particulier $x_0 = 0$. Il est facile de voir qu'avec une telle condition initiale, $\beta - x_n$ reste non négative, et donc que P_e ne peut être qu'égal à β . Il y a donc dans ce cas une probabilité positive mais strictement inférieure à 1 d'extinction.

EXEMPLE 2.3.5: L'ART DE COMPTER. Les fonctions génératrices sont un outil fort utilisé en Combinatoire, qui est l'art de compter. Elles jouent un rôle analogue dans le calcul des probabilités, lorsqu'on a à faire un dénombrement de toutes les situations possibles. Un exemple illustre ce point.

Soit X_1, X_2, X_3, X_4, X_5 , et X_6 des variables aléatoires indépendantes uniformément distribuées sur $\{0, 1, \dots, 9\}$. Chacune représente le tirage d'un nombre à six chiffres dans une loterie. Le but ici est de calculer la probabilité pour que la somme des trois premiers

chiffres soit égale à la somme des trois derniers. Pour ce faire nous allons calculer la fonction génératrice de la variable entière non négative $Y = 27 + X_1 + X_2 + X_3 - X_4 - X_5 - X_6$. (L'addition de 27 garantit que Y est non négative.) Il est clair que la probabilité cherchée est $P(Y = 27)$, c'est-à-dire le coefficient de s^{27} dans le développement en série entière de la fonction génératrice de Y , que nous devons donc calculer. On a

$$\begin{aligned} E[s^{X_i}] &= \frac{1}{10}(1 + s + \cdots + s^9) \\ &= \frac{1}{10} \frac{1 - s^{10}}{1 - s}, \end{aligned}$$

$$\begin{aligned} E[s^{-X_i}] &= \frac{1}{10} \left(1 + \frac{1}{s} + \cdots + \frac{1}{s^9} \right) \\ &= \frac{1}{10} \frac{1 - s^{-10}}{1 - s^{-1}} = \frac{1}{10} \frac{1}{s^9} \frac{1 - s^{10}}{1 - s}, \end{aligned}$$

et

$$\begin{aligned} E[s^Y] &= E\left[s^{27 + \sum_{i=1}^3 X_i - \sum_{i=4}^6 X_i}\right] \\ &= E\left[s^{27} \prod_{i=1}^3 s^{X_i} \prod_{i=4}^6 s^{-X_i}\right] \\ &= s^{27} \prod_{i=1}^3 E[s^{X_i}] \prod_{i=4}^6 E[s^{-X_i}]. \end{aligned}$$

Donc,

$$g_Y(s) = \frac{1}{10^6} \frac{(1 - s^{10})^6}{(1 - s)^6}.$$

Comme

$$(1 - s^{10})^6 = 1 - \binom{6}{1}s^{10} + \binom{6}{2}s^{20} - \cdots + s^{60}$$

et

$$(1 - s)^{-6} = 1 + \binom{6}{5}s + \binom{7}{5}s^2 + \binom{8}{5}s^3 + \cdots,$$

(rappelons la formule du binôme négative

$$(1 - s)^{-p} = 1 + \binom{p}{p-1}s + \binom{p+1}{p-1}s^2 + \binom{p+2}{p-1}s^3 + \cdots)$$

on trouve

$$P(Y = 27) = \frac{1}{10^6} \left(\binom{32}{5} - \binom{6}{1} \binom{22}{5} + \binom{6}{2} \binom{12}{5} \right),$$

le coefficient de s^{27} dans le développement de $g_Y(s)$.

2.4 Exercices

Exercice 2.4.1.

Soit X une variable aléatoire discrète à valeurs réelles. Montrez que si X^2 est intégrable, alors X l'est aussi.

Exercice 2.4.2. VARIABLE FINIE D'ESPÉRANCE INFINIE.

Donnez un exemple de variable aléatoire à valeurs entières finies dont l'espérance est infinie.

Exercice 2.4.3.

Montrez que $E[(X - m_X)^2] = E[X^2] - m_X^2$, où X est une variable aléatoire discrète à valeurs réelles de carré intégrable dont la moyenne est m_X .

Exercice 2.4.4.

Calculez directement (sans passer par les fonctions génératrices) la moyenne et la variance d'une variable aléatoire de Poisson de paramètre λ .

Exercice 2.4.5.

Calculez directement (sans passer par les fonctions génératrices) la moyenne et la variance d'une variable aléatoire binomiale d'ordre n et de paramètre p .

Exercice 2.4.6. * POISSON PAIR ET IMPAIR.

Soit X une variable de Poisson de moyenne $\theta > 0$. Quelle est la probabilité que X soit pair ? impair ?

Exercice 2.4.7. LA VARIABLE ALÉATOIRE GÉOMÉTRIQUE N'A PAS DE MÉMOIRE.

On dit que la variable aléatoire T à valeurs entières positives est une *variable aléatoire géométrique* de paramètre p , $0 < p < 1$, si

$$P(T = n) = pq^{n-1} .$$

Montrez que, pour tout $n_0 \geq 0$ et tout $n > 1$:

$$P(T \geq n_0 + n | T > n_0) = P(T \geq n) .$$

Exercice 2.4.8. FORMULE TÉLÉSCOPIQUE.

Soit X une variable aléatoire prenant ses valeurs dans \mathbb{N} . Montrer que :

$$E[X] = \sum_{n=0}^{\infty} P(X > n) .$$

Exercice 2.4.9. * LANCER DE DÉS.

On lance trois dés. Quelle est la probabilité pour que le chiffre indiqué par l'un des dés soit égale à la somme des chiffres indiqués par les deux autres ? On donnera deux solutions, une directe et l'autre utilisant les fonctions génératrices.

Exercice 2.4.10.

Calculez la moyenne et la variance d'une variable aléatoire de Poisson de paramètre λ en utilisant l'expression de la fonction génératrice de cette variable.

Exercice 2.4.11.

Calculer la moyenne et la variance d'une variable aléatoire binomiale d'ordre n et de paramètre p , $0 < p < 1$, en utilisant l'expression de la fonction génératrice de cette variable.

Exercice 2.4.12.

On rappelle que la variable T à valeurs entières positives est dite géométrique de paramètre p , $0 < p < 1$ si

$$P(T = n) = pq^{n-1} \quad (n \geq 1) .$$

Calculer sa fonction génératrice, et à l'aide de cette dernière, la moyenne et la variance de T .

Exercice 2.4.13. * LE BLUE PINKO D'AUSTRALIE.

L'incroyable (et hautement improbable) oiseau d'Australie, connu sous le nom de "blue pinko" pond un nombre aléatoire T d'œufs, soit bleus, soit roses. T est une variable aléatoire de Poisson de moyenne $\lambda > 0$. Les œufs d'une même couvée prennent leur couleur indépendamment les uns des autres et du nombre total d'œufs. La probabilité d'un œuf rose est $p \in (0, 1)$. Calculer la fonction génératrice du nombre total d'œufs roses et en déduire la distribution de ce nombre. Démontrez que le nombre d'œufs roses et le nombre d'œufs bleus sont indépendants.

Exercice 2.4.14. IDENTITÉ DE WALD.

Soit T, X_1, X_2, \dots une suite de variables aléatoires intégrables indépendantes à valeurs dans \mathbb{N}_+ . Les X_i , $i \geq 1$, sont identiquement distribuées. On définit $S = X_1 + \dots + X_T$. Montrez que

$$E[S] = E[X_1]E[T] .$$

Exercice 2.4.15. MOYENNE ET VARIANCE DU PROCESSUS DE BRANCHEMENT.

Calculez, pour le processus de branchement, la moyenne m_n et la variance σ_n^2 du nombre d'individus X_n de la n -ème génération, lorsque $X_0 = 1$. Calculez ces mêmes quantités lorsque $X_0 = k > 1$.

Exercice 2.4.16.

Soit X une variable aléatoire discrète à valeurs dans \mathcal{X} , de distribution de probabilité $(p(x), x \in \mathcal{X})$. Montrez que $P(p(X) > 0) = 1$.

Exercice 2.4.17. L'ENTOMOLOGISTE.

Chaque individu d'une certaine espèce d'insectes a, indépendamment des autres, une probabilité θ d'être une femelle. Un entomologiste cherche à capturer $M > 1$ femelles, et arrête donc d'en capturer dès qu'il a atteint le quota fixé. Quelle est la distribution X du nombre d'insectes qu'il aura à capturer dans ce but ?

Exercice 2.4.18. * LE RETOUR DE L'ENTOMOLOGISTE.

(suite de l'Exercice 2.4.17.) Quelle est la distribution de probabilité de Y , le plus petit nombre d'insectes que l'entomologiste doit capturer pour obtenir au moins M femelles et au moins N mâles ?

Chapitre 3

Vecteurs aléatoires

3.1 Distribution des vecteurs aléatoires

Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire réel à n dimensions, c'est-à-dire un vecteur dont toutes les composantes X_i sont des variables aléatoires à valeurs réelles. Soit $f_X : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction non négative telle que

$$\int_{\mathbb{R}^n} f_X(x) dx = 1 .$$

On dit que X admet la *densité de probabilité* f_X si pour tous $a, b \in \mathbb{R}^n$,

$$P(a_1 \leq X_1 \leq b_1, \dots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f_X(x_1, \dots, x_n) dx_1 \dots dx_n$$

Pour des raisons d'économie dans les notations, on notera souvent l'égalité ci-dessus

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

et semblablement pour d'autres expressions faisant intervenir des vecteurs.

Soit X et Y deux vecteurs aléatoires réels de dimensions respectives ℓ et m , et tels que le vecteur aléatoire (X, Y) de dimension $\ell + m$ admette la densité de probabilité $f_{X,Y}(x, y)$. Alors le vecteur X admet la densité de probabilité $f_X(x)$ donnée par

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy .$$

En effet, pour tous a, b , vecteurs réels de dimension ℓ , on a, par définition de la densité de probabilité,

$$\begin{aligned} P(a \leq X \leq b) &= P((X, Y) \in [a, b] \times \mathbb{R}^m) \\ &= \int_a^b \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dx dy \\ &= \int_a^b \left(\int_{-\infty}^{+\infty} f_{X,Y}(x, y) dy \right) dx . \end{aligned}$$

On appelle f_X la *densité de probabilité marginale* de (X, Y) sur les ℓ -premiers composantes. Le cas des densités marginales sur des composantes quelconques se traite de la même manière.

Espérance mathématique

Dans le cas où la fonction $g : \mathbb{R}^n \rightarrow \mathbb{R}$ soit est non négative, soit vérifie la condition d'intégrabilité

$$\int_{-\infty}^{+\infty} |g(x)| f_X(x) dx < \infty ,$$

on définit l'espérance de $g(X)$, notée $E[g(X)]$ par

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx .$$

Sous la condition d'intégrabilité ci-dessus, on dit que $g(X)$ est *intégrable*, et l'espérance $E[g(X)]$ est alors finie. Si g est simplement non négative, cette espérance peut être infinie.

On définira l'espérance aussi dans le cas suivant : l'une des espérances $E[g^+(X)]$ et $E[g^-(X)]$ est infinie, l'autre étant finie. On pose alors $E[g(X)] = E[g^+(X)] - E[g^-(X)]$, qui est une quantité de valeur absolue infinie.

Dans le cas où $E[g^+(X)]$ et $E[g^-(X)]$ sont toutes deux infinies, on renonce à définir $E[g(X)]$.

Si g est à valeurs complexes, de partie réelle g_R et de partie imaginaire g_I , on définit

$$E[g(X)] = E[g_R(X)] + E[g_I(X)]$$

si et seulement si $g_R(X)$ et $g_I(X)$ sont intégrables, c'est-à-dire, si et seulement si $E[|g(X)|] < \infty$.

Les propriétés de linéarité et de monotonie de l'intégrale entraînent la linéarité et la monotonie de l'espérance ainsi définie (voir Théorème 2.1.2).

Vecteur moyenne et matrice de covariance

Si les variables aléatoires du type $g(X) = X_i$ sont intégrables, on définit le vecteur (colonne) m_X , appelé *vecteur moyenne* de X , par

$$m_X = \begin{pmatrix} m_{X_1} \\ \vdots \\ m_{X_n} \end{pmatrix} = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix} = E[X].$$

Si de plus les variables aléatoires du type $g(X) = X_i X_j$ sont intégrables¹, on définit la *matrice de covariance* Σ_X du vecteur aléatoire X par

$$\Sigma_X = \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} & \cdots & \sigma_{X_1 X_n} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 & \cdots & \sigma_{X_2 X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_n X_1} & \sigma_{X_n X_2} & \cdots & \sigma_{X_n}^2 \end{pmatrix},$$

où

$$\sigma_{X_i X_j} = E[(X_i - m_{X_i})(X_j - m_{X_j})] \quad \text{et} \quad \sigma_{X_i}^2 = E[(X_i - m_{X_i})^2].$$

Dans le cas d'une variable aléatoire réelle X telle que $E[X^2] < \infty$, on parle de la *variance* $V(X) = E[(X - m_X)^2]$, également notée σ_X^2 , où $\sigma_X \geq 0$ est appelé l'*écart-type*.

Nous allons maintenant présenter les variables aléatoires avec densité de probabilité les plus fréquemment rencontrées.

Variable uniforme

Définition 3.1.1 *On dit que la variable aléatoire réelle X est uniforme sur l'intervalle fini $[a, b]$ si elle admet la densité*

$$f_X(x) = \frac{1}{b-a} 1_{\{a \leq x \leq b\}}.$$

On note ceci : $X \sim \mathcal{U}(a, b)$.

Des calculs élémentaires montrent que la moyenne d'une variable uniforme sur l'intervalle fini $[a, b]$ est

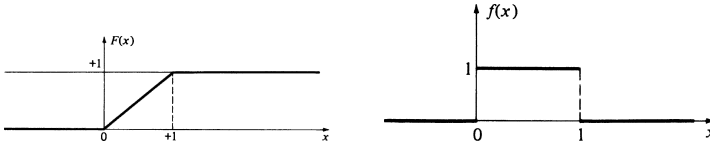
$$m_X = \frac{a+b}{2}$$

et que sa variance est donnée par la formule

$$\sigma_X^2 = \frac{(b-a)^2}{12}.$$

La variable uniforme sur $[0, 1]$ est dite *variable uniforme standard*.

¹On verra dans la Section 3.3 que pour que les variables X_i et $X_i X_j$ soient intégrables, il suffit que les variables X_i^2 et X_j^2 le soient.



Variable uniforme standard
(fonction de répartition et densité de probabilité)

Variable exponentielle

Définition 3.1.2 La variable aléatoire réelle non négative X est dite exponentielle (de paramètre λ) si elle admet la densité

$$f_X(x) = \lambda e^{-\lambda x} 1_{\{x \geq 0\}}.$$

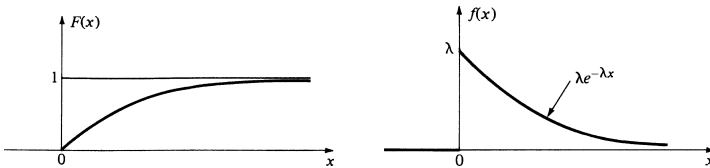
On note ceci : $X \sim \mathcal{E}(\lambda)$.

On vérifie que X est bien une densité, c'est-à-dire $\int_0^\infty \lambda e^{-\lambda x} dx = 1$. D'autre part la fonction de répartition $F_X(x) = P(X \leq x)$ est donnée par

$$F_X(x) = 1 - e^{-\lambda x} 1_{\{x \geq 0\}}.$$

L'égalité $1/\lambda = \int_0^\infty x e^{-\lambda x} dx$ montre que sa moyenne est

$$m_X = 1/\lambda.$$



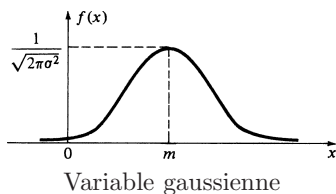
Variable exponentielle
(fonction de répartition et densité de probabilité)

Variable gaussienne

Définition 3.1.3 La variable aléatoire réelle X est dite gaussienne de moyenne m et de variance σ^2 si elle admet la densité de probabilité

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x-m)^2}{\sigma^2}}.$$

On note ceci : $X \sim \mathcal{N}(m, \sigma^2)$.



On peut vérifier que $f_X(X)$ est bien une densité, c'est-à-dire :

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}} dx = 1 .$$

De même, on peut vérifier que m et σ^2 sont bien la moyenne et la variance de X , c'est-à-dire :

$$m = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}} dx$$

$$\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-m)^2 e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}} dx .$$

(Exercice 3.5.1.)

Variable de loi gamma

Définition 3.1.4 On dit que la variable aléatoire réelle non négative X suit une loi gamma de paramètres α et β , ($\alpha > 0, \beta > 0$) si elle admet la densité

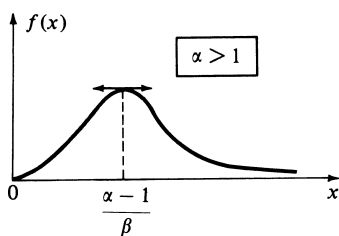
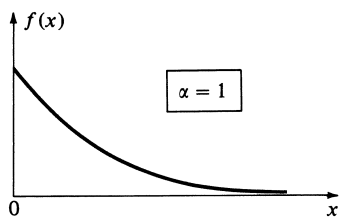
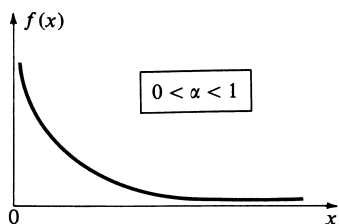
$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} 1_{\{x \geq 0\}}$$

où la fonction Γ est définie par

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du .$$

On note ceci : $X \sim \gamma(\alpha, \beta)$.

Il y a trois types de lois gamma, selon que le paramètre α est égal, supérieur ou inférieur à 1 (voir les Figures ci-dessous).



Lorsque $\alpha = 1$, on retrouve la loi exponentielle de paramètre β .

Une intégration par parties montre que la fonction Γ vérifie l'équation fonctionnelle

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \quad (\alpha > 1).$$

Comme $\Gamma(1) = 1$, on a en particulier

$$\Gamma(n) = (n - 1)!$$

On a aussi :

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

La variable aléatoire X de loi gamma a les caractéristiques suivantes (voir l'Exercice 3.5.2) :

$$m_X = \frac{\alpha}{\beta}$$

$$\sigma_X^2 = \frac{\alpha}{\beta^2}$$

Définition 3.1.5 La loi gamma de paramètres $\alpha = \frac{n}{2}$ et $\beta = \frac{1}{2}$, c'est-à-dire de densité de probabilité

$$f_X(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x} 1_{\{x \geq 0\}}.$$

s'appelle la loi du khi-deux à n degrés de liberté. On note ceci : $X \sim \chi_n^2$. La loi du khi-deux à un degré s'appelle parfois loi de Rayleigh :

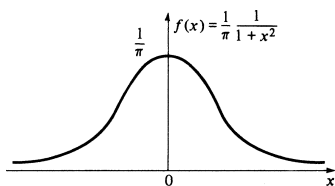
$$f_X(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{1}{2}x} 1_{\{x \geq 0\}}.$$

Variable de Cauchy

Définition 3.1.6 Si la variable aléatoire X admet la densité de probabilité

$$f_X(x) = \frac{1}{\pi(1+x^2)},$$

on dit que X est une variable aléatoire de Cauchy. On note ceci : $X \sim \text{Cauchy}$.



La moyenne d'une variable aléatoire de Cauchy n'est pas définie puisque cette variable n'est pas intégrable. En effet :

$$\begin{aligned} E[|X|] &= \int_{-\infty}^{+\infty} |x| f_X(dx) \\ &= \frac{2}{\pi} \int_{-\infty}^{+\infty} \frac{|x|}{1+x^2} dx = +\infty. \end{aligned}$$

EXEMPLE 3.1.1: UN PETIT EXEMPLE JUSTE POUR SE DIVERTIR. Un quidam choisit deux nombres réels, et ne vous dit pas lesquels pour l'instant. Il écrit chacun des nombres sur un morceau de papier, et il place les morceaux de papier dans deux boîtes différentes, qu'il prend soin de fermer. Il vous demande de choisir une de ces boîtes au hasard, selon le résultat d'un lancer de pièce de monnaie (non truquée), ce que vous faites. Il ouvre alors la boîte, vous fait lire le morceau de papier qu'elle contient, et vous pose la question suivante : est-ce le plus grand des nombres qu'il a choisis ?

Existe-t-il une stratégie de réponse qui vous garantit que vous avez plus de chances de donner la bonne réponse que la mauvaise ? A priori, il semble que non, et que vous avez autant de chances de vous tromper que de tomber sur la bonne réponse. Et pourtant...

Voici comment procéder pour gruger la Fortune aveugle. Choisissez un nombre de référence Z de manière aléatoire, avec la densité de probabilité (quelconque) f . Si Z est strictement plus grand que le nombre dévoilé, pariez que le nombre resté secret est plus grand que celui qui a été dévoilé. Nous allons montrer que c'est une stratégie avantageuse.

Soit a et b , $a < b$ les nombres choisis par le quidam. Notons X le nombre que vous avez pu lire. La probabilité que vous avez deviné juste est

$$\begin{aligned} P(X = a, Z \geq a) + P(X = b, Z < b) &= P(X = a)P(Z \geq a) + P(X = b)P(Z < b) \\ &= \frac{1}{2}P(Z \geq a) + \frac{1}{2}P(Z < b) \\ &= \frac{1}{2} \int_a^{+\infty} f(x)dx + \frac{1}{2} \int_{-\infty}^b f(x)dx \\ &= \frac{1}{2} + \frac{1}{2} \int_a^b f(x)dx. \end{aligned}$$

Or ceci est strictement supérieur à $\frac{1}{2}$ dès que f est partout positive.

Fonction caractéristique

Définition 3.1.7 La fonction $\varphi_X : \mathbb{R}^n \rightarrow \mathbb{C}$ définie par

$$\varphi_X(u) = E[e^{iu^T X}]$$

est la fonction caractéristique du vecteur aléatoire X .

Les exemples classiques qui suivent concernent des variables aléatoires.

EXEMPLE 3.1.2: FONCTION CARACTÉRISTIQUE DE LA VARIABLE GAUSSIENNE. La fonction caractéristique de la variable gaussienne $X \sim \mathcal{N}(m, \sigma^2)$ est (voir Exercice 3.5.1)

$$\varphi_X(u) = e^{ium - \frac{1}{2}\sigma^2 u^2}.$$

EXEMPLE 3.1.3: FONCTION CARACTÉRISTIQUE DE LA VARIABLE UNIFORME. La fonction caractéristique de la variable aléatoire uniforme $X \sim \mathcal{U}(a, b)$ est, comme le montre un calcul immédiat,

$$\varphi_X(u) = \frac{e^{iub} - e^{iua}}{iu(b-a)}.$$

EXEMPLE 3.1.4: FONCTION CARACTÉRISTIQUE DE LA VARIABLE GAMMA. La fonction caractéristique de la variable aléatoire $X \sim \gamma(\alpha, \beta)$ est, comme le montrerait une intégration dans le plan complexe,

$$\varphi_X(u) = \left(1 - \frac{iu}{\beta}\right)^{-\alpha}.$$

En particulier, la fonction caractéristique de la variable exponentielle $X \sim \mathcal{E}(\lambda)$ est

$$\varphi_X(u) = \frac{\lambda}{\lambda - iu}.$$

La fonction caractéristique mérite son nom car elle caractérise la loi d'un vecteur aléatoire :

Théorème 3.1.1 *Si deux n -vecteurs aléatoires X et Y ont la même fonction caractéristique, alors*

$$P(X \leq x) = P(Y \leq x) \quad (x \in \mathbb{R}^n).$$

La démonstration complète de ce résultat sort du cadre de ce cours élémentaire. Dans le cas particulier où X et Y admettent des densités de probabilité continues f_X et f_Y respectivement, et où leur fonction caractéristique commune est de module intégrable, le résultat découle du théorème d'inversion de Fourier, puisque dans les conditions ci-dessus, f_X et f_Y sont les transformées de Fourier inverses de la même fonction :

$$f_X(x) = f_Y(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-ix^T u} \varphi(u) du.$$

Vecteurs aléatoires indépendants

Définition 3.1.8 *On dit que les vecteurs aléatoires X_1, \dots, X_ℓ à valeurs dans $\mathbb{R}^{n_1}, \dots, \mathbb{R}^{n_\ell}$ respectivement, sont indépendants (dans leur ensemble) si pour tous sous-ensembles $A_1 \subset \mathbb{R}^{n_1}, \dots, A_\ell \subset \mathbb{R}^{n_\ell}$ on a*

$$P\left(\bigcap_{i=1}^{\ell} \{X_i \in A_i\}\right) = \prod_{i=1}^{\ell} P(X_i \in A_i).$$

Théorème 3.1.2 *Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire de \mathbb{R}^n admettant une densité de probabilité f_X et telle que les variables aléatoires X_1, \dots, X_n soit indépendantes. On a alors*

$$f_X(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i),$$

où les f_{X_i} sont les densités de probabilité des X_i . Inversement, si f_X a la forme produit $\prod_{i=1}^n f_{X_i}(x_i)$, où les f_i sont des densités de probabilité, alors X_1, \dots, X_n sont indépendantes, et les f_i sont les densités de probabilité des X_i .

On ne démontrera pas ce théorème ici (voir l'Exercice 6.4.18). Son contenu est intuitif si l'on assimile $f_X(x_1, \dots, x_n)\Delta x_1 \dots \Delta x_n$ à $P(\cap_{i=1}^n \{x_i < X_i \leq x_i + \Delta x_i\})$ et $f_{X_i}(x_i)\Delta x_i$ à $P(x_i \leq X_i + \Delta x_i)$, car l'indépendance entraîne alors

$$P(\cap_{i=1}^n \{x_i < X_i \leq x_i + \Delta x_i\}) = \prod_{i=1}^n P(x_i < X_i \leq x_i + \Delta x_i) .$$

c'est-à-dire

$$f_X(x_1, \dots, x_n)\Delta x_1 \dots \Delta x_n = f_{X_1}(x_1)\Delta x_1 \dots f_{X_n}(x_n)\Delta x_n .$$

Le théorème est vrai, *mutatis mutandis*, aussi lorsque les X_i sont des vecteurs.

Théorème 3.1.3 *Soit X et Y deux vecteurs aléatoires indépendants, à valeurs dans \mathbb{R}^ℓ et \mathbb{R}^m , et admettant les densités de probabilité f_X et f_Y . Soit $g : \mathbb{R}^\ell \rightarrow \mathbb{C}$ et $h : \mathbb{R}^m \rightarrow \mathbb{C}$ deux fonctions mesurables telles que $g(X)$ et $h(Y)$ soient intégrables. Alors $g(X)h(Y)$ est intégrable et*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)] .$$

Cette dernière égalité vaut également lorsque g et h sont des fonctions réelles non négatives.

Démonstration. Il s'agit d'une conséquence immédiate du théorème de Fubini puisque la densité de (X, Y) est le produit des densités de X et Y . \square

Voici un autre critère d'indépendance, faisant cette fois-ci intervenir les fonctions caractéristiques.

Théorème 3.1.4 *Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire admettant la densité de probabilité f_X et la fonction caractéristique φ_X . Pour que les variables aléatoires X_1, \dots, X_n soient indépendantes dans leur ensemble, il faut et il suffit que*

$$\varphi_X(u_1, \dots, u_n) = \prod_{i=1}^n \varphi_{X_i}(u_i) ,$$

où $\varphi_{X_i}(u_i)$ est la fonction caractéristique de X_i .

Démonstration. Supposons l'indépendance. D'après le Théorème 3.1.3,

$$\varphi_X(u) = E[\prod_{j=1}^n e^{iu_j X_j}] = \prod_{j=1}^n E[e^{iu_j X_j}] = \prod_{j=1}^n \varphi_{X_j}(u_j) .$$

Inversement, supposons que la fonction caractéristique du vecteur se factorise comme dans l'énoncé du théorème, les φ_{X_j} étant les fonctions caractéristiques des X_j . Considérons maintenant n variables aléatoires $\tilde{X}_1, \dots, \tilde{X}_n$ indépendantes dans leur ensemble et telles que pour chaque j , \tilde{X}_j ait la même loi que X_j . Définissons

$\tilde{X} = (\tilde{X}_j, \dots, \tilde{X}_n)$. D'après la partie directe du théorème,

$$\varphi_{\tilde{X}}(u_1, \dots, u_n) = \prod_{j=1}^n \varphi_{\tilde{X}_j}(u_j) ,$$

et donc comme $\varphi_{\tilde{X}_j}(u_j) \equiv \varphi_{X_j}(u_j)$,

$$\varphi_{\tilde{X}}(u) = \varphi_X(u) .$$

Comme la fonction caractéristique caractérise la loi, X et \tilde{X} ont même loi. D'après Théorème 3.1.2, \tilde{X} admet la densité de probabilité

$$f_{\tilde{X}}(x_1, \dots, x_n) = \prod_{j=1}^n f_{X_j}(x_j) .$$

Le vecteur X admet donc lui aussi la densité de probabilité $\prod_{j=1}^n f_{X_j}(x_j)$, ce qui prouve l'indépendance des X_1, \dots, X_n , d'après le Théorème 3.1.2. \square

Le théorème précédent est également vrai pour des vecteurs, avec une preuve analogue. Par exemple, si Y et Z sont deux vecteurs de dimensions ℓ et p respectivement, une condition nécessaire et suffisante d'indépendance de Y et Z est

$$E[e^{i(v^T Y + w^T Z)}] = E[e^{i v^T Y}] E[e^{i w^T Z}] \quad (v \in \mathbb{R}^\ell, w \in \mathbb{R}^p) .$$

Somme de variables aléatoires indépendantes

Théorème 3.1.5 *Soit X et Y deux variables aléatoires indépendantes admettant les densités de probabilité f_X et f_Y respectivement. La somme $Z = X + Y$ admet pour densité de probabilité le produit de convolution de f_X et f_Y :*

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) dy = \int_{-\infty}^{+\infty} f_Y(z-x) f_X(x) dx .$$

Démonstration. La densité de probabilité du vecteur (X, Y) est $f_X(x) f_Y(y)$, d'après le Théorème 3.1.2. On a donc si on note $A = \{(x, y) ; x + y \leq z\}$

$$\begin{aligned} P(Z \leq z) &= P(X + Y \leq z) \\ &= P((X, Y) \in A) \\ &= E[1_A(X, Y)] \\ &= \int \int_A f_X(x) f_Y(y) dx dy \end{aligned}$$

$$\begin{aligned}
&= \int \int 1_A(x, y) f_X(x) f_Y(y) \, dx \, dy \\
&= \int \int 1_{\{x+y \leq z\}} f_X(x) f_Y(y) \, dx \, dy \\
&= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{z-x} f_Y(y) \, dy \right) f_X(x) \, dx \\
&= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^z f_Y(y-x) \, dy \right) f_X(x) \, dx \\
&= \int_{-\infty}^z \left(\int_{-\infty}^{+\infty} f_Y(y-x) f_X(x) \, dx \right) dy .
\end{aligned}$$

On a donc bien

$$P(Z \leq z) = \int_{-\infty}^z \left(\int_{-\infty}^z f_Y(y-x) f_X(x) \, dx \right) dy .$$

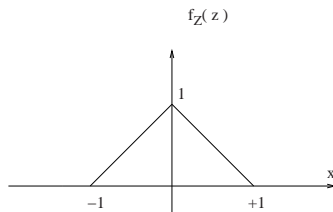
□

EXEMPLE 3.1.5: SOMME DE 2 VARIABLES DE MÊME DISTRIBUTION UNIFORME. Soit X et Y deux variables aléatoires indépendantes uniformes sur $[-\frac{1}{2}, +\frac{1}{2}]$. La densité de probabilité de $Z = X + Y$ est

$$f_Z(z) = \int_{-\frac{1}{2}}^{+\frac{1}{2}} 1_{[-\frac{1}{2}, +\frac{1}{2}]}(z-x) \, dx .$$

On obtient

$$f_Z(z) = \begin{cases} 0 & \text{si } z \leq -1 \\ z+1 & \text{si } -1 \leq z \leq 0 \\ -z+1 & \text{si } 0 \leq z \leq 1 \\ 0 & \text{si } z \geq +1 . \end{cases}$$



Variance d'une somme de variables indépendantes

Théorème 3.1.6 Soit X et Y deux variables aléatoires indépendantes admettant les densités f_X et f_Y et de variances σ_X^2 et σ_Y^2 . La somme $Z = X + Y$ admet pour variance la somme des variances de X et Y :

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 .$$

Démonstration. Il s'agit de calculer $\sigma_Z^2 = E[\{(X + Y) - (m_X + m_Y)\}^2]$. D'après la linéarité de l'espérance,

$$\begin{aligned} \sigma_Z^2 &= E[(X - m_X)^2] + 2E[(X - m_X)(Y - m_Y)] + E[(Y - m_Y)^2] \\ &= \sigma_X^2 + 2E[(X - m_X)(Y - m_Y)] + \sigma_Y^2 , \end{aligned}$$

et d'après le Théorème 3.1.3.

$$E[(X - m_X)(Y - m_Y)] = E[X - m_X] E[Y - m_Y] = 0 .$$

□

Somme aléatoire de variables indépendantes

Théorème 3.1.7 Soit T, X_1, X_2, \dots une famille de variables aléatoires indépendantes, où T est à valeurs dans \mathbb{N} et les X_i , $i \geq 1$, sont identiquement distribuées de fonction caractéristique commune φ_X . Alors la fonction caractéristique φ_S de la somme aléatoire $S = X_1 + \dots + X_T$ ($S = 0$ si $T = 0$) est donnée par la formule

$$\varphi_S(u) = g_T(\varphi_X(u)) , \tag{3.1}$$

où g_T est la fonction génératrice de T .

Démonstration. On a $\varphi_S(u) = E[e^{iuS}] = E[e^{iu(X_1 + \dots + X_T)}]$. Or :

$$e^{iu(X_1 + \dots + X_T)} = \sum_{n=0}^{\infty} (1_{\{T=n\}} e^{iu(X_1 + \dots + X_T)}) = \sum_{n=0}^{\infty} (1_{\{T=n\}} e^{iu(X_1 + \dots + X_n)})$$

On a donc

$$E[e^{iu(X_1 + \dots + X_T)}] = \sum_{n=0}^{\infty} E[1_{\{T=n\}} e^{iu(X_1 + \dots + X_n)}] .$$

Comme T et (X_1, \dots, X_n) sont indépendantes, cette dernière quantité est égale à

$$\sum_{n=1}^{\infty} E[1_{\{T=n\}}] E[e^{iu(X_1 + \dots + X_n)}] .$$

Mais $E[1_{\{T=n\}}] = P(T = n)$ (d'après (2.6)) et d'autre part, comme X_1, \dots, X_n sont indépendantes,

$$\begin{aligned} E[e^{iu(X_1 + \dots + X_n)}] &= E\left[\prod_{j=1}^n e^{iuX_j}\right] \\ &= \prod_{j=1}^n E[e^{iuX_j}] = \varphi_X(u)^n. \end{aligned}$$

On a donc

$$\varphi_S(u) = \sum_{n=1}^{\infty} P(T = n) \varphi_X(u)^n = g_T(\varphi_X(u)).$$

□

Formule du changement de variables

Comment calculer la densité de probabilité d'un vecteur aléatoire Y qui s'exprime comme fonction d'un autre vecteur aléatoire ? Parfois, les calculs à effectuer sont directs et simples, comme le montre l'exemple suivant.

EXEMPLE 3.1.6: DISTRIBUTION DE RAYLEIGH. $Y = X^2$ où X suit une loi gaussienne standard ($\mathcal{N}(0, 1)$). Si $y \leq 0$, $P(Y \leq y) = 0$, tandis que si $y \geq 0$,

$$P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}),$$

ou encore (toujours pour $y \geq 0$) :

$$F_Y(y) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{y}}^{+\sqrt{y}} e^{-\frac{1}{2}x^2} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-\frac{1}{2}x^2} dx.$$

Pour obtenir la densité de probabilité f_Y il suffit de dériver la fonction de répartition F_Y . On obtient dans le cas présent la densité de la loi $\gamma(\frac{1}{2}, \frac{1}{2})$, c'est-à-dire la loi du χ^2 à un degré de liberté, ou loi de Rayleigh :

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}} 1_{\{y \geq 0\}}.$$

Il se présente des situations où les calculs sont moins immédiats. La formule de changement de variables, que nous allons donner, rend parfois de grands services.

Soit $X = (X_1, \dots, X_n)$ un n -vecteur aléatoire admettant une densité de probabilité f_X et soit $Y = (Y_1, \dots, Y_n)$ un n -vecteur aléatoire de la forme ;

$$Y = g(X) \quad \text{c'est-à-dire} \quad (Y_1, \dots, Y_n) = (g(X_1), \dots, g_n(X_n)),$$

où g est une fonction suffisamment régulière. Plus précisément, la fonction g satisfait aux hypothèses suivantes :

(i) g est définie sur un ouvert U de \mathbb{R}^n tel que

$$P(X \in U) = \int_U f_X(x) dx = 1 ,$$

et g est bijective (son inverse étant notée g^{-1})

(ii) les dérivées partielles premières de g existent et elles sont continues en tout point de U .

(iii) la fonction “déterminant jacobien” de la transformation $y = g(x)$, qui est la fonction Dg définie par

$$Dg = \det \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \dots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial x_1} & \dots & \frac{\partial g_n}{\partial x_n} \end{pmatrix} ,$$

ne s'annule pas sur U .

(Un résultat d'Analyse nous dit que $V = g(U)$ est alors un ouvert, et que l'application inverse $g^{-1} : V \rightarrow U$ vérifie les mêmes propriétés que g).

De plus, on a la formule de changement de variables de l'Analyse, à savoir que si $u : U \rightarrow \mathbb{R}$ est non négative,

$$\int_U u(x) dx = \int_{g(U)} u(g^{-1}(y)) |Dg^{-1}(y)| dy$$

Théorème 3.1.8 *Sous les conditions énoncées plus haut, Y admet la densité de probabilité*

$$f_Y(y) = f_X(g^{-1}(y)) |Dg^{-1}(y)| 1_U(g^{-1}(y)) .$$

Démonstration. Calculons l'espérance de $h(Y)$ où $h : U \rightarrow \mathbb{R}$ est non négative. On a

$$E[h(Y)] = E[h(g(X))] = \int_U h(g(x)) f_X(x) dx ,$$

et donc en appliquant la formule de changement de variables de l'Analyse avec $u(x) = h(g(x))$,

$$E[h(Y)] = \int_{g(U)} h(y) f_X(g^{-1}(y)) |Dg^{-1}(y)| dy .$$

En particulier, avec $h(y) = 1_{\{y \leq a\}} = 1_{\{y_1 \leq a_1, \dots, y_n \leq a_n\}}$,

$$P(Y \leq a) = \int_{-\infty}^a f_X(g^{-1}(y)) |Dg^{-1}(y)| 1_U(g^{-1}(y)) dy .$$

(On rappelle la notation $\int_{-\infty}^a = \int_{-\infty}^{a_1} \dots \int_{-\infty}^{a_n}$.)

□

EXEMPLE 3.1.7: Le couple $X = (X_1, X_2)$ admet la densité uniforme sur le carré $[0, 1]^2$:

$$f_X(x_1, x_2) = \begin{cases} 1 & \text{si } (x_1, x_2) \in [0, 1] \times [0, 1] \\ 0 & \text{autrement.} \end{cases}$$

On veut calculer la densité du vecteur $Y = (Y_1, Y_2)$ donné par

$$Y_1 = \frac{X_1}{X_2}, Y_2 = X_1 + X_2$$

lorsqu'on connaît la densité de probabilité de $X = (X_1, X_2)$. La transformation inverse g^{-1} faisant passer de Y à X est

$$x_1 = \frac{y_1 y_2}{y_1 + 1}, \quad x_2 = \frac{y_2}{y_1 + 1},$$

d'où

$$Dg^{-1}(y) = \det \begin{pmatrix} \frac{y_2}{(y_1+1)^2} & \frac{y_1}{y_1+1} \\ \frac{-y_2}{(y_1+1)^2} & \frac{1}{y_1+1} \end{pmatrix} = \frac{y_2}{(y_1+1)^2}.$$

On applique le Théorème 3.1.8 avec $U = (0, 1) \times (0, 1)$ pour obtenir

$$f_Y(y) = \begin{cases} \frac{y_2}{(y_1+1)^2} & \text{si } (y_1, y_2) \in V = g(U) \\ 0 & \text{si } (y_1, y_2) \notin V = g(U). \end{cases}$$

où

$$V = g((0, 1) \times (0, 1)) = \{y_2 \leq 1 + \inf\{y_1, y_1^{-1}\}\}.$$

EXEMPLE 3.1.8: CHANGEMENT DE VARIABLES LINÉAIRE. Soit X un n -vecteur aléatoire réel et $Y = AX + b$ où A est une matrice $n \times n$ réelle *invertible* et b est un n -vecteur réel. La densité de Y s'obtient immédiatement par application du Théorème 3.1.8 :

$$f_Y(y) = \frac{1}{|\det A|} f_X(A^{-1}(y - b)).$$

Il y a des cas d'intérêt pratique où la fonction g n'est pas bijective mais pour lesquels l'ouvert U se décompose en plusieurs (disons 2 pour l'exemple) ouverts disjoints :

$$U = U_1 + U_2,$$

tels que les restrictions g_1 et g_2 de g à U_1 et U_2 respectivement satisfont aux hypothèses admises pour g dans le Théorème 3.1.8. Dans ce cas, la même méthode s'applique, à ceci près qu'il faut maintenant dissocier l'intégrale

$$\int_U h(g(x))f_X(x) dx = \int_{U_1} h(g_1(x))f_X(x) dx + \int_{U_2} h(g_2(x))f_X(x) dx$$

et appliquer la formule de changement de variables de l'Analyse à chacune des parties séparément, ce qui donne

$$E[h(Y)] = \int_{g_1(U_1)} h(y)f_X(g_1^{-1}(y))|Dg_1^{-1}(y)| dy \\ + \int_{g_2(U_2)} h(y)f_X(g_2^{-1}(y))|Dg_2^{-1}(y)| dy,$$

et donc

$$f_Y(y) = f_X(g_1^{-1}(y))|Dg_1^{-1}(y)| 1_{U_1}(g_1^{-1}(y)) \\ + f_X(g_2^{-1}(y))|Dg_2^{-1}(y)| 1_{U_2}(g_2^{-1}(y)).$$

EXEMPLE 3.1.9: Le couple $X = (X_1, X_2)$ est un couple de variables aléatoires indépendantes, X_1 suivant une loi $\mathcal{E}(1)$:

$$f_{X_1}(x_1) = e^{-x_1} 1_{\{x_1 > 0\}}$$

et X_2 une loi $\gamma(2, 1)$:

$$f_{X_2}(x_2) = x_2 e^{-x_2} 1_{\{x_2 > 0\}}$$

On a donc :

$$f_X(x_1, x_2) = x_1 e^{-(x_1+x_2)} 1_{\{x_1 x_2 > 0\}}.$$

On effectue le changement de variables $(x_1, x_2) \rightarrow (y_1, y_2)$ défini par

$$y_1 = \inf(x_1, x_2), \quad y_2 = \sup(x_1, x_2).$$

On se trouve donc dans la situation où la transformation g n'est pas bijective sur l'intérieur du support de f_X . Définissons

$$U_1 = U \cap \{x_2 > x_1\}, \quad U_2 = U \cap \{x_1 < x_2\},$$

où $U = \{x_1 > 0, x_2 > 0\} - \{x_1 = x_2\}$ est un ouvert tel que $P(X \in U) = 1$. La restriction g_1 de g à U_1 est l'identité, tandis que la restriction de g à U_2 est la transposition $(x_1, x_2) \rightarrow (x_2, x_1)$. Pour g_2 aussi bien que pour g_1 , le déterminant du jacobien est 1. Appliquons la formule du changement de variables modifiée en remarquant que $V_1 = V_2 = \{y_1 < y_2\}$.

On obtient :

$$f_Y(y_1, y_2) = (y_1 + y_2) e^{-(y_1+y_2)} 1_{\{0 < y_1 < y_2\}}.$$

Ordonnancement d'un vecteur aléatoire

Soit X_1, \dots, X_n des variables aléatoires réelles IID de densité de probabilité commune f . Sous ces hypothèses, la probabilité pour que X_1, \dots, X_n prennent des valeurs différentes est égale à 1. Notons que ceci est vrai dès que le vecteur $X = (X_1, \dots, X_n)$ admet une densité de probabilité. En effet, et par exemple,

$$P(X_1 = X_2) = \int \int_{\{(x_1, x_2); x_1 = x_2\}} f_{(X_1, X_2)}(x_1, x_2) dx_1 dx_2 = 0$$

puisque l'ensemble $A = \{(x_1, x_2); x_1 = x_2\}$ a une aire nulle.

On peut donc, presque sûrement, définir sans ambiguïté les variables aléatoires Z_1, \dots, Z_n obtenues en réordonnant les X_1, \dots, X_n en ordre croissant :

$$\begin{cases} Z_i \in \{X_1, \dots, X_n\}, \\ Z_1 < Z_2 < \dots < Z_n. \end{cases}$$

En particulier, $Z_1 = \min(X_1, \dots, X_n)$ et $Z_n = \max(X_1, \dots, X_n)$.

Théorème 3.1.9 *La densité de probabilité du vecteur réordonné $Z = (Z_1, \dots, Z_n)$ est*

$$f_Z(z_1, \dots, z_n) = n! \left\{ \prod_{j=1}^n f(z_j) \right\} 1_C(z_1, \dots, z_n),$$

où

$$C = \{(z_1, \dots, z_n) \in \mathbb{R}^n; z_1 < z_2 < \dots < z_n\}.$$

Démonstration. Soit σ la permutation de $\{1, \dots, n\}$ qui ordonne X_1, \dots, X_n en ordre croissant, c'est-à-dire,

$$X_{\sigma(i)} = Z_i$$

(σ est une permutation aléatoire de $\{1, 2, \dots, n\}$). Pour tout $A \subseteq \mathbb{R}^n$,

$$P(Z \in A) = P(Z \in A \cap C) = P(X_\sigma \in A \cap C) = \sum_{\sigma_o} P(X_{\sigma_o} \in A \cap C, \sigma = \sigma_o),$$

où la somme porte sur toutes les permutations de $\{1, \dots, n\}$. Observant que $X_{\sigma_o} \in A \cap C$ implique $\sigma = \sigma_o$,

$$P(X_{\sigma_o} \in A \cap C, \sigma = \sigma_o) = P(X_{\sigma_o} \in A \cap C)$$

et donc, puisque la distribution de probabilité de X_{σ_o} ne dépend pas de la permutation fixée (non aléatoire) σ_o (les X_i sont IID),

$$P(X_{\sigma_o} \in A \cap C) = P(X \in A \cap C).$$

Donc,

$$\begin{aligned} P(Z \in A) &= \sum_{\sigma_o} P(X \in A \cap C) = n! P(X \in A \cap C) \\ &= n! \int_{A \cap C} f_X(x) dx = \int_A n! f_X(x) 1_C(x) dx. \end{aligned}$$

□

EXEMPLE 3.1.10: LE VOLUME DE LA PYRAMIDE. En guise d'exemple, nous allons appliquer le résultat obtenu ci-dessus pour obtenir la formule

$$\int_a^b \cdots \int_a^b 1_C(z_1, \dots, z_n) dz_1 \cdots dz_n = \frac{(b-a)^n}{n!}. \quad (3.2)$$

En effet, lorsque les X_i sont uniformément distribuées sur $[a, b]$.

$$f_Z(z_1, \dots, z_n) = \frac{n!}{(b-a)^n} 1_{[a,b]^n}(z_1, \dots, z_n) 1_C(z_1, \dots, z_n). \quad (3.3)$$

Le résultat suit parce que $\int_{\mathbb{R}^n} f_Z(z) dz = 1$.

3.2 Échantillonnage de distributions de probabilité

Lorsqu'on cherche à simuler un système stochastique, le problème se pose de générer des variables ou vecteurs aléatoires ayant une distribution de probabilité donnée. On dit alors qu'on veut *échantillonner* cette distribution. Pour ce faire, on dispose d'un *générateur aléatoire* qui fournit à la demande une variable aléatoire uniformément distribuée sur $[0, 1]$, ou une suite IID de telles variables. La question de savoir comment fabriquer un tel générateur à l'aide d'un ordinateur, qui est une machine déterministe, est hors de notre propos, et sera supposée résolue de manière satisfaisante.

Nous allons passer en revue les deux méthodes classiques permettant de construire une variable Z de fonction de répartition donnée

$$F(z) = P(Z \leq z).$$

La méthode de l'inverse, cas des variables discrètes

Dans le cas où Z est une variable discrète de distribution $P(Z = a_i) = p_i (0 \leq i \leq K)$, le principe de base de l'algorithme d'échantillonnage est le suivant

(1) Générer $U \sim \mathcal{U}([0, 1])$

(2) Définir $Z = a_\ell$ si $p_0 + p_1 + \dots + p_{\ell-1} \leq U \leq p_0 + p_1 + \dots + p_\ell$.

C'est la *méthode de l'inverse*. (Cette terminologie sera justifiée bientôt.)

Une version "naïve" de l'algorithme mettant en oeuvre cette méthode est le suivant : on fait successivement les tests $U \leq p_0?$, $U \leq p_0 + p_1?$, \dots , jusqu'à ce que le test soit positif. Le nombre moyen d'itérations nécessaires est $\sum_{i \geq 0} (i+1)p_i = 1 + E[Z]$. Il peut être diminué, et l'exemple suivant donne une façon de le faire.

EXEMPLE 3.2.1: ÉCHANTILLONNAGE DE LA DISTRIBUTION DE POISSON. On veut échantillonner la distribution de Poisson de moyenne $\theta > 0$. L'algorithme naïf demande $1 + \theta$ itérations en moyenne. Pour l'améliorer, on réordonne les états, et on part du plus probable, i_0 , à partir duquel on explore successivement $i_0 - 1$, $i_0 + 1$, $i_0 - 2$, $i_0 + 2$, etc. Lorsque θ est grand, i_0 est proche de θ , et le nombre moyen d'itérations est, en gros, $E[|Z - \theta|]$. Pour simplifier l'analyse, on supposera que θ est un entier. La variable $Z - \theta$ a la même distribution que

$$\sum_{j=1}^{\theta} (Y_j - 1),$$

où les Y_j sont des variables de Poisson indépendantes, de moyenne commune égale à 1. Lorsque θ est large, le théorème de la limite gaussienne (que nous verrons plus tard) dit que

$$\frac{Z - \theta}{\sqrt{\theta}} = \frac{\sum_{j=1}^{\theta} (Y_j - 1)}{\sqrt{\theta}}$$

est approximativement distribuée selon une loi gaussienne de moyenne 0 et de variance 1. Donc

$$\begin{aligned} E[|Z - \theta|] &\simeq \sqrt{\theta} E[|\mathcal{N}(0, 1)|] \\ &\simeq 0.82\sqrt{\theta}, \end{aligned}$$

que l'on comparera avec θ .

La méthode de l'inverse, cas des variables absolument continues

Dans le cas de variables de fonctions de répartition continues, la méthode de l'inverse prend la forme suivante (qui justifie la terminologie).

Théorème 3.2.1 *La variable aléatoire*

$$Z = F^{-1}(U),$$

où F^{-1} est l'inverse de la fonction de répartition continue F et $U \sim \mathcal{U}([0, 1])$, admet F pour fonction de répartition.

Démonstration. On a bien

$$\begin{aligned} P(Z \leq z) &= P(F^{-1}(U) \leq z) \\ &= P(U \leq F(z)) \\ &= F(z). \end{aligned}$$

□

EXEMPLE 3.2.2: ÉCHANTILLONNAGE DE LA DISTRIBUTION EXPONENTIELLE. On veut échantillonner la distribution exponentielle de moyenne λ^{-1} , dont la fonction de répartition est :

$$F(z) = 1 - e^{-\lambda z} \quad (z \geq 0).$$

La solution de $y = 1 - e^{-\lambda z}$ est $z = -\frac{1}{\lambda} \ln(1 - y) = F^{-1}(y)$, et donc on posera $Z = -\frac{1}{\lambda} \ln(1 - U)$. Ou encore, puisque U et $1 - U$ ont la même distribution,

$$Z = -\frac{1}{\lambda} \ln(U).$$

La méthode de l'inverse nécessite le calcul de l'inverse de la fonction de répartition, ce qui n'est pas toujours facile. D'autre part, elle ne s'applique pas aux vecteurs aléatoires. Ces défauts sont éliminés dans la méthode suivante.

La méthode d'acceptation-réjection

On cherche à échantillonner la densité de probabilité sur f sur \mathbb{R}^d . On suppose qu'on peut générer une suite IID $\{Y_n\}_{n \geq 1}$ de vecteurs aléatoires de \mathbb{R}^d admettant une densité de probabilité g qui vérifie, pour tout $x \in \mathbb{R}$,

$$\frac{f(x)}{g(x)} \leq c$$

pour une constante c (nécessairement supérieure ou égale à 1). On dispose également d'une suite IID $\{U_n\}_{n \geq 1}$ de variables uniformément distribuées sur $[0, 1]$.

Théorème 3.2.2 Soit τ le premier indice $n \geq 1$ tel que

$$U_n \leq \frac{f(Y_n)}{cg(Y_n)}$$

et posons

$$Z = Y_\tau.$$

Alors :

(a) Z admet la densité de probabilité f , et :

(b) $E[\tau] = c$.

Démonstration. On a

$$P(Z \leq x) = P(Y_\tau \leq x) = \sum_{n \geq 1} P(\tau = n, Y_n \leq x).$$

Soit A_k l'événement $\{U_k > \frac{f(Y_k)}{cg(Y_k)}\}$. On a :

$$\begin{aligned} P(\tau = n, Y_n \leq x) &= P(A_1, \dots, A_{n-1}, \overline{A_n}, Y_n \leq x) \\ &= P(A_1) \cdots P(A_{n-1}) P(\overline{A_n}, Y_n \leq x). \end{aligned}$$

$$\begin{aligned} P(\overline{A_k}) &= \int_{\mathbb{R}^d} P\left(U_k \leq \frac{f(y)}{cg(y)}\right) g(y) dy \\ &= \int_{\mathbb{R}^d} \frac{f(y)}{cg(y)} g(y) dy \\ &= \int_{\mathbb{R}^d} \frac{f(y)}{c} dy = \frac{1}{c}. \end{aligned}$$

$$\begin{aligned} P(\overline{A_k}, Y_k \leq x) &= \int_{\mathbb{R}^d} P\left(U_k \leq \frac{f(y)}{cg(y)}\right) 1_{\{y \leq x\}} g(y) dy \\ &= \int_{-\infty}^x \frac{f(y)}{cg(y)} g(y) dy \\ &= \int_{-\infty}^x \frac{f(y)}{c} dy \\ &= \frac{1}{c} \int_{-\infty}^x f(y) dy. \end{aligned}$$

Donc

$$P(Z \leq x) = \sum_{n \geq 1} \left(1 - \frac{1}{c}\right)^{n-1} \frac{1}{c} \int_{-\infty}^x f(y) dy = \int_{-\infty}^x f(y) dy.$$

Aussi, en utilisant les calculs ci-dessus,

$$\begin{aligned} P(\tau = n) &= P(A_1, \dots, A_{n-1}, \overline{A_n}) \\ &= P(A_1) \cdots P(A_{n-1})P(\overline{A_n}) \\ &= \left(1 - \frac{1}{c}\right)^{n-1} \frac{1}{c}, \end{aligned}$$

d'où il suit que $E[\tau] = c$. □

Les deux méthodes ci-dessus n'épuisent pas toutes les ressources disponibles. De multiples astuces sont utilisées pour rendre l'échantillonnage simple et efficace. En voici un exemple, qu'on peut facilement généraliser :

EXEMPLE 3.2.3: On cherche à échantillonner la densité de probabilité $f(x) = \frac{1}{2} \times 2e^{-2x} + \frac{1}{2} \times 3e^{-3x}$. Voici une proposition : On utilise U_1 et U_2 deux variables uniformément distribuées sur $[0, 1]$. On définit alors

$$Z = 1_{U_1 \leq \frac{1}{2}} \times \left(-\frac{1}{2} \log U_2\right) + 1_{U_1 < \frac{1}{2}} \times \left(-\frac{1}{3} \log U_2\right)$$

On a :

$$\begin{aligned} P(Z \leq x) &= P(U_1 \leq \frac{1}{2}, -\frac{1}{2} \log U_2 \leq x) + P(U_1 < \frac{1}{2}, -\frac{1}{3} \log U_2 \leq x) \\ &= P(U_1 \leq \frac{1}{2})P(-\frac{1}{2} \log U_2 \leq x) + P(U_1 < \frac{1}{2})P(-\frac{1}{3} \log U_2 \leq x) \\ &= \frac{1}{2}(1 - e^{-2x}) + \frac{1}{2}(1 - e^{-3x}). \end{aligned}$$

En dérivant cette fonction de répartition, on obtient bien la densité de probabilité f . _____

3.3 Corrélation et régression linéaire

Variables aléatoires de carré intégrable

Si X et Y sont deux variables aléatoires réelles de carré intégrable et si a et b sont des nombres réels, alors $aX + bY$ est aussi de carré intégrable. En effet

$$|aX + bY|^2 \leq |a|^2|X|^2 + 2|ab||XY| + |b|^2|Y|^2,$$

et donc (propriétés de monotonie et de linéarité de l'espérance²)

$$E[|aX + bY|^2] \leq |a|^2E[|X|^2] + 2|ab|E[|XY|] + |b|^2E[|Y|^2].$$

²Ces propriétés ont été démontrées dans le cas des variables discrètes ou admettant une densité de probabilité. Elles sont vraies dans le cas général, comme on le verra dans le Chapitre 6. Les résultats de cette section sont donc généraux.

On aura prouvé que la quantité $E[|aX + bY|^2]$ est finie si on montre que $E[|XY|] < \infty$. Or ceci découle de l'inégalité $2|XY| \leq |X|^2 + |Y|^2$ et des propriétés de linéarité et de monotonie de l'espérance.

Nous aurons à nouveau besoin de la notion de “presque-sûrement”, cette fois-ci dans un cadre plus général que celui des variables aléatoires discrètes.

Définition 3.3.1 On dit que la propriété \mathcal{P} relative à un vecteur aléatoire X est vérifiée P -presque sûrement si

$$P(\{\omega; X(\omega) \text{ vérifie } \mathcal{P}\}) = 1,$$

et on note ceci

$$X \text{ vérifie } \mathcal{P} \quad P - p.s..$$

Par exemple, $g_1(X) < g_2(X)$ P -p.s. veut dire que $P(g_1(X) < g_2(X)) = 1$.

Théorème 3.3.1 Si X est une variable aléatoire réelle non négative telle que $E[X] = 0$, alors $X = 0$, p.s.

Démonstration. Ceci découle, par exemple, de l'inégalité de Markov

$$P(|X| \geq \varepsilon) \leq \frac{E[X]}{\varepsilon} \quad (\varepsilon > 0),$$

qui donne $P(|X| \geq \varepsilon) = 0$, pour tout $\varepsilon > 0$. Ceci implique que $X = 0$, presque sûrement. En effet le complémentaire de l'ensemble $\{X = 0\}$ est l'union des ensembles $\{|X| \geq \frac{1}{n}\}$ et donc :

$$\begin{aligned} 1 - P(X = 0) &= P\left(\bigcup_{n=1}^{\infty} \left\{|X| \geq \frac{1}{n}\right\}\right) \\ &\leq \sum_{n=1}^{\infty} P\left(|X| \geq \frac{1}{n}\right) = 0. \end{aligned}$$

□

Inégalité de Schwarz

Théorème 3.3.2 Soit X et Y deux variables aléatoires complexes de carré du module intégrable. On a alors l'inégalité de Schwarz

$$E[|XY|] \leq E[|X|^2]^{\frac{1}{2}} E[|Y|^2]^{\frac{1}{2}}. \quad (3.4)$$

(Ceci prouve en particulier que le produit XY est intégrable.)

Démonstration. Il suffit de faire la démonstration quand X et Y sont non négatives. Pour tout λ réel ,

$$(X + \lambda Y)^2 \geq 0,$$

et donc (monotonie et linéarité)

$$E[X]^2 + 2\lambda E[XY] + \lambda^2 E[Y^2] \geq 0 . \quad (3.5)$$

Cette dernière inégalité étant vraie pour tout λ réel, le discriminant du binôme en λ doit être négatif, ce qui s'écrit

$$E[XY]^2 - E[X^2]E[Y^2] \leq 0 .$$

□

Coefficient de corrélation

Définition 3.3.2 Soit X et Y deux variables aléatoires de carré intégrable de variances strictement positives. Le coefficient de corrélation entre X et Y est la quantité

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} . \quad (3.6)$$

(On notera le plus souvent par ρ ce coefficient de corrélation.)

D'après l'inégalité de Schwarz,

$$-1 \leq \rho \leq +1 . \quad (3.7)$$

Si $\rho > 0$, on dit que X et Y sont *positivement corrélées*, et si $\rho < 0$, on dit que X et Y sont *négativement corrélées*.

Si $\rho = 0$, c'est-à-dire si $\text{cov}(X, Y) = 0$, on dit que X et Y sont *non corrélées*. Un tel cas peut se produire lorsque X et Y sont des variables aléatoires indépendantes car alors

$$E[(X - m_X)(Y - m_Y)] = E[X - m_X]E[Y - m_Y] = 0 ,$$

mais cela n'est pas le seul cas. La non corrélation n'entraîne en général pas l'indépendance.

Théorème 3.3.3 Si $|\rho| = 1$, alors X et Y sont liés par une relation affine :

$$aX + bY + c = 0 \quad (P \text{ p.s.}) .$$

Démonstration. En effet $|\rho| = 1$ équivaut à l'égalité dans l'inégalité de Schwarz écrite pour $X - m_X$ et $Y - m_Y$. Il existe donc a et b tels que $a(X - m_X) + b(Y - m_Y) = 0$. □

Matrices de covariance

On considère maintenant des *vecteurs* aléatoires

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix},$$

dont toutes les coordonnées sont de carré intégrable. On dira que de tels vecteurs aléatoires sont de carré intégrable. La *matrice d'intercorrélation* de X et Y est l'espérance de la matrice

$$(X - m)(Y - m)^T = \begin{pmatrix} X_1 - m_{X_1} \\ \vdots \\ X_n - m_{X_n} \end{pmatrix} (Y_1 - m_{Y_1}, \dots, Y_m - m_{Y_m}),$$

c'est-à-dire la matrice $n \times m$

$$\Sigma_{XY} = \begin{pmatrix} \text{cov}(X_1, Y_1) & \text{cov}(X_1, Y_2) & \dots & \text{cov}(X_1, Y_m) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \text{cov}(X_n, Y_1) & \text{cov}(X_n, Y_2) & \dots & \text{cov}(X_n, Y_m) \end{pmatrix},$$

On a évidemment

$$\Sigma_{XY} = \Sigma_{XY}^T. \quad (3.8)$$

En particulier, la matrice Σ_{XX} est une matrice carrée de $n \times n$ symétrique. On la dénotera par Σ_X au lieu de Σ_{XX} . On l'appelle la *matrice de covariance* de X .

Théorème 3.3.4 *La matrice de covariance Σ_X est symétrique. Elle est de plus non négative : Pour tout vecteur réel $u = (u_1, \dots, u_n)^T$,*

$$u^T \Sigma_X u \geq 0.$$

Démonstration. En effet,

$$\begin{aligned} u^T \Sigma_X u &= \sum_{i=1}^n \sum_{j=1}^n u_i u_j \sigma_{X_i X_j} \\ &= \sum_{i=1}^n \sum_{j=1}^n u_i u_j E[(X_i - m_{X_i})(X_j - m_{X_j})] \\ &= E \left[\left(\sum_{i=1}^n u_i (X_i - m_{X_i}) \right) \left(\sum_{j=1}^n u_j (X_j - m_{X_j}) \right) \right] \\ &= E[Z^2] \geq 0, \end{aligned}$$

où $Z = \sum_{i=1}^n u_i (X_i - m_{X_i})$. □

Transformations linéaires

Soit X un n -vecteur aléatoire de carré intégrable sur lequel on effectue une transformation affine

$$Z = CX + b ,$$

où C est une matrice $m \times n$ et b un m -vecteur. Le vecteur moyenne de Y est

$$m_Z = Cm_X + b ,$$

d'où

$$Z - m_Z = C(X - m_X) .$$

La matrice de covariance de Y est donc

$$E[(Z - m_Z)(Z - m_Z)^T] = E[C(X - m_X)(X - m_X)^T C^T] ,$$

c'est-à-dire après explicitation des calculs,

$$\Sigma_Z = C\Sigma_X C^T .$$

De même, si X et Y sont respectivement un n -vecteur et un m -vecteur aléatoires de carré intégrable et de matrice d'intercorrélation Σ_{XY} , alors les vecteurs

$$U = CX + b, \quad V = D.Y + e ,$$

où C , b , D et e sont des matrices $\ell \times n$, $\ell \times 1$, $k \times m$ et $k \times 1$ respectivement, admettent pour matrice d'intercorrélation

$$\Sigma_{UV} = C\Sigma_{XY}D^T .$$

Revenons au vecteur Z défini quelques lignes plus haut, et choisissons pour C la matrice orthonormale (avec en particulier la propriété $C^{-1} = C^T$) qui diagonalise la matrice de covariance Σ_X :

$$C\Sigma_X C^T = \text{diag}(\sigma_1^2, \sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0) ,$$

où r est le rang de Σ_X et où les σ_i^2 sont les valeurs propres de Σ_X répétées autant de fois que leur ordre de multiplicité. Le vecteur $Z = C(X - m_X)$ résultant a donc ses $n - r$ dernières coordonnées presque-sûrement nulles, car $E[(Z_i - m_{Z_i})^2] = 0$ entraîne que $P(Z_i = m_{Z_i}) = 1$. De plus les coordonnées Z_i ne sont pas corrélées entre elles. On en extraira de l'analyse ci-dessus le :

Théorème 3.3.5 *Si la matrice de covariance Σ_X est de rang r , alors, presque sûrement, le vecteur X se trouve dans un hyperplan de \mathbb{R}^n de dimension r .*

Un problème d'optimisation

La fin de cette section est consacrée à la recherche de la meilleure approximation d'une variable aléatoire Y par une combinaison affine de n variables aléatoires données X_1, X_2, \dots, X_n . La combinaison affine

$$\hat{Y} = \hat{a}_0 + \sum_{i=1}^n \hat{a}_i X_i,$$

où les $\hat{a}_i \in \mathbb{R}$ sera jugée la meilleure approximation de Y si l'écart *quadratique* entre Y et \hat{Y} est inférieur à l'écart quadratique entre Y et toute autre approximation qui est une combinaison affine des X_i :

$$E[|Y - \hat{Y}|^2] \leq E \left[\left| Y - \left(a_0 + \sum_{i=1}^n a_i X_i \right) \right|^2 \right] \quad (a_i \in \mathbb{R} \quad (0 \leq i \leq n)).$$

Un tel problème, qui se pose lorsque la seule information concernant une variable aléatoire Y provient de l'observation de variables aléatoires X_i statistiquement liées à Y (voir l'exemple de théorie du signal traité dans l'Exercice 3.5.21), admet une solution simple qui fait l'objet de la présente section. Dans le Chapitre 4, on donnera une solution théorique d'un problème plus général : la recherche de la meilleure approximation quadratique de Y par une variable aléatoire fonction des X_i . Ce problème diffère du précédent en ce que l'approximation n'est pas nécessairement une fonction affine de l'observation.

Le problème abordé dans cette section celui de la *régression linéaire* de Y par X_i ($1 \leq i \leq n$). Il est a priori moins intéressant que le problème général de l'approximation (non nécessairement affine) qui vient d'être décrit, mais sa solution a un avantage considérable : elle ne fait intervenir la loi de probabilité de $Z = (Y, X_1, \dots, X_n)$ que par sa moyenne m_Z et sa matrice de covariance Σ_Z , tandis que la résolution du problème général nécessite la connaissance de toute la loi.

Le cas unidimensionnel

On cherche la meilleure approximation au sens des moindres carrés d'une variable aléatoire Y comme fonction affine d'une autre variable aléatoire X . En d'autres termes on cherche deux nombres réels a et b qui minimisent l'écart quadratique entre Y et son approximation $aX + b$:

$$E[|Y - (aX + b)|^2].$$

On peut tout de suite restreindre notre recherche à des a et b tels que $b = am_X + m_Y$. En effet :

Lemme 3.3.4 *Soit Z une variable aléatoire de carré intégrable. Pour tout nombre réel c ,*

$$E[(Z - m_Z)^2] \leq E[(Z - c)^2].$$

(C'est autour de sa moyenne qu'une variable aléatoire est la moins dispersée.)

Démonstration. L'inégalité à démontrer se réduit à $-m_Z^2 \leq -2cm_Z + c^2$ c'est-à-dire $(c - m_Z)^2 \geq 0$. \square

On est donc ramené au problème de trouver un nombre a qui minimise :

$$f(a) = E[(Y - m_Y)^2 + a(X - m_X)^2] .$$

Comme

$$f(a) = \text{Var}(X) - 2a \text{cov}(X, Y) + a^2 \text{Var}(X) ,$$

on a

$$f'(a) = -2 \text{cov}(X, Y) + 2a \text{Var}(X) .$$

Supposons dans un premier temps que $\text{Var}(X) > 0$. Un extremum de $f(a)$ est nécessairement atteint pour

$$a = \frac{\text{cov}(X, Y)}{\text{Var}(X)} .$$

Il s'agit en fait d'un minimum car $f''(a) = \text{Var}(X) > 0$. On trouve donc comme meilleure approximation linéaire de Y

$$\hat{Y} = m_Y + \frac{\text{cov}(X, Y)}{\text{Var}(X)}(X - m_X) .$$

Le cas $\text{Var}(X) = 0$ est trivial. On a en effet $X - m_X = 0$ presque-sûrement et tout revient alors à trouver un nombre b tel que $E[(Y - b)^2]$ soit minimum. D'après le lemme on trouve $b = m_Y$. La formule qui vient d'être donnée reste donc valable.

Interprétation géométrique

La *droite de régression* de Y par rapport à X est définie par

$$y = m_Y + \frac{\text{cov}(X, Y)}{\text{Var}(X)}(x - m_X) .$$

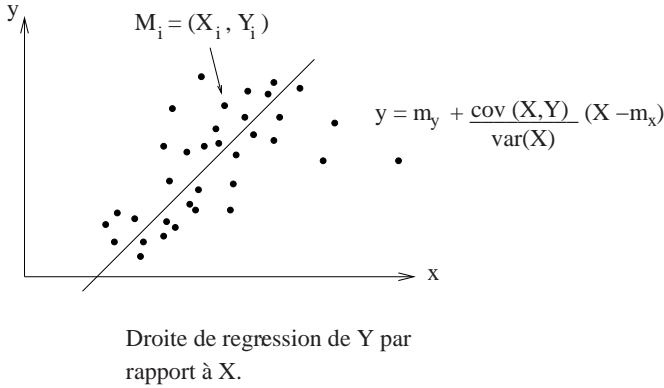
Cette droite passe par le point moyen (m_X, m_Y) .

Soit M le point aléatoire du plan, de coordonnées (X, Y) et soit Z_D la distance oblique (parallèlement à l'axe des ordonnées) de M à une droite D . Le résultat précédent dit que la droite de régression minimise $E[Z_D^2]$.

Soit maintenant n tirages indépendants de points $M_i = (X_i, Y_i)$, $(1 \leq i \leq n)$ selon une loi identique, celle de (X, Y) . Soit $Z_{D,i}$, $(1 \leq i \leq n)$, les distances obliques à la droite D associées à ces tirages. On verra plus tard la loi forte des grands nombres qui énonce dans ce cas particulier :

$$E[Z_D^2] = \lim_{n \uparrow \infty} \frac{1}{n} \sum_{i=1}^n Z_{D,i}^2 .$$

On peut donc dire que la droite de régression minimise asymptotiquement la dispersion $\frac{1}{n} \sum_{i=1}^n Z_{D,i}^2$ du nuage de points (M_1, \dots, M_n) autour d'elle.



Cas multidimensionnel

Soit X un n -vecteur aléatoire de carré intégrable et Y une variable aléatoire réelle de carré intégrable. Parmi toutes les approximations linéaires du type

$$\hat{Y} = m_Y + a^T(X - m_X)$$

où $a^T = (a_1, a_2, \dots, a_n)$, nous allons en chercher une qui minimise l'écart quadratique entre Y et son approximation \hat{Y} . En d'autres termes, on cherche à minimiser

$$f(a) = E[|(Y - m_Y) - a^T(X - m_X)|^2] .$$

Cette forme se réduit à

$$f(a) = \Sigma_Y - 2\Sigma_{YX} a + a^T \Sigma_X a .$$

Supposons que Σ_X est définie positive, et donc qu'il existe un inverse Σ_X^{-1} . Le vecteur a minimisant cette forme quadratique est alors donné par

$$a^T = \Sigma_{YX} \Sigma_X^{-1} .$$

La meilleure approximation linéaire-quadratique est donc

$$\hat{Y} = m_Y + \Sigma_{YX} \Sigma_X^{-1} (X - m_X) . \quad (3.9)$$

On verra dans l'Exercice 3.5.20 que l'erreur résiduelle est donnée par la formule :

$$E[|\hat{Y} - Y|^2] = \Sigma_Y - \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} . \quad (3.10)$$

3.4 Vecteurs aléatoires gaussiens

Pourquoi les gaussiennes ?

La popularité des modèles gaussiens tient à ce qu'ils se prêtent bien au calcul et de ce fait les problèmes relatifs à de tels modèles reçoivent souvent une solution on

analytique complète sous une forme agréable. Cet avantage est dû principalement aux propriétés de stabilité suivantes : d'une part les transformations affines des vecteurs gaussiens produisent des vecteurs gaussiens, d'autre part lorsque deux vecteurs sont gaussiens dans leur ensemble (voir la définition 3.4.4), le *conditionnement* de l'un par l'autre conduit à un vecteur gaussien qui est une combinaison du vecteur conditionnant (le conditionnement fait l'objet du chapitre 4).

Cependant, ceci ne suffirait pas à justifier l'usage des modèles gaussiens. Il y a une raison profonde qui fait que la distribution gaussienne apparaît *naturellement* dans les phénomènes aléatoires dont la base physique est de nature microscopique et qu'on observe à l'échelle macroscopique, comme par exemple le bruit de fond dans les émetteurs ou les récepteurs radioélectriques dû à l'agitation thermique des électrons (échelle *microscopique*) et qui se traduit par un courant électrique (échelle *macroscopique*). Les petites contributions supposées indépendantes (mais cette hypothèse n'est pas absolument nécessaire) de chaque électron, s'ajoutent pour former un courant gaussien, et cela quelle que soit la distribution statistique des contributions individuelles. En voici la raison profonde :

Imaginons que l'on ait à modéliser un phénomène macroscopique dont la mesure conduit à une variable aléatoire X de moyenne nulle. Des expériences répétées ont permis d'en évaluer empiriquement la variance σ^2 et une analyse physique du phénomène microscopique sous-jacent a conduit à penser que X est la somme d'un "très grand" nombre de "petites" variables aléatoires indépendantes et indistinctement distribuées, et de moyenne nulle :

$$X = Y_1 + \cdots + Y_n .$$

La variance commune des Y_i est donc, d'après le théorème d'addition des variances, égale à $\frac{\sigma^2}{n}$. Si on effectue le changement $X_i = Y_i\sqrt{n}$, la variable X s'écrit

$$X = \frac{X_1 + \cdots + X_n}{\sqrt{n}} ,$$

où les X_i sont indépendantes dans leur ensemble, indistinctement distribuées et de variance σ^2 . On sait simplement que n est très grand, mais on n'a aucune raison de choisir telle ou telle valeur. Dans ces conditions, il semble naturel de donner à n une valeur infinie. Il faut entendre par là la chose suivante : on choisit pour distribution de X la limite de la distribution de $\frac{X_1 + \cdots + X_n}{\sqrt{n}}$ lorsque n tend vers l'infini :

$$P(X \leq x) = \lim_{n \uparrow \infty} P\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}} \leq x\right) ,$$

Il se trouve, et c'est le contenu du fameux *théorème de la loi gaussienne limite* (qui fait l'objet du chapitre 7) que la limite du deuxième membre de l'égalité ci-dessus est

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\frac{y^2}{\sigma^2}} dy .$$

La variable aléatoire X est donc gaussienne.

Aucune hypothèse n'a été faite *sur la forme* de la distribution des contributions microscopiques Y_i , ce qui explique pourquoi les variables aléatoires gaussiennes sont omniprésentes.

Deux définitions équivalentes

Rappelons d'abord la définition d'une *variable aléatoire gaussienne réduite* ou *standard* : c'est une variable aléatoire réelle X de densité de probabilité

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (x \in \mathbb{R}) .$$

Sa moyenne est nulle, sa variance est égale à 1 et sa fonction caractéristique est $e^{-\frac{1}{2}u^2}$. La variable aléatoire

$$Y = \sigma X + m$$

a pour moyenne m et pour variance σ^2 et si $\sigma^2 > 0$, sa densité est donnée par

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-m}{\sigma}\right)^2} .$$

La définition de Y garde un sens si $\sigma = 0$, et on a alors $Y = m$. Ceci suggère une légère extension de la définition d'une variable gaussienne qui inclut ce cas limite.

Définition 3.4.1 *Une variable aléatoire gaussienne au sens large est une variable aléatoire réelle dont la fonction caractéristique est*

$$\varphi(u) = e^{ium} e^{-\frac{1}{2}\sigma^2 u^2} .$$

où m est un nombre réel et σ est un nombre réel positif ou nul.

Le cas où $\sigma = 0$ correspond à une variable aléatoire “déterministe” égale à m .

Nous allons donner la première définition d'un vecteur gaussien.

Définition 3.4.2 *On dit que le vecteur aléatoire $X = (X_1, \dots, X_n)$ à composantes réelles est un vecteur aléatoire gaussien si pour tous réels a_0, a_1, \dots, a_n , la variable aléatoire $Y = a_0 + \sum_{j=1}^n a_j X_j$ est une variable aléatoire gaussienne au sens large.*

Calculons la fonction caractéristique $\varphi_X(u) = E[e^{iu^T X}]$ d'un tel vecteur aléatoire. La variable aléatoire $Z = u^T X$ est gaussienne et donc

$$\varphi_X(u) = \varphi_Z(1) = e^{im_Z} e^{-\frac{1}{2}\sigma_Z^2} .$$

Il reste à calculer m_Z et σ_Z . On trouve

$$m_Z = E[Z] = E[u^T X] = u^T m_X .$$

et

$$\begin{aligned}\sigma_Z^2 &= E[(Z - m_Z)^2] = E[u^T(X - m_X)(X - m_X)^T u] \\ &= u^T E[T(X - m_X)(X - m_X)^T] u = u^T \Sigma_X u,\end{aligned}$$

d'où le résultat :

$$\varphi_X(u) = e^{iu^T m_X - \frac{1}{2}u^T \Sigma_X u},$$

où Σ_X est la matrice de covariance de X .

Voici maintenant une autre définition d'un vecteur gaussien, qui est l'extension naturelle de la définition 3.4.1 d'une variable gaussienne.

Définition 3.4.3 *On dit que le vecteur aléatoire $X = (X_1, \dots, X_n)$ à composantes réelles est un vecteur aléatoire gaussien s'il admet pour fonction caractéristique*

$$\varphi_X(u) = e^{iu^T m_X - \frac{1}{2}u^T \Sigma_X u},$$

où m_X est sa moyenne et Σ_X sa matrice de covariance.

Les deux définitions sont équivalentes. Le calcul qui suit la définition 3.4.2 montre qu'un vecteur gaussien au sens de cette définition l'est aussi au sens de la définition 3.4.3. Supposons maintenant que X est gaussien selon la définition 3.4.3. Soit donc une combinaison affine réelle $Y = a_0 + \sum_{j=1}^n a_j X_j$. On doit montrer que cette variable est gaussienne au sens de la définition étendue 3.4.1. On peut se contenter de le faire dans le cas $a_0 = 0$. On a donc $Y = a^T X$. Soit v un réel.

$$E[e^{ivY}] = E[e^{iv(a^T X)}] = E[e^{iu^T X}]$$

où $u = va$. On a donc

$$E[e^{ivY}] = e^{iva^T m_X - \frac{1}{2}v^2 a^T \Sigma_X a}.$$

Comme $m_Y = a^T m_X$ et $\sigma_Y^2 = a^T \Sigma_X a$, on a bien

$$E[e^{ivY}] = e^{ivm_Y - \frac{1}{2}v^2 \sigma_Y^2}.$$

Densité d'un vecteur gaussien non dégénéré

La matrice de covariance Σ_X d'un vecteur aléatoire de \mathbb{R}^n est une matrice symétrique non négative. D'après un résultat classique d'algèbre linéaire, une telle matrice admet une décomposition (non unique)

$$\Sigma_X = AA^T,$$

où A est une matrice $n \times r$ de rang r , où le nombre r est également le rang de Σ_X . Si $r = n$, c'est-à-dire si Σ_X est inversible, A est une matrice $n \times n$ qu'on peut choisir inversible.

Théorème 3.4.1 Soit X un vecteur aléatoire gaussien de \mathbb{R}^n , non dégénéré, c'est-à-dire, de covariance Σ_X inversible. Soit m_X sa moyenne. Un tel vecteur aléatoire admet la densité de probabilité

$$f_X(x) = \frac{1}{(2\pi)^{n/2}(\det \Sigma_X)^{1/2}} e^{-\frac{1}{2}(x-m_X)^T \Sigma_X^{-1}(x-m_X)} .$$

D'autre part le vecteur

$$Y = A^{-1}(X - m_X) ,$$

où A est une matrice carrée inversible telle que $\Sigma_X = AA^T$ est un vecteur gaussien standard, c'est-à-dire dont les coordonnées sont des gaussiennes réduites indépendantes dans leur ensemble.

Démonstration. Soit A une matrice $n \times n$ inversible telle que $\Sigma_X = AA^T$. Le vecteur aléatoire

$$Z = A^{-1}(X - \mu_X)$$

est gaussien, en tant que combinaison affine de vecteur gaussien. D'autre part $m_Z = 0$ et sa matrice de covariance est l'unité, comme il découle du calcul suivant :

$$\sigma_Z^2 = A^{-1}\Sigma_X(A^{-1})^T = A^{-1}AA^T(A^T)^{-1} .$$

Il admet donc pour fonction caractéristique

$$\varphi_Z(u_1, \dots, u_n) = \prod_{j=1}^n e^{-\frac{1}{2}u_j^2} .$$

Chaque Z_j en tant que combinaison affine des X_ℓ est une variable gaussienne. D'autre part, comme on vient de voir, $m_{Z_j} = 0$ et $\sigma_{Z_j}^2 = 1$. La fonction caractéristique de Z_j est donc donnée par

$$\varphi_{Z_j}(u_j) = e^{-\frac{1}{2}u_j^2} .$$

On a donc que

$$\varphi_Z(u_1, \dots, u_n) = \prod_{j=1}^n \varphi_{Z_j}(u_j) ,$$

ce qui revient à dire que les variables aléatoires Z_j ($1 \leq j \leq n$) sont indépendantes dans leur ensemble. La densité de probabilité du vecteur Z est donc le produit des densités de probabilité des variables aléatoires Z_j :

$$f_Z(z_1, \dots, z_n) = \prod_{j=1}^n f_{Z_j}(z_j) .$$

Mais les Z_j étant des gaussiennes standard,

$$f_Z(z_1, \dots, z_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^n z_j^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}z^T z} .$$

La densité de X est donc (formule du changement de variables dans le cas des transformations linéaires ; voir le chapitre 3) :

$$f_X(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(A^{-1}\tilde{x})^T (A^{-1}\tilde{x})} |\det(A^{-1})| ,$$

où $\tilde{x} = x - \mu_X$. Or

$$(A^{-1}\tilde{x})^T (A^{-1}\tilde{x}) = \tilde{x}^T (A^{-1}) A^{-1} \tilde{x} = \tilde{x}^T (AA')^{-1} \tilde{x} = \tilde{x} \Sigma_X^{-1} \tilde{x}$$

et

$$\det A^{-1} = (\det \Sigma_X)^{-\frac{1}{2}},$$

d'où le résultat annoncé. \square

Théorème 3.4.2 *Soit X un n -vecteur aléatoire gaussien non-dégénéré dont la matrice de covariance est diagonale :*

$$\Sigma_X = \text{diag} \{ \sigma_1^2, \dots, \sigma_n^2 \}$$

où les σ_i^2 sont des nombres strictement positifs. Alors les coordonnées X_i ($1 \leq i \leq n$) sont des variables gaussiennes indépendantes et de variances σ_i^2 ($1 \leq i \leq n$) respectivement.

Démonstration. La fonction caractéristique de X s'écrit

$$\varphi_X(u) = \prod_{j=1}^n (e^{iu_j m_{X_j} - \frac{1}{2} \sigma_j^2 u_j^2}) .$$

Or $e^{iu_j m_{X_j} - \frac{1}{2} \sigma_j^2 u_j^2}$ n'est autre que la fonction caractéristique de X_j puisque

$$\varphi_{X_j}(u_j) = \varphi_X(0, \dots, 0, u_j, 0, \dots, 0) .$$

En particulier X_j est une variable aléatoire gaussienne de variance σ_j^2 et de moyenne m_{X_j} . Comme

$$\varphi_X(u) = \prod_{j=1}^n \varphi_{X_j}(u_j) ,$$

les X_j sont indépendants dans leur ensemble (critère d'indépendance de la fonction caractéristique). \square

EXEMPLE 3.4.1: VECTEUR GAUSSIEN À DEUX DIMENSIONS. A titre d'illustration des résultats précédents, nous allons considérer le cas bi-dimensionnel et rechercher une transformation affine explicite $Z = A^{-1}(X - \mu_X)$ qui fasse de Z un vecteur gaussien standard.

Soit donc $X = (X_1, X_2)$ un vecteur gaussien bidimensionnel de moyenne nulle et de matrice de covariance

$$\Sigma_X = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

Noter que $\sigma_{12} = \sigma_{21}$. On suppose que $\det \Sigma_X = \sigma_1^2 \sigma_2^2 - \sigma_{21}^2 > 0$. La matrice de covariance est donc inversible et son inverse est

$$\Sigma_X^{-1} = (\sigma_1^2 \sigma_2^2 - \sigma_{21}^2)^{-1} \begin{pmatrix} \sigma_2^2 & -\sigma_{21} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix},$$

d'où

$$x^T \Sigma_X^{-1} x = \frac{\sigma_1^2 x_1^2 - 2\sigma_{12} x_1 x_2 + \sigma_2^2 x_2^2}{\sigma_1^2 \sigma_2^2 - \sigma_{21}^2} = \frac{\frac{x_1^2}{\sigma_1^2} - 2\rho \frac{x_1 x_2}{\sigma_1 \sigma_2} + \frac{x_2^2}{\sigma_2^2}}{1 - \rho^2},$$

où

$$\rho = \frac{\sigma_{21}}{\sigma_1 \sigma_2}$$

est le coefficient de corrélation de X_1 et X_2 . Finalement :

$$\begin{aligned} f_X(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1 x_2}{\sigma_1 \sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}Q(x_1, x_2)}. \end{aligned}$$

Pour trouver une transformation affine qui transforme le 2-vecteur X en un 2-vecteur dont les coordonnées sont des gaussiennes standard indépendantes, on complète les carrés dans $Q(x_1, x_2)$. On obtient par exemple $Q(x_1, x_2) = \frac{x_1/\sigma_1 - x_2/\sigma_2}{\sqrt{1-\rho^2}} + \frac{x_2^2}{\sigma_2^2}$. Le choix suivant de Z résout le problème posé :

$$\begin{cases} Z_1 = \left(\frac{X_1}{\sigma_1} - \rho\frac{X_2}{\sigma_2}\right)\frac{1}{\sqrt{1-\rho^2}} \\ Z_2 = \frac{X_2}{\sigma_2}. \end{cases}$$

Vecteur gaussien dégénéré

Lorsque la matrice de covariance Σ_X est de rang r strictement inférieur à n , on sait (Théorème 3.3.5) que le vecteur aléatoire X est dans un hyperplan de \mathbb{R}^n de dimension r et n'admet donc pas de densité de probabilité.

Indépendance et non-corrélation dans les vecteurs gaussiens

On dit que deux vecteurs aléatoires (pas nécessairement gaussiens) sont non-corrélés si leur matrice de corrélation

$$\Sigma_{XY} = E[(X - m_X)(Y - m_Y)^T]$$

est indistinctement nulle. (La matrice Σ_{XY} est une matrice $n \times p$ où n et p sont les dimensions de X et Y respectivement. Son terme (i, j) est la covariance de X_i et Y_j

$$\sigma_{X_i Y_j} = E[(X_i - m_{X_i})(Y_j - m_{Y_j})]$$

Les notions de corrélation et de dépendance sont liées mais pas équivalentes. L'indépendance entraîne la non-corrélation, mais la réciproque n'est pas vraie en général, sauf pour les vecteurs gaussiens où elle est "*à moitié vraie*". Il faut bien prendre garde à ne pas tomber dans l'erreur qui consiste à dire que si X et Y sont deux vecteurs gaussiens non corrélés, alors ils sont indépendants. En effet :

EXEMPLE 3.4.2: UN CONTRE-EXEMPLE. X est une variable aléatoire gaussienne réduite et $Y = UX$, où U est une variable aléatoire telle que $P(U = 1) = P(U = -1) = \frac{1}{2}$, indépendante de X . La variable aléatoire $Y = UX$ a la même distribution que X car $P(Y \leq x) = P(UX \leq x) = P(X \leq x, U = 1) + P(-X \leq x, U = -1) = \frac{1}{2} P(X \leq x) + \frac{1}{2} P(X \geq -x) = P(X \leq x)$. D'autre part X et Y ne sont pas corrélées, en effet

$$E[XY] = E[UX^2] = E[U]E[X^2] = 0 ,$$

puisque X et U sont indépendantes et U a une moyenne nulle. Les variables aléatoires X et Y sont donc gaussiennes et non-corrélées. Mais elles ne sont pas indépendantes, comme on peut le voir par exemple en constatant que $E[X^2 Y^2]$ n'est pas égale à $E[X^2]E[Y^2] = 1$. En effet, $E[X^2 Y^2] = E[X^4] = 3$.

Avant d'énoncer le résultat correct, il nous faut une définition.

Définition 3.4.4 Les vecteurs aléatoires Y_1, \dots, Y_k sont dits gaussiens dans leur ensemble si toute combinaison affine de leurs coordonnées est une variable aléatoire gaussienne.

En d'autres termes le "grand" vecteur Y obtenu en mettant bout à bout les vecteurs Y_1, \dots, Y_k est un vecteur gaussien.

Théorème 3.4.3 Si X et Y sont deux vecteurs gaussiens dans leur ensemble qui sont non-corrélés, alors X et Y sont indépendants.

Démonstration. Comme X et Y sont gaussiens dans leur ensemble, le vecteur $Z = (X, Y)$ est gaussien. Par définition de la non-corrélation de X et Y , $\Sigma_{YX}^T = \Sigma_{XY} = E[(X - m)(Y - m)^T] = 0$ et donc la matrice de covariance Σ_Z prend la forme

$$\Sigma_Z = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix}.$$

En particulier $\det \Sigma_Z = \det \Sigma_X \det \Sigma_Y$ et $w^T \Sigma_Z w = u^T \Sigma_X u + v^T \Sigma_Y v$ où $w^T = (u^T, v^T)$ et u et v ont les dimensions appropriées. La fonction caractéristique de Z s'écrit donc sous forme produit $\varphi_Z(w) = \varphi_X(u)\varphi_Y(v)$ ce qui démontre l'indépendance de X et Y . \square

3.5 Exercices

Exercice 3.5.1. CARACTÉRISTIQUES DE LA DISTRIBUTION GAUSSIENNE.

Montrez que

$$\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi},$$

et vérifiez que la densité $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$:

- (i) $\int_{-\infty}^{+\infty} f_X(x) dx = 1$
- (ii) $m = \int_{-\infty}^{+\infty} x f_X(x) dx$
- (iii) $\sigma^2 = \int_{-\infty}^{+\infty} (x - m)^2 f_X(x) dx$
- (iv) $e^{i u m - \frac{1}{2} \sigma^2 u^2} = \int_{-\infty}^{+\infty} e^{i u x} f_X(x) dx$.

Exercice 3.5.2. CARACTÉRISTIQUES DE LA DISTRIBUTION GAMMA.

Montrer que pour $\alpha > 0, \beta > 0$, la fonction

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} 1_{\{x \geq 0\}}$$

est une densité de probabilité. Montrez que si une v.a. X admet cette fonction pour densité de probabilité alors

- (i) $m_X = \frac{\alpha}{\beta}$
- (ii) $\sigma_X^2 = \frac{\alpha}{\beta^2}$
- (iii) $\varphi_X(u) = E[e^{i u X}] = (1 - \frac{i u}{\beta})^{-\alpha}$.

Exercice 3.5.3. SOMME DE GAMMAS.

Soit X_1 et X_2 deux variables aléatoires qui suivent des lois $\gamma(\alpha_1, \beta)$ et $\gamma(\alpha_2, \beta)$ ($\alpha_1 > 0, \alpha_2 > 0, \beta > 0$). Montrez que si X_1 et X_2 sont indépendantes, alors

$$X_1 + X_2 \sim \gamma(\alpha_1 + \alpha_2, \beta).$$

Exercice 3.5.4.

Soit X une variable aléatoire de densité de probabilité f_X et soit Y une variable aléatoire discrète prenant les valeurs réelles y_1, y_2, \dots avec les probabilités p_1, p_2, \dots . Montrez que si X et Y sont indépendantes $Z = X + Y$ admet une densité de probabilité, et calculez cette dernière.

Exercice 3.5.5. SOMME DE CARRÉS DE GAUSSIENNES STANDARD INDÉPENDANTES.

Montrez que si X_1, X_2, \dots, X_n sont des variables aléatoires gaussiennes standard indépendantes, alors

$$X_1^2 + \dots + X_n^2 \stackrel{D}{\sim} \gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

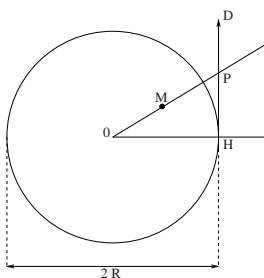
(c'est-à-dire une loi du khi-deux à n degrés de liberté).

Exercice 3.5.6. LA DISTRIBUTION EXPONENTIELLE N'A PAS DE MÉMOIRE.

Montrez que si X suit une loi exponentielle, alors $P(X \geq x + y | X \geq y) = P(X \geq x)$.

Exercice 3.5.7. * DANS LE DISQUE UNITÉ.

Un point M est tiré au hasard, uniformément, sur un disque de centre 0 et de rayon R .



Quelle est la fonction de répartition de la variable aléatoire X égale à la longueur du segment OM ? Quelle est la densité de probabilité de la variable aléatoire Y égale à la longueur du segment HP ? On note Θ l'angle (mesuré dans le sens trigonométrique) entre les segments OH et OM . En particulier $\Theta \in [0, 2\pi)$. Montrez que Θ et X sont des variables aléatoires indépendantes, et calculez leurs densités de probabilité.

Exercice 3.5.8. * DENSITÉ DE PROBABILITÉ DE LA VALEUR ABSOLUE.

Montrez que si X admet une densité f_X symétrique ($f_X(x) = f_X(-x)$ pour tout $x \in \mathbb{R}$), alors la densité $f_{|X|}$ de $|X|$ est reliée à celle de X par

$$f_X(x) = \frac{1}{2} f_{|X|}(|x|) .$$

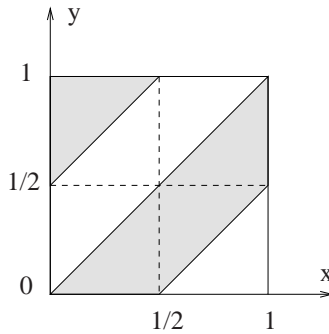
Exercice 3.5.9.

a. Soit X_1, \dots, X_n n ($n \geq 2$) des variables aléatoires réelles indépendantes et identiquement distribuées, de fonction de répartition commune F . Quelles sont les fonctions de répartition de $Y = \max(X_1, \dots, X_n)$ et $Z = \min(X_1, \dots, X_n)$?

b. Calculez les densités de probabilité de Y et Z lorsque $X_i \sim \mathcal{E}(\lambda)$.

Exercice 3.5.10. *

Soit X et Y deux variables aléatoires telles de densité de probabilité $f_{X,Y}$ constante sur la région en pointillé, nulle ailleurs.



Quelle est la valeur de la constante ? Calculez les densités de X et de Y . En déduire que X et Y ne sont pas indépendantes. Montrez que les fonctions caractéristiques de X , Y et $X + Y$ vérifient pour tout $u \in \mathbb{R}$:

$$\varphi_X(u)\varphi_Y(u) = \varphi_{X+Y}(u). \quad (\star)$$

Commentaire : La condition (\star) est une condition nécessaire d'indépendance, mais pas une condition suffisante, comme on vient de le démontrer par un contre-exemple.

Exercice 3.5.11. *

Soit X et Y deux variables aléatoires indépendantes. X admet la fonction caractéristique $\varphi_X(u)$ et Y est une variable aléatoire discrète prenant les valeurs $0, 1, 2, \dots$ avec les probabilités p_0, p_1, p_2, \dots . Quelle est la fonction caractéristique $\varphi_Z(u)$ de la variable aléatoire $Z = YX$?

Exercice 3.5.12. *

Soit X_1 et X_2 deux variables aléatoires indépendantes, identiquement distribuées selon la loi exponentielle de paramètre λ . Trouvez la distribution du vecteur $Y = (Y_1, Y_2)$ donné par

$$Y_1 = \frac{X_1}{X_2}, \quad Y_2 = X_1^2 + X_2^2.$$

Exercice 3.5.13. *

Soit X_1 et X_2 deux variables aléatoires indépendantes $X_i \sim \gamma(\alpha_i, 1)$ ($i = 1, 2$), où $\alpha_i > 0$ ($i = 1, 2$). Calculez la densité de probabilité de $Y_1 = X_1/(X_1 + X_2)$ et celle de $Y_1 = X_1 + X_2$. Que peut-on dire de Y_1 et Y_2 ?

Exercice 3.5.14. *

De quelle variable aléatoire la fonction suivante est-elle la fonction caractéristique ?

$$\varphi(u) = \exp e^{-|u|} - 1$$

(SVP, pas de transformation de Fourier...)

Exercice 3.5.15. PROPRIÉTÉS DES FONCTIONS DE RÉPARTITION.

Soit X une variable aléatoire réelle définie sur l'espace de probabilité (Ω, \mathcal{F}, P) . Montrez que sa fonction de répartition F_X vérifie les propriétés suivantes :

- (i) l'application $x \rightarrow F_X(x)$ est croissante
- (ii) $\lim_{x \downarrow -\infty} F_X(x) = 0$
- (iii) $\lim_{x \uparrow +\infty} F_X(x) = +1$
- (iv) l'application $x \rightarrow F_X(x)$ est continue à droite.

Exercice 3.5.16. LA I-ÈME PLUS PETITE.

La situation et les notations sont celles du Théorème 3.1.9. Calculez la distribution de probabilité de Z_i , la i -ème plus petite variable parmi X_1, \dots, X_n , où les X_i sont indépendantes et uniformément distribuées sur $[0, 1]$.

Exercice 3.5.17. NON CORRÉLÉES MAIS PAS INDÉPENDANTES.

Soit Z une variable aléatoire uniformément distribuée sur $[0, 2\pi]$. Montrez que les variables aléatoires $X = \sin Z$ et $Y = \cos Z$ ne sont pas corrélées, mais qu'elles ne sont pas indépendantes.

Exercice 3.5.18. *

Soit X et Y deux variables aléatoires réelles de carré intégrable, et soit $a > 0, b, c > 0, d \in \mathbb{R}$. Montrez que $\rho(aX + b, cY + d) = \frac{ac}{|ac|} \rho(X, Y)$.

Exercice 3.5.19. AXES CONJUGUÉS ET GAUSSIENNES.

Trouvez la droite de régression de X_2 par rapport à X_1 dans le cas où

$$f_{X_1 X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)}.$$

Montrez que c'est l'axe conjugué de $0x_1$ pour l'ellipse d'équidensité :

$$Q(x_1, x_2) = \frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2} = \lambda^2.$$

Exercice 3.5.20.

Démontrez la formule (3.10).

Exercice 3.5.21. SIGNAL PLUS BRUIT.

Un signal aléatoire n -dimensionnel S est noyé dans un bruit n -dimensionnel B , le signal reçu X étant de la forme

$$X = S + B$$

(on dit alors que le bruit est additif). Le signal et le bruit ne sont pas corrélés.

(1) Calculez la meilleure approximation quadratique \hat{S} de S comme fonction affine de X .

(2) Montrez que

$$E[(\hat{S} - S)(\hat{S} - S)^T] = \Sigma_B \Sigma_S (\Sigma_S + \Sigma_B)^{-1}.$$

Exercice 3.5.22. * INÉGALITÉ DU TRIANGLE.

Soit X et Y deux variables aléatoires réelles de carré intégrable. Démontrez l'inégalité du triangle

$$E[(X + Y)^2]^{1/2} \leq E[X^2]^{1/2} + E[Y^2]^{1/2}$$

Exercice 3.5.23.

Calculez la fonction caractéristique de la variable X qui admet la densité

$$f_X(x) = \frac{a}{2} e^{-a|x|}.$$

(exponentielle symétrique). Utilisez ce résultat pour obtenir la fonction caractéristique de la distribution de Cauchy.

Exercice 3.5.24. *

Soit $\{X_n\}_{n \geq 1}$ une suite de variables indépendantes distribuées selon la loi de Cauchy, et soit T une variable aléatoire à valeurs entières positives, indépendante de cette suite. Montrez que

$$X = \frac{X_1 + \cdots + X_T}{T}$$

suit également une loi de Cauchy.

Exercice 3.5.25.

X_1 et X_2 sont deux variables gaussiennes réduites indépendantes. Calculez la fonction caractéristique de $Z = \frac{X_1^2 - X_2^2}{4}$.

Exercice 3.5.26.

Soit X une gaussienne standard.

- (1) Montrez que $Y_a = X1_{\{|X|<a\}} - X1_{\{|X|\geq a\}}$ ($a > 0$) est une gaussienne standard.
- (2) Montrez que (X, Y_a) n'est pas un vecteur gaussien.
- (3) En choisissant a tel que $\frac{1}{\sqrt{2\pi}} \int_0^a t^2 e^{-\frac{t^2}{2}} dt = \frac{1}{4}$, montrez que $\text{cov}(X, Y_a) = 0$.

Chapitre 4

Espérance conditionnelle

4.1 Définitions et propriétés élémentaires

Cas avec densité de probabilité

Soit X et Y deux vecteurs aléatoires, de dimension ℓ et m respectivement. On suppose que le vecteur $Z = (X, Y)$ admet une densité de probabilité $f_{X,Y}$. On notera f_Y la densité de probabilité de Y , qu'on obtient par la formule

$$f_Y(y) = \int_{\mathbb{R}^\ell} f_{X,Y}(x, y) dx.$$

Définition 4.1.1 Pour $y \in \mathbb{R}^\ell$ fixé tel que $f_Y(y) > 0$, la fonction de x définie par

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad (4.1)$$

est appelée la densité de probabilité conditionnelle de X sachant (étant donné, conditionnellement à)¹ $Y = y$.

On constate en effet que si y est fixé tel que $f_Y(y) > 0$, alors $f_{X|Y}(x|y)$ est, en tant que fonction de x , une densité de probabilité. En effet, $f_{X|Y}(x|y) \geq 0$ et

$$\int_{\mathbb{R}^\ell} f_{X|Y}(x|y) dx = \frac{\int_{\mathbb{R}^\ell} f_{X,Y}(x, y) dx}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1.$$

Voici une caractérisation de la densité conditionnelle qui est à la fois évidente et utile.

¹Ces alternatives à “sachant” seront utilisées indifféremment dans la suite.

Théorème 4.1.1 Soit X et Y deux vecteurs aléatoires de dimension ℓ et m respectivement et tels que $Z = (X, Y)$ admette la densité de probabilité $f_{X,Y}$. Si on sait trouver deux fonctions $a : \mathbb{R}^m \rightarrow \mathbb{R}_+$ et $b : \mathbb{R}^{\ell \times m} \rightarrow \mathbb{R}_+$ telles que pour tous $x \in \mathbb{R}^\ell$, $y \in \mathbb{R}^m$,

$$f_{XY}(x, y) = a(y) b(x, y)$$

et

$$\int_{\mathbb{R}^\ell} b(x, y) dx = 1,$$

alors $a(y) = f_Y(y)$ et $b(x, y) = f_{X|Y}(x|y)$.

Démonstration. Il suffit de remarquer que

$$\begin{aligned} f_Y(y) &= \int_{\mathbb{R}^\ell} f_{X,Y}(x, y) dx = \int_{\mathbb{R}^\ell} a(y) b(x, y) dx \\ &= a(y) \int_{\mathbb{R}^\ell} b(x, y) dx = a(y). \end{aligned}$$

□

EXEMPLE 4.1.1: Soit (X_1, X_2) un vecteur gaussien bidimensionnel de densité de probabilité

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left(\frac{x_1^2}{\sigma_1^2} - 2\rho \frac{x_1 x_2}{\sigma_1 \sigma_2} + \frac{x_2^2}{\sigma_2^2} \right)}.$$

Comme la variable aléatoire X_1 est une gaussienne centrée de variance σ_1^2 (voir l'Exemple 3.4.1) :

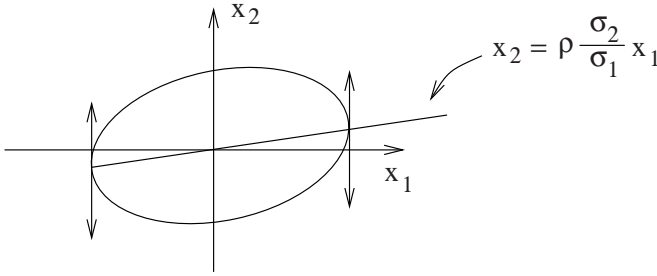
$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2} \frac{x_1^2}{\sigma_1^2}},$$

la densité conditionnelle de X_2 sachant que $X_1 = x_1$ est, en appliquant la formule (4.1),

$$f_{X_2|X_1}(x_2|x_1) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2} \frac{(x_2 - \rho \frac{\sigma_2}{\sigma_1} x_1)^2}{\sigma_2^2(1-\rho^2)}}.$$

C'est une densité gaussienne de moyenne $\rho \frac{\sigma_2}{\sigma_1} x_1$ et de variance $\sigma_2^2(1-\rho^2)$.

La figure ci-dessous relie la moyenne de cette densité de probabilité conditionnelle à l'axe conjugué de l'axe $0x_2$ dans une ellipse d'équiprobabilité (par exemple, celle d'équation $f_{X_1, X_2}(x_1, x_2) = 1$).



Définition 4.1.2 Soit $g : \mathbb{R}^\ell \times \mathbb{R}^m \rightarrow \mathbb{R}$ une fonction ou bien non négative, ou bien telle que la variable aléatoire $g(X, Y)$ est intégrable. On définit pour tout y tel que $f_Y(y) > 0$

$$h(y) = \int_{\mathbb{R}^\ell} g(x, y) f_{X|Y}(x|y) dx, \quad (4.2)$$

et on appelle cette quantité, notée $E[g(X, Y) | Y = y]$ ou $E^{Y=y}[g(X, Y)]$, l'espérance conditionnelle de $g(X, Y)$ sachant $Y = y$. La variable aléatoire $h(Y)$ est appelée espérance conditionnelle de $g(X, Y)$ sachant Y , et elle est notée $E[g(X, Y) | Y]$ ou $E^Y[g(X, Y)]$.

On verra dans le Chapitre 6 une définition plus générale de l'espérance conditionnelle. On démontrera alors que la fonction h est bien définie (l'intégrale définissant cette quantité a un sens), sauf évidemment sur l'ensemble $A = \{y; f_Y(y) = 0\}$. En fait, on peut donner une valeur arbitraire à cette fonction sur cet ensemble particulier, car la probabilité que $Y \in A$ est nulle. En effet

$$P(Y \in A) = P(f_X(Y) = 0) = \int_{\mathbb{R}^m} 1_{\{f_Y(y)=0\}} f_Y(y) dy = 0.$$

On verra aussi que si $g(X, Y)$ est intégrable, alors $E^Y[g(X, Y)]$ l'est aussi.

EXEMPLE 4.1.2: Soit (X_1, X_2) le vecteur gaussien de l'Exemple 4.1.1. La quantité

$$E[X_2 | X_1 = x_1] = \int_{-\infty}^{+\infty} x_2 f_{X_2|X_1}(x_2 | x_1) dx_2$$

est la moyenne conditionnelle de X_2 étant donné $X_1 = x_1$, c'est-à-dire la moyenne de la distribution décrite par la densité conditionnelle de X_2 étant donné $X_1 = x_1$. Or, comme on l'a vu, cette densité est gaussienne de moyenne $\rho \frac{\sigma_2}{\sigma_1} x_1$, d'où l'expression de l'espérance conditionnelle de X_2 étant donné X_1 :

$$E[X_2 | X_1] = \rho \frac{\sigma_2}{\sigma_1} X_1.$$

Les cas mixtes

Nous avons considéré le cas où le vecteur (X, Y) admet une densité de probabilité, qu'on appellera le cas "densité-densité". Nous allons maintenant dire un mot des cas "densité-discret" et "discret-densité", c'est-à-dire des cas où l'une des variables aléatoires X ou Y est discrète et l'autre admet, conditionnellement à la première, une densité de probabilité.

Soit donc (X, Y) un vecteur aléatoire où Y est une variable aléatoire discrète, dont on peut supposer sans restreindre la généralité qu'elle prend des valeurs entières positives, et X est un vecteur aléatoire de \mathbb{R}^ℓ . La loi de Y est décrite par sa distribution :

$$p_k = P(Y = k) .$$

Supposons que, conditionnellement à $Y = k$, X admette une densité de probabilité $f(x|k)$, c'est-à-dire :

$$P(X \in A | Y = k) = \int_A f(x|k) dx .$$

La loi du couple (X, Y) est entièrement déterminée par les probabilités p_k et les densités de probabilité $f(x|k)$. En effet, d'après la règle des causes totales,

$$P(X \in A, Y \in B) = \sum_{k \in B} p_k \int_A f(x|k) dx .$$

La densité $f(x|k)$ étant la densité de X conditionnée par $Y = k$, on définit pour toute variable $g(X, Y)$ qui est soit non négative, soit intégrable,

$$h(k) = \int_{\mathbb{R}^\ell} g(x, k) f(x|k) dx ,$$

puis l'espérance conditionnelle de $g(X, Y)$ sachant $X = k$:

$$E[g(X, Y) | Y = k] = h(k) ,$$

et enfin l'espérance conditionnelle de $g(X, Y)$ sachant X :

$$E[g(X, Y) | Y] = h(Y) .$$

Voyons maintenant le cas "discret-densité" : Y et X sont comme on vient de voir, mais nous allons nous intéresser cette fois-ci au conditionnement de Y par X (et non plus de X par Y). On définira la distribution conditionnelle de Y étant donné $X = x$ comme étant la distribution discrète

$$p_{k|x} = \frac{p_k f(x|k)}{\sum_{k \geq 1} p_k f(x|k)} .$$

On notera que $\sum_{k \geq 1} p_k f(x|k)$ n'est autre que la densité marginale de X , puisque d'après la règle des causes totales,

$$\begin{aligned} P(X \in A) &= \sum_{k \geq 1} P(Y = k) P(X \in A | Y = k) \\ &= \sum_{k \geq 1} p_k \int_A f(x|k) dx \\ &= \int_A \left(\sum_{k \geq 1} p_k f(x|k) \right) dx. \end{aligned}$$

En suivant le chemin maintenant familier, on définira

$$h(x) = \sum_{k \geq 1} g(x, k) p_{k|x},$$

puis l'espérance conditionnelle de $g(X, Y)$ (supposée soit non négative, soit intégrable) étant donné $X = x$:

$$E[g(X, Y) | X = x] = h(x),$$

et enfin l'espérance conditionnelle de $g(X, Y)$ étant donné X :

$$E[g(X, Y) | X] = h(X).$$

Propriétés élémentaires de l'espérance conditionnelle

Les démonstrations ne seront faites que pour les cas densité-densité. Leur adaptation aux autres cas est immédiate. (De plus, les preuves seront donnés au chapitre 6 dans une situation plus générale.)

L'espérance conditionnelle a, comme l'espérance, les propriétés de linéarité et de monotonie :

Théorème 4.1.2 Linéarité. Soit λ_1 et λ_2 des nombres réels. On a

$$E[\lambda_1 g_1(X, Y) + \lambda_2 g_2(X, Y) | Y] = \lambda_1 E[g_1(X, Y) | Y] + \lambda_2 E[g_2(X, Y) | Y],$$

dès que les deux membres de cette égalité sont bien définis (et en particulier, ne font pas intervenir de forme indéterminée $+\infty - \infty$).

Théorème 4.1.3 Monotonie. On a

$$P(g_1(X, Y) \leq g_2(X, Y)) = 1 \Rightarrow E[g_1(X, Y) | Y] \leq E[g_2(X, Y) | Y],$$

dès que les espérances conditionnelles en question sont bien définies.

Les démonstrations de ces deux théorèmes sont évidentes. La propriété suivante dit que "l'espérance de l'espérance conditionnelle est égale à l'espérance". Plus précisément :

Théorème 4.1.4 Si $g(X, Y)$ est intégrable, alors $E[g(X, Y)|Y]$ est intégrable. Si $g(X, Y)$ est soit non négative, soit intégrable,

$$E[E[g(X, Y)|Y]] = E[g(X, Y)] .$$

Démonstration.

$$\begin{aligned} E[|E[g(X, Y)|Y]|] &= E[|h(Y)|] \\ &= \int_{\mathbb{R}^m} |h(y)| f_Y(y) dy \\ &= \int_{\mathbb{R}^m} \left| \int_{\mathbb{R}^\ell} g(x, y) f_{X|Y}(x|y) dx \right| f_Y(y) dy \\ &\leq \int_{\mathbb{R}^m} \left(\int_{\mathbb{R}^\ell} |g(x, y)| f_{X|Y}(x|y) dx \right) f_Y(y) dy \\ &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^\ell} |g(x, y)| f_{X|Y}(x|y) f_Y(y) dx dy \\ &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^\ell} |g(x, y)| f_{X,Y}(x, y) dx dy \\ &= E[|g(X, Y)|] < \infty . \end{aligned}$$

Ceci prouve que si $g(X, Y)$ est intégrable, alors $E[g(X, Y)|Y]$ est intégrable. Un calcul similaire conduit à l'égalité annoncée. \square

Théorème 4.1.5 Si X et Y sont deux vecteurs aléatoires indépendants, alors pour toute fonction $g_1 : \mathbb{R}^\ell \rightarrow \mathbb{R}$ qui est soit non négative, soit telle que $g_1(X)$ est intégrable,

$$E[g_1(X)|Y] = E[g_1(X)] .$$

Démonstration. Les vecteurs X et Y étant indépendants, $f_{X,Y}(X, Y) = f_X(x)f_Y(y)$, et donc $f_{X|Y}(x|y) = f_X(x)$, ce qui conduit au résultat annoncé au vu de la formule (4.2). \square

Les fonctions de Y se conduisent vis-à-vis de l'espérance conditionnelle sachant Y comme des constantes. Plus précisément :

Théorème 4.1.6 Soit $v : \mathbb{R} \rightarrow \mathbb{R}$ une fonction. Supposons que l'une des deux conditions suivantes est satisfaite :

- (i) $g(X, Y)$ est intégrable et v est bornée, ou :
- (ii) g et v sont non négatives.

Alors

$$E[g(X, Y)v(Y)|Y] = E[g(X, Y)|Y]v(Y) .$$

En particulier, si $g_2 : \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction non négative ou telle que $g_2(Y)$ est intégrable,

$$E[g_2(Y)|Y] = g_2(Y) .$$

Démonstration. Par définition, $E[g(X, Y)v(Y)|Y] = u(Y)$, où

$$\begin{aligned} u(y) &= \int_{\mathbb{R}^\ell} g(x, y)v(y)f_{X|Y}(x|y)dx \\ &= v(y) \int_{\mathbb{R}^\ell} g(x, y)f_{X|Y}(x|y)dx = v(y)E[g(X, Y)|Y = y]. \end{aligned}$$

□

Considérons le cas où la dimension m du vecteur conditionnant Y est strictement supérieure à 1 et où ce vecteur admet une décomposition $Y = (Y_1, Y_2)$ où les dimensions de Y_1 et Y_2 sont m_1 et m_2 respectivement ($m_1 + m_2 = m$). On note alors

$$E[g(X, Y)|Y] = E[g(X, Y_1, Y_2)|Y_1, Y_2].$$

Si on prend l'espérance conditionnelle sachant Y_2 de cette dernière variable aléatoire, on obtient le même résultat que si on avait directement pris l'espérance conditionnelle de $g(X, Y) = g(X, Y_1, Y_2)$ sachant Y_2 . Plus précisément, voici la *formule des conditionnements successifs* :

Théorème 4.1.7 *Si $g(X, Y_1, Y_2)$ est soit intégrable, soit non négative, on a :*

$$E[E[g(X, Y_1, Y_2)|Y_1, Y_2]|Y_2] = E[g(X, Y_1, Y_2)|Y_2].$$

Démonstration. (Dans le cas densité-densité par exemple). On vérifie facilement, en utilisant la définition des densités de probabilité conditionnelle, que

$$f_{X|Y_1, Y_2}(x|y_1, y_2)f_{Y_1|Y_2}(y_1|y_2) = f_{X, Y_1|Y_2}(x, y_1|y_2).$$

On a alors, en notant $E[g(X, Y_1, Y_2)|Y_1, Y_2] = h(Y_1, Y_2)$,

$$\begin{aligned} &E[E[g(X, Y_1, Y_2)|Y_1, Y_2]|Y_2 = y_2] \\ &= E[h(Y_1, Y_2)|Y_2 = y_2] \\ &= \int_{\mathbb{R}^{m_1}} h(y_1, y_2)f_{Y_1|Y_2}(y_1|y_2)dy_1 \\ &= \int_{\mathbb{R}^{m_1}} \left(\int_{\mathbb{R}^\ell} g(x, y_1, y_2)f_{X|Y_1, Y_2}(x|y_1, y_2)dx \right) f_{Y_1|Y_2}(y_1|y_2)dy_1 \\ &= \int_{\mathbb{R}^{m_1}} \int_{\mathbb{R}^\ell} g(x, y_1, y_2)f_{X, Y_1|Y_2}(x, y_1|y_2) dx dy_1 \\ &= E[g(X, Y_1, Y_2)|Y_2 = y_2]. \end{aligned}$$

□

Régression non-linéaire

L'espérance conditionnelle est la solution d'un problème de minimisation. En effet, dans le cas densité-densité par exemple :

Théorème 4.1.8 *Si $g(X, Y)$ est de carré intégrable, la variable aléatoire $h(Y) = E[g(X, Y)|Y]$ est de carré intégrable et elle minimise l'écart quadratique $E[|v(Y) - g(X, Y)|^2]$ parmi toutes les fonctions $v : \mathbb{R}^m \rightarrow \mathbb{R}$ telles que $v(Y)$ soit de carré intégrable. Soit h' une autre fonction avec les mêmes propriétés que h . Alors $P(h'(Y) = h(Y)) = 1$.*

Démonstration. Fixons y . L'inégalité de Schwarz appliquée aux deux fonctions de x de carré intégrable $g(x, y) f_{X|Y}(x|y)^{1/2}$ et $f_{X|Y}(x|y)^{1/2}$ donne

$$\begin{aligned} |h(y)|^2 &= \left| \int_{\mathbb{R}^\ell} g(x, y) f_{X|Y}(x|y) dx \right|^2 \\ &\leq \left(\int_{\mathbb{R}^\ell} g(x, y)^2 f_{X|Y}(x|y) dx \right) \left(\int_{\mathbb{R}^\ell} f_{X|Y}(x|y) dx \right) \\ &= \int_{\mathbb{R}^\ell} g(x, y)^2 f_{X|Y}(x|y) dx , \end{aligned}$$

et donc

$$\begin{aligned} E[|h(Y)|^2] &= \int_{\mathbb{R}^m} |h(y)|^2 f_Y(y) dy \\ &\leq \int_{\mathbb{R}^m} \left(\int_{\mathbb{R}^\ell} g(x, y)^2 f_{X|Y}(x|y) dx \right) f_Y(y) dy \\ &= \int_{\mathbb{R}^\ell} \int_{\mathbb{R}^m} g(x, y)^2 f_{X,Y}(x, y) dx dy = E[g(X, Y)^2] < \infty . \end{aligned}$$

Ceci prouve que $h(Y) = E[g(X, Y)|Y]$ est de carré intégrable. Soit maintenant v une fonction telle que décrite dans l'énoncé du théorème. Observons tout d'abord que $g(X, Y)v(Y)$ est intégrable (en tant que produit de deux variables de carré intégrable) et que, de ce fait, la même preuve que celle du Théorème 4.1.6 conduit au résultat :

$$E[g(X, Y)v(Y)|Y] = E[g(X, Y)|Y]v(Y) ,$$

et donc (Théorème 4.1.4)

$$E[E[g(X, Y)|Y]v(Y)] = E[g(X, Y)v(Y)] .$$

On a donc, en notant $E[g(X, Y)|Y] = h(Y)$

$$\begin{aligned} E[|v(Y) - g(X, Y)|^2] &= E[v(Y)^2] + E[g(X, Y)^2] - 2E[v(Y)g(X, Y)] \\ &= E[v(Y)^2] + E[g(X, Y)^2] - 2E[v(Y)h(Y)] . \end{aligned}$$

Semblablement,

$$\begin{aligned} E[|h(Y) - g(X, Y)|^2] &= E[h(Y)^2] + E[g(X, Y)^2] - 2E[h(Y)h(Y)] \\ &= E[g(X, Y)^2] - E[h(Y)^2]. \end{aligned}$$

Il suffit donc de montrer que

$$E[v(Y)^2] - 2E[v(Y)h(Y)] \geq -E[h(Y)^2],$$

qui se réduit en effet à :

$$E[(v(Y) - h(Y))^2] \geq 0.$$

Pour démontrer l'unicité, on remplace dans les calculs précédents v par h' , et l'inégalité par l'égalité, ce qui donne

$$E[(h'(Y) - h(Y))^2] = 0,$$

et donc $P(h'(Y) - h(Y) = 0) = 1$ (Théorème 3.3.1). \square

4.2 Conditionnement des vecteurs gaussiens

Nous allons généraliser la formule de l'Exemple 4.1.2 donnant l'espérance conditionnelle de X_2 par rapport à X_1 lorsque (X_1, X_2) est gaussien. Il nous faut d'abord définir l'espérance conditionnelle d'un *vecteur* aléatoire X par rapport à un autre *vecteur* aléatoire Y . Par définition,

$$E[X|Y] = \begin{pmatrix} E[X_1|Y] \\ \vdots \\ E[X_n|Y] \end{pmatrix},$$

où X_1, \dots, X_n sont les composantes de X .

Soit X et Y deux vecteurs gaussiens dans leur ensemble, de moyennes m_X et m_Y , de matrices de covariance Σ_X et Σ_Y respectivement. La matrice d'inter-covariance de ces deux vecteurs est notée

$$\Sigma_{XY} = E[(X - m_X)(Y - m_Y)^T].$$

(On se rappellera au cours des calculs que cette matrice est symétrique.)

Théorème 4.2.1 *On suppose de plus que Y n'est pas dégénéré ($\Sigma_Y > 0$). L'espérance conditionnelle $E[X|Y]$ est alors donnée par*

$$E[X|Y] = m_X + \Sigma_{XY} \Sigma_Y^{-1} (Y - m_Y), \quad (4.3)$$

et la matrice de covariance de $\tilde{X} = X - E[X|Y]$ est

$$\Sigma_{\tilde{X}} = \Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX}. \quad (4.4)$$

D'autre part, \tilde{X} et Y sont des vecteurs aléatoires indépendants.

Démonstration. Considérons le vecteur

$$U = X - m_X - \Sigma_{XY}\Sigma_Y^{-1} (Y - m_Y).$$

Il est clair que $E[U] = 0$. En tant que combinaison linéaire de vecteurs X et Y gaussiens dans leur ensemble, le vecteur (U, Y) est gaussien. En particulier, si on montre que U et Y ne sont pas corrélés, ils sont indépendants (Théorème 3.4.3). Il en résultera, d'après le Théorème 4.1.5, que

$$E[U|Y] = E[U] = 0.$$

Comme $E[U|Y] = E[X|Y] - m_X - \Sigma_{XY}\Sigma_Y^{-1}(Y - m_Y)$, on a bien la formule (4.3) et $U = \tilde{X} - m_X$. D'autre part,

$$\begin{aligned} \Sigma_U &= E[UU^T] = E[(X - m_X - \Sigma_{XY}\Sigma_Y^{-1}(Y - m_Y))(X - m_X - \Sigma_{XY}\Sigma_Y^{-1}(Y - m_Y))^T] \\ &= E[(X - m_X)(X - m_X)^T] - E[(X - m_X)(Y - m_Y)^T \Sigma_Y^{-1} \Sigma_{YX}] \\ &\quad - E[\Sigma_{XY}\Sigma_Y^{-1}(Y - m_Y)(X - m_X)^T] \\ &\quad + E[\Sigma_{XY}\Sigma_Y^{-1}(Y - m_Y)(Y - m_Y)^T \Sigma_Y^{-1} \Sigma_{YX}], \end{aligned}$$

c'est-à-dire

$$\Sigma_U = \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}. \quad (4.5)$$

On a donc démontré la formule (4.4).

Reste à vérifier que U et Y ne sont pas corrélés :

$$\begin{aligned} E[U(Y - m_Y)^T] &= E[(X - m_X)(Y - m_Y)^T] - \Sigma_{XY}\Sigma_Y^{-1}E[(Y - m_Y)(Y - m_Y)^T] \\ &= \Sigma_{XY} - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_Y = 0. \end{aligned}$$

□

On a vu dans un chapitre précédent que la régression linéaire de X par rapport à Y est

$$\hat{X} = m_X + \Sigma_{XY}\Sigma_Y^{-1}(Y - m_Y).$$

On a donc, dans le cas gaussien, l'égalité de la régression linéaire et de l'espérance conditionnelle.

4.3 Tests d'hypothèses bayésiennes

États de la nature et observation

Dans le formalisme des *tests d'hypothèses* bayésiennes, les rôles principaux sont tenus par l'*état de la nature* Θ , une variable aléatoire prenant ses valeurs dans un ensemble fini $\{1, 2, \dots, K\}$, et l'*observation* X , un vecteur aléatoire réel de dimension m . On appelle

$\{1, 2, \dots, K\}$ l'espace des états de la nature, ou encore l'espace des hypothèses (c'est ce dernier terme que nous utiliserons dans le présent contexte).

Nous supposons que la loi du couple (Θ, X) est donnée par

$$P(\Theta = i, X \in A) = \mu(i) \int_A f(x|i) dx, \quad (4.6)$$

où $\mu = (\mu(1), \dots, \mu(K))$ est un vecteur de probabilité et $f(x|i)$ est, pour i fixé, une densité de probabilité sur \mathbb{R}^m . La distribution μ est celle de la variable Θ puisque

$$P(\Theta = i) = P(\Theta = i, X \in \mathbb{R}^m) = \mu(i) \int_{\mathbb{R}^m} f(x|i) dx = \mu(i).$$

D'autre part, $f(x|i)$ est la densité conditionnelle de l'observation X sachant que l'hypothèse i a lieu, puisque

$$P(X \in A | \Theta = i) = \frac{P(\Theta = i, X \in A)}{P(\Theta = i)} = \int_A f(x|i) dx.$$

On appelle parfois $\mu(i)$ la probabilité *a priori* (avant toute observation) de l'hypothèse i . La densité de probabilité de X est

$$f(x) = \sum_{i=1}^K \mu(i) f(x|i).$$

En effet,

$$P(X \in A) = \sum_{i=1}^K P(\Theta = i, X \in A) = \int_A \left\{ \sum_{i=1}^K \mu(i) f(x|i) \right\} dx.$$

La probabilité *a posteriori* de l'hypothèse i après observation de la valeur $X = x$ de l'observation est la fonction de i et de x :

$$\mu(i|x) = \frac{\mu(i) f(x|i)}{f(x)}. \quad (4.7)$$

EXEMPLE 4.3.1: SIGNAL PLUS BRUIT. Supposons que $f(x|i)$ ait la forme

$$f(x|i) = g(x - S(i)),$$

où $g(x)$ est une densité de probabilité sur \mathbb{R}^m et $S(i) = (S_1(i), \dots, S_m(i))$ est un vecteur déterministe de \mathbb{R}^m . Tout se passe comme si l'observation X était la somme

$$X = S(\Theta) + B, \quad (4.8)$$

où $B = (B_1, \dots, B_m)$ est un vecteur aléatoire de densité de probabilité $g(x)$ indépendant de Θ . En effet, dans ce cas, en notant $A - S(i) := \{x \in \mathbb{R}^m; x + S(i) \in A\}$,

$$\begin{aligned} P(X \in A | \Theta = i) &= P(B \in A - S(i) | \Theta = i) \\ &= P(B \in A - S(i)) \\ &= \int_{A - S(i)} g(x) dx \\ &= \int_A g(x - S(i)) dx. \end{aligned}$$

EXEMPLE 4.3.2: COMMUNICATION EN PRÉSENCE DE BRUIT ADDITIF, TAKE 1. Voici une interprétation possible de l'égalité (4.8). Les vecteurs X et B proviennent de l'échantillonnage de fonctions aléatoires (famille de variables aléatoires indexées par le temps) $\{X(t)\}_{t \in [0, T]}$ et $\{B(t)\}_{t \in [0, T]}$, et le vecteur $S(\Theta)$ provient de l'échantillonnage d'une fonction aléatoire $\{S(\Theta, t)\}_{t \in [0, T]}$, où $\{S(i, t)\}_{t \in [0, T]}$ est pour chaque "message" $i \in \{1, 2, \dots, K\}$ une fonction déterministe. Plus précisément :

$$\begin{aligned} X_j &= X(j\tau) \\ B_j &= B(j\tau) \\ S_j(\Theta) &= S(\Theta, j\tau), \end{aligned}$$

où τ est la période d'échantillonnage (ou encore $f = \frac{1}{\tau}$ est la fréquence d'échantillonnage, exprimée en *hertz* si la période est exprimée en secondes) et $T = m\tau$ est le temps d'observation. Un *message* aléatoire $\Theta \in \{1, \dots, K\}$ engendré par une *source* (qui produit i avec la probabilité $\mu(i)$) doit être transmis à travers un *canal physique*, disons, pour fixer les idées, un faisceau hertzien. Ce message abstrait ne peut pas être transmis tel quel dans un canal physique, il faut en faire une *modulation* $\{S(i, t)\}_{t \in [0, T]}$ (par exemple, $S(t, i) = \cos(2\pi f_i t)$ si on emploie une modulation de fréquence). C'est cette modulation qui est envoyée dans les airs sous forme d'onde électromagnétique. A l'autre extrémité du système de communication parvient une version de $\{S(\Theta, t)\}_{t \in [0, T]}$ perturbée additivement par un bruit $\{B(t)\}_{t \in [0, T]}$ (bruit du récepteur, parasitage intentionnel ou non, etc.) :

$$X(t) = S(\Theta, t) + B(t) .$$

L'observation $(X(t), t \in [0, T])$ est alors échantillonnée, et sur la base du vecteur aléatoire obtenu, on doit décider quel message i a été envoyé. On retrouve ainsi le modèle général de l'exemple précédent.

Probabilité d'erreur et région de décision

Soit $\mathcal{A} = (A_1, \dots, A_K)$ une K -partition de l'espace \mathbb{R}^m . On associe à cette partition la *règle de décision*

$$x \in A_i \rightarrow \text{hypothèse } i ,$$

ce qui veut dire : si la valeur expérimentale x de X est dans la région A_i , on décide que Θ a la valeur i (l'hypothèse i a lieu). Cette règle de décision n'est pas infaillible, et la probabilité d'erreur correspondante est donnée par :

$$P_E(\mathcal{A}) = \sum_{i=1}^K P(X \notin A_i, \Theta = i) .$$

On a donc

$$P_E(\mathcal{A}) = \sum_{i=1}^K \mu(i) \int_{\bar{A}_i} f(x|i) dx .$$

Le problème mathématique qui se pose maintenant est de trouver la K -partition \mathcal{A}^* optimale en ce sens que

$$P_E(\mathcal{A}^*) \leq P_E(\mathcal{A})$$

pour toute K -partition \mathcal{A} . La solution est très simple. Il suffit de réécrire $P_E(\mathcal{A})$ en observant que $\int_{\bar{A}_i} f(x|i) dx = 1 - \int_{A_i} f(x|i)$ et en utilisant l'égalité $\sum_{i=1}^K \mu(i) = 1$:

$$P_E(\mathcal{A}) = 1 - \sum_{i=1}^K \mu(i) \int_{A_i} f(x|i) dx$$

($\sum_{i=1}^K \mu(i) \int_{A_i} f(x|i) dx$ est la probabilité de tomber juste), ou encore

$$P_E(\mathcal{A}) = 1 - \int_{\mathbb{R}^m} \left(\sum_{i=1}^K \mu(i) f(x|i) 1_{A_i}(x) \right) dx .$$

Le résultat suivant découle immédiatement de cette dernière expression de la probabilité d'erreur.

Théorème 4.3.1 *Toute partition \mathcal{A}^* telle que*

$$x \in A_i^* \Rightarrow \mu(i) f(x|i) = \max_k \mu(k) f(x|k)$$

minimise la probabilité d'erreur. La décision optimale est alors

$$\hat{\Theta} = \operatorname{argmin}_i \mu(i) f(X|i) .$$

N'importe quelle façon de choisir l'hypothèse i qui atteint le maximum de probabilité *a posteriori* $\mu(k|x)\mu(k)f(x|k)$ minimise la probabilité d'erreur. Il peut y avoir en effet plusieurs indices assurant le maximum.

EXEMPLE 4.3.3: COMMUNICATION EN PRÉSENCE DE BRUIT ADDITIF, TAKE 2. (suite de l'Exemple 4.3.2) On suppose que le bruit échantillonné B est un m -vecteur gaussien de moyenne nulle et de covariance Σ strictement positive :

$$g(x) = \frac{1}{(2\pi)^{m/2}(\det \Sigma)^{1/2}} e^{-\frac{1}{2}x^T \Sigma^{-1}x},$$

et donc, pour toute hypothèse i ,

$$f(x|i) = \frac{1}{(2\pi)^{m/2}(\det \Sigma)^{1/2}} e^{-\frac{1}{2}(x-S(i))^T \Sigma^{-1}(x-S(i))}.$$

Le test bayésien minimisant la probabilité d'erreur compare les quantités $\mu(i)f(X|i)$ ou, ce qui revient au même, les logarithmes de ces quantités. En supprimant dans ces logarithmes les termes ne dépendant pas de l'hypothèse i , on arrive au test optimal. Celui-ci consiste à choisir l'hypothèse i qui maximise

$$\log \mu(i) - \frac{1}{2}S(i)^T \Sigma^{-1}S(i) + X^T \Sigma^{-1}S(i).$$

Supposons maintenant que le bruit B est un bruit “blanc” gaussien, c'est-à-dire un vecteur gaussien standard : $\Sigma = I$. Aussi, pour simplifier, supposons que les hypothèses sont *a priori* équiprobables. Alors, le test optimal consiste à choisir i qui maximise

$$X^T S(i) - \frac{1}{2}S^T(i)S(i),$$

ou encore, de manière équivalente,

$$\|X - S(i)\|^2.$$

On choisit donc le vecteur moyenne le plus proche de l'observation X .

Observations discrètes

La théorie ci-dessus, qui a été détaillée pour une observation admettant une densité de probabilité, s'applique, *mutatis mutandis*, à une observation prenant ses valeurs dans un ensemble \mathcal{X} dénombrable. Il suffit de remplacer là où il le faut $f(x|i)$ par $p(x|i)$ où

$$p(x|i) = P(X = x|\Theta = i)$$

et les intégrales par des sommes. La stratégie optimale est

$$\hat{\Theta} = \operatorname{argmin}_i \mu(i)p(X|i).$$

EXEMPLE 4.3.4: HYPOTHÈSES POISSONNIENNES. L'observation est une variable aléatoire de Poisson X , et sous l'hypothèse i , celle-ci a pour moyenne $\lambda_i > 0$:

$$p(x|i) = e^{-\lambda_i} \frac{(\lambda_i)^x}{x!} \quad (x \in \mathbb{N}).$$

Si on suppose de plus les k hypothèses a priori équiprobables, la stratégie optimale est celle qui consiste à choisir l'hypothèse i qui maximise $\mu(i)e^{-\lambda_i} \frac{\lambda_i^X}{X!} = \frac{1}{k} e^{-\lambda_i} \frac{\lambda_i^X}{X!}$, ou encore (en passant aux logarithmes népériens, et en supprimant les termes qui ne dépendent pas de i), celle qui maximise

$$(-\lambda_i + X \log \lambda_i) .$$

Le test de Neyman–Pearson.

C'est l'existence de probabilités *a priori* des hypothèses qui vaut à la théorie des tests d'hypothèses du paragraphe précédent l'appellation "bayésienne". Ces probabilités peuvent être "objectives" ou "subjectives". Nous n'aborderons pas les problèmes philosophiques qui se cachent derrière cette terminologie. Disons simplement que si le test est utilisé un grand nombre de fois, comme c'est souvent le cas dans les systèmes de communications (Exemples 4.3.2 et 4.3.3), le modélisateur dispose de statistiques sur la source de messages. On est alors dans une situation bayésienne. Dans d'autres cas, le modélisateur ne dispose pas de données suffisantes sur la fréquence des hypothèses, ou bien le test ne servira qu'une seule fois, ou bien encore, l'environnement n'est pas stable et les probabilités *a priori* des hypothèses varient d'une utilisation à l'autre. Une façon de s'en sortir est d'admettre son ignorance en adoptant la distribution *a priori* uniforme sur les hypothèses. Les doutes qui peuvent subsister quant au bien fondé d'un choix particulier de distribution *a priori* des hypothèses sont levés si la stratégie optimale dépend "peu" de ce choix. En d'autres termes, le test est "robuste" (la robustesse est un vaste sujet de la statistique que nous n'aborderons pas).

Il existe des situations où il faut absolument sortir du cadre bayésien. C'est le cas par exemple en détection (disons la détection radar pour fixer les idées) où il y a deux hypothèses : y a-t-il ou non dans le domaine aérien couvert par le radar un objet volant ? Le principe de symétrie dans l'ignorance semble imposer la distribution *a priori* uniforme, c'est-à-dire une probabilité $\frac{1}{2}$ pour chacune des deux hypothèses, mais les radaristes préfèrent adopter une autre approche, et un autre critère que le critère de la probabilité d'erreur globale. C'est le *critère de Neyman et Pearson* que nous allons décrire.

On a une observation $X \in \mathbb{R}^m$ qui admet sous l'hypothèse $i \in \{0, 1\}$ la densité de probabilité $f(x|i)$. On note A_0 et A_1 les domaines où l'on décide 0 et 1 respectivement. Bien entendu $\{A_0, A_1\}$ est une partition de \mathbb{R}^m . L'*erreur de première espèce* consiste à supposer que c'est l'hypothèse 0 qui a lieu alors que c'est en fait l'hypothèse 1. L'erreur symétrique (on dit 1 alors que c'est 0) est appelée *erreur de seconde espèce*. Les probabilités d'erreur de première espèce et de deuxième espèce, $P_E(1)$ et $P_E(2)$ respectivement, admettent les expressions

$$P_E(1) = \int_{A_0} f(x|1) dx, \quad P_E(2) = \int_{A_1} f(x|0) dx .$$

Le critère de Neyman-Pearson consiste à se fixer un seuil $\alpha \in (0, 1)$ et à minimiser la probabilité d'erreur de première espèce sous la contrainte que celle de seconde espèce reste inférieure au seuil α :

$$\min \left\{ \int_{A_0} f(x|1) dx ; \int_{A_1} f(x|0) dx \leq \alpha \right\}.$$

EXEMPLE 4.3.5: LE RADAR, TAKE 1. Ici

$$X = \theta S + B, \quad (4.9)$$

où B est un vecteur aléatoire de densité de probabilité g , indépendant de l'hypothèse $\theta \in \{0, 1\}$, et S est un vecteur déterministe. Les vecteurs S et B peuvent provenir de l'échantillonnage d'une fonction déterministe $S(t)$ et d'une fonction aléatoire $B(t)$, appelées respectivement le signal utile et le bruit. Le signal $S(t)$ prend, dans le cas d'un radar, la forme d'une brève impulsion, le "bip", émise par l'antenne et éventuellement réfléchi sur une cible. Nous ignorerons le problème de l'atténuation, ou plutôt, nous considérerons que $S(t)$ est en fait le signal retour. S'il n'y a pas de cible (hypothèse 0), le signal observé $X(t)$ ne contient que le bruit $B(t)$:

$$X(t) = B(t).$$

S'il y a une cible (hypothèse 1),

$$X(t) = S(t) + B(t).$$

C'est bien ce que traduit (4.9) en termes de signaux échantillonnés. Il faut cependant noter que θ n'est pas une variable aléatoire, comme dans la situation bayésienne, mais un simple paramètre prenant les valeurs 0 ou 1 : $\theta = 0 \equiv$ hypothèse H_0 , $\theta = 1 \equiv$ hypothèse H_1 . On a, dans cet exemple

$$f(x|0) = g(x) \quad , \quad f(x|1) = g(x - S).$$

Dans le cas du radar $P_E(1)$ est la probabilité de *non-détection* puisqu'on décide qu'il n'y a pas de cible alors qu'il y en a une, tandis que $P_E(2)$ est la probabilité de *fausse alarme*. Pour éviter la saturation des ordinateurs de gestion du trafic aérien, on se fixe un taux de fausse alarme à ne pas dépasser, et on cherche évidemment à maximiser, sous cette contrainte, la probabilité de détection.

Nous allons donner la stratégie optimale. Pour cela, on introduit la fonction *rapport de vraisemblance* de l'hypothèse 1 contre l'hypothèse 0, défini par :

$$L(x) = \frac{f(x|1)}{f(x|0)}.$$

On fera l'hypothèse

$$f(x|0) = 0 \Rightarrow f(x|1) = 0$$

et la convention $\frac{0}{0} = 1$, ce qui suffit à donner un sens à $L(x)$ pour tout x .

Théorème 4.3.2 Le lemme de Neyman-Pearson. *Tout ensemble $A_0^* \subset \mathbb{R}^n$ de la forme*

$$A_0^* = \{x | L(x) \leq \sigma\} , \quad (4.10)$$

où σ est un nombre non négatif choisi de telle manière que

$$\int_{A_0^*} f(x|0) dx = \alpha , \quad (4.11)$$

est optimal, en ce sens que pour tout autre ensemble $A_0 \subset \mathbb{R}^n$ tel que

$$\int_{A_0} f(x|0) dx \leq \alpha ,$$

on a

$$\int_{A_1^*} f(x|1) dx \geq \int_{A_1} f(x|1) dx .$$

Démonstration. Observons que

$$A_0^* = A_0^* \cap \overline{A_0} + A_0^* \cap A_0, \quad A_0 = A_0 \cap \overline{A_0^*} + A_0 \cap A_0^*$$

et que par définition

$$\begin{aligned} x \in A_0^* &\Leftrightarrow f(x|0) \geq \frac{1}{\sigma} f(x|1) \\ x \in \overline{A_0^*} &\Leftrightarrow f(x|0) < \frac{1}{\sigma} f(x|1) . \end{aligned}$$

On a donc, avec des simplifications évidentes dans les notations

$$\begin{aligned} \int_{A_0^*} f_1 - \int_{A_0} f_1 &= \int_{A_0^* \cap \overline{A_0}} f_1 - \int_{A_0 \cap \overline{A_0^*}} f_1 \\ &\leq \sigma \left(\int_{A_0^* \cap \overline{A_0}} f_0 - \int_{A_0 \cap \overline{A_0^*}} f_0 \right) \\ &= \sigma \left(\int_{A_0^*} f_0 - \int_{A_0} f_0 \right) \leq 0 , \end{aligned}$$

puisque $\int_{A_0} f_0 \geq 1 - \alpha$ et $\int_{A_0^*} f_0 = 1 - \alpha$. □

EXEMPLE 4.3.6: LE RADAR, TAKE 2. Voyons comment tout cela s'applique à la situation correspondant à deux hypothèses gaussiennes ne différant que par la moyenne :

$$f(x|0) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} , \quad f(x|1) = \frac{1}{\sqrt{2\pi}} e^{-(x-m)^2/2} .$$

Le rapport de vraisemblance est dans ce cas

$$L(x) = e^{-m^2/2} e^{+mx} .$$

On voit que $L(x) \leq \sigma$ est défini par

$$x \leq \left(\log \sigma + \frac{m^2}{2} \right) / m ,$$

et donc σ est déterminé de façon unique par :

$$\frac{1}{\sqrt{2\pi}} \int_{(\log \sigma + \frac{m^2}{2})/m}^{+\infty} e^{-1/2 x^2} dx = \alpha .$$

4.4 Exercices

Exercice 4.4.1.

Soit (X, Y) un vecteur aléatoire de \mathbb{R}^2 distribué uniformément dans le disque fermé de rayon unité et de centre O. Calculez la densité de probabilité conditionnelle de X étant donné Y .

Exercice 4.4.2.

Soit X_1, \dots, X_n des variables aléatoires indépendantes et identiquement distribuées, de densité de probabilité commune $f(x)$. On pose :

$$Y = \max(X_1, \dots, X_n), \quad Z = \min(X_1, \dots, X_n) .$$

Calculez la densité de probabilité conditionnelle $f_{Z|Y}(z|y)$ et l'espérance conditionnelle $E[Z|Y]$. Terminez les calculs dans le cas où chaque X_i est uniformément distribuée sur $[0, 1]$.

Exercice 4.4.3. SOLEIL, LÈVE-TOI !

A l'origine des temps, un nombre p a été choisi au hasard entre 0 et 1, sans que cette valeur ait été révélée à l'humanité. Depuis, le soleil se lève chaque jour avec la probabilité (inconnue) p . Les jours précédents n'influent pas sur le lever du soleil du jour présent. Sachant que depuis l'origine des temps le soleil s'est levé tous les jours, soit N fois (pour le chiffre exact, consultez votre gourou), quelle est la probabilité pour qu'il se lève demain ?

Exercice 4.4.4.

Donnez la densité de probabilité conditionnelle de Y étant donné $X = x$ dans le cas où (X, Y) admet la densité de probabilité.

$$f_{X,Y}(x, y) = \frac{x}{\sqrt{2\pi}} e^{-\frac{1}{2}x(y+x)} 1_{\{x \geq 0, y \geq 0\}} .$$

Exercice 4.4.5.

Soit X_1 et X_2 deux variables aléatoires de Poisson indépendantes, de moyennes $\lambda_1 > 0$ et $\lambda_2 > 0$. Montrez que X_1 suit, conditionnellement à $X_1 + X_2 = n$, une loi binomiale d'ordre n .

Exercice 4.4.6. INFORMATION ÉQUIVALENTE.

Soit X une variable aléatoire de carré intégrable, et soit Y_1 et Y_2 deux vecteurs aléatoires. On suppose qu'il existe deux fonctions f et g telles que

$$Y_1 = f(Y_2) \quad , \quad Y_2 = g(Y_1) .$$

Cette hypothèse exprime que Y_1 et Y_2 contiennent la même information. Il est donc naturel que le conditionnement par l'une soit équivalent au conditionnement par l'autre. On demande de démontrer qu'il en est bien ainsi, c'est-à-dire :

$$E[X|Y_1] = E[X|Y_2] .$$

(Le résultat est vrai aussi lorsque X est une variable aléatoire non négative ou intégrable.)

Exercice 4.4.7.

Soit X , U et V trois vecteurs gaussiens dans leur ensemble. On suppose que U et V sont non-corrélés et que Σ_U et Σ_V sont des matrices strictement positives. Montrez que

$$E[X|U, V] = E[X|U] + E[X|V] - m_X .$$

Exercice 4.4.8. DÉCODEUR OPTIMAL.

On dispose d'une observation aléatoire $X \in \{0,1\}^m$ et de K hypothèses H_i ($1 \leq i \leq K$). Sous l'hypothèse H_i , l'observation a la forme $X = v_i \oplus B$, où v_i est un vecteur déterministe de $\{0,1\}^m$ et B est un "bruit blanc informatique", c'est-à-dire, $B = (B_1, \dots, B_m)$ est une suite IID de variables aléatoires à valeurs dans $\{0,1\}$ avec $P(B_i = 1) = p$, $0 < p < \frac{1}{2}$. Ici, \oplus est l'addition modulo 2 ($1 \oplus 1 = 0$) composante par composante. Les K hypothèses sont supposées équiprobables.

Montrez que la stratégie de décision minimisant la probabilité d'erreur est la suivante : décider pour l'hypothèse H_i qui minimise la distance de Hamming entre l'observation X et le vecteur v_i , où on appelle distance de Hamming entre deux vecteurs de $\{0,1\}^m$ le nombre de composantes par lesquelles ils diffèrent.

Exercice 4.4.9. *

Soit $\{A_n\}_{1 \leq n \leq N}$ et $\{B_n\}_{1 \leq n \leq N}$ deux suites IID indépendantes de variables gaussiennes de moyennes μ_A et μ_B et de variances σ_A^2 et σ_B^2 . Soit $X = \{X_n\}_{1 \leq n \leq N}$ un vecteur de la forme

$$X_n = \Theta_n A_n + (1 - \Theta_n) B_n$$

où $\Theta = \{\Theta_n\}_{1 \leq n \leq N}$ est une suite IID de variables à valeurs équiprobables dans $\{0, 1\}$, et indépendante de $\{A_n\}_{1 \leq n \leq N}$ et $\{B_n\}_{1 \leq n \leq N}$. Trouvez la valeur de estimée $\hat{\Theta} \in \{0, 1\}^n$ de Θ en fonction de l'observation X qui minimise la probabilité d'erreur.

Exercice 4.4.10. * D'OÙ VIENT-IL ?

Un point $Y \in \mathbb{R}^2$ est choisi au hasard (distribution uniforme) dans le carré $[-1, +1] \times [-1, +1]$. Au temps 0 un mobile est placé en ce point, et dès lors, il se meut dans une direction aléatoire, l'angle $\Theta \in [0, 2\pi]$ caractérisant cette direction ayant la densité de probabilité $f(\theta) = \frac{\theta}{2\pi}$. Sa vitesse aléatoire dans cette direction est une variable exponentielle, $V \sim \mathcal{E}(\lambda)$, de moyenne λ^{-1} . Les variables X , Θ , V sont indépendantes. Soit $X \in \mathbb{R}^2$ sa position au bout d'un temps unité. Calculez l'estimée $E[Y | X]$ de la position de départ.

Chapitre 5

Information et entropie

5.1 L'inégalité de Gibbs

Dans ses travaux sur le codage et la transmission de données dans un canal bruité, l'ingénieur américain Claude Shannon a introduit le concept de *quantité d'information* contenue dans une variable aléatoire ou associée à une distribution de probabilité. Il a mis en évidence le rôle fondamental qu'elle joue dans les communications, en particulier pour la compression des données (que nous aborderons dans ce chapitre) et pour la correction des erreurs de transmission.

Définition 5.1.1 Soit X une variable aléatoire à valeurs dans un ensemble fini \mathcal{X} , de distribution de probabilité $(p(x), x \in \mathcal{X})$. Son entropie est, par définition, la quantité

$$H(X) = -E[\log(p(X))] = -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$

avec la convention $0 \log 0 = 0$.

La base du logarithme doit être choisie une fois pour toutes. En base D , on notera $H(X) = H_D(X)$. En base 2, on exprimera l'entropie en *bits*, et en *nats* en base e .

EXEMPLE 5.1.1: $\mathcal{X} = \{0, 1\}$, $P(X = 1) = p$. On a alors $H_2(X) = h_2(p)$, où

$$h_2(p) = -p \log_2 p - (1 - p) \log_2 (1 - p).$$

Cette fonction est concave et son maximum ($=1$) est atteint pour $p = \frac{1}{2}$.

L'inégalité (dite *de Gibbs*) du théorème suivant joue un grand rôle en théorie de l'information.

Théorème 5.1.1 *Soit $(p(x), x \in \mathcal{X})$ et $(q(x), x \in \mathcal{X})$ deux distributions de probabilité sur \mathcal{X} . On a l'inégalité*

$$-\sum_{x \in \mathcal{X}} p(x) \log p(x) \leq -\sum_{x \in \mathcal{X}} p(x) \log q(x),$$

avec égalité si et seulement si $p(x) = q(x)$ pour tout $x \in \mathcal{X}$.

Démonstration. On peut supposer que $q(x) > 0$ pour tout x tel que $p(x) > 0$ (sinon, l'inégalité est triviale, le deuxième membre étant alors infini). On se ramène donc au cas où \mathcal{X} ne contient que des x tels que $p(x) > 0$, et où q est une sous-probabilité ($\sum_{x \in \mathcal{X}} q(x) \leq 1$) telle que $q(x) > 0$ pour tout x . En utilisant le fait que $\log z \leq z - 1$ pour tout $z > 0$, avec égalité si et seulement si $z = 1$, on a

$$\begin{aligned} \sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} &\leq \sum_{x \in \mathcal{X}} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \\ &= \sum_{x \in \mathcal{X}} q(x) - \sum_{x \in \mathcal{X}} p(x) \leq 0. \end{aligned}$$

L'égalité a lieu seulement lorsque $\frac{q(x)}{p(x)} = 1$ pour tout $x \in \mathcal{X}$. □

Théorème 5.1.2 *Soit X une variable aléatoire discrète à valeurs dans l'ensemble fini \mathcal{X} . Alors, notant $|\mathcal{X}|$ le nombre d'éléments dans \mathcal{X} :*

$$0 \leq H(X) \leq \log |\mathcal{X}|.$$

On a $H(X) = 0$ si et seulement si X est déterministe, et $H(X) = \log |\mathcal{X}|$ si et seulement si X est uniformément distribuée sur \mathcal{X} .

En particulier si $D = |\mathcal{X}|$, et si le logarithme est à base D ,

$$H_D(X) \leq 1.$$

Démonstration. L'inégalité de gauche est évidente. Celle de droite découle de l'inégalité de Gibbs avec $q(x) = \frac{1}{|\mathcal{X}|}$.

La valeur 0 n'est possible que si pour chaque $x \in \mathcal{X}$, $p(x) \log p(x) = 0$, c'est-à-dire si $p(x) = 0$ ou $p(x) = 1$. Comme la somme des $p(x)$ doit être égale à 1, il y a un et un seul $x \in \mathcal{X}$ pour lequel $p(x) = 1$, et donc X est déterministe.

L'égalité $H(X) = \log |\mathcal{X}|$ est équivalente à

$$-\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\sum_{x \in \mathcal{X}} p(x) \log \left(\frac{1}{|\mathcal{X}|} \right),$$

ce qui n'est possible (d'après le Théorème 8.45 appliqué avec $q(x) = \frac{1}{|\mathcal{X}|}$) que si $p(x) = \frac{1}{|\mathcal{X}|}$ pour tout $x \in \mathcal{X}$. \square

Théorème 5.1.3 Soient X_1, \dots, X_n des variables aléatoires indépendantes, respectivement à valeurs dans les ensembles finis $\mathcal{X}_1, \dots, \mathcal{X}_n$, et d'entropies $H(X_1), \dots, H(X_n)$, alors :

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i).$$

En particulier, si X_1, \dots, X_n sont IID,

$$H(X_1, \dots, X_n) = nH(X_1).$$

Démonstration. On note $p_i(x_i) = P(X_i = x_i)$ pour tout $x_i \in \mathcal{X}_i$. L'hypothèse d'indépendance se lit

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n p_i(x_i),$$

et par conséquent

$$\begin{aligned} H(X_1, \dots, X_n) &= -E \left[\log \prod_{i=1}^n p_i(X_i) \right] \\ &= -E \left[\sum_{i=1}^n \log p_i(X_i) \right] \\ &= -\sum_{i=1}^n E [\log p_i(X_i)] \\ &= \sum_{i=1}^n H(X_i). \end{aligned}$$

\square

5.2 Suites typiques et compression des données

Soit \mathcal{X} un ensemble fini de cardinal D , et soient X_1, \dots, X_n des variables aléatoires IID à valeurs dans \mathcal{X} , de distribution commune $(p(x), x \in \mathcal{X})$ et d'entropie commune

$$H_D = E [-\log_D p(X_1)].$$

Le vecteur $X^{(n)} = (X_1, \dots, X_n)$ a pour distribution

$$p(x^{(n)}) = p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i).$$

Théorème 5.2.1 *Pour $\varepsilon > 0$, on définit l'ensemble ε -typique d'ordre n*

$$A_\varepsilon^{(n)} = \left\{ x^{(n)} ; \left| -\frac{1}{n} \sum_{i=1}^n \log_D p(x_i) - H_D \right| \leq \varepsilon \right\}.$$

On a

$$\lim_{n \rightarrow \infty} P(X^{(n)} \in A_\varepsilon^{(n)}) = 1$$

et

$$|A_\varepsilon^{(n)}| \leq D^{n(H_D + \varepsilon)}.$$

Démonstration.

$$P(X^{(n)} \in A_\varepsilon^{(n)}) = P\left(\left| -\frac{1}{n} \sum_{i=1}^n \log_D p(X_i) - H_D \right| \leq \varepsilon\right).$$

La loi faible des grands nombres donne

$$-\frac{1}{n} \sum_{i=1}^n \log_D p(x_i) \xrightarrow{Pr} -E[\log_D p(X_1)] = H_D,$$

c'est-à-dire : pour tout $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P\left(\left| -\frac{1}{n} \sum_{i=1}^n \log_D p(x_i) - H_D \right| > \varepsilon\right) = 0,$$

ce qui est le premier résultat annoncé. Par définition de $A_\varepsilon^{(n)}$, si $x^{(n)}$ appartient à cet ensemble,

$$D^{-n(H_D + \varepsilon)} \leq p(x^{(n)}),$$

et par conséquent,

$$P(X^{(n)} \in A_\varepsilon^{(n)}) = \sum_{x^{(n)} \in A_\varepsilon^{(n)}} p(x^{(n)}) \geq D^{-n(H_D + \varepsilon)} |A_\varepsilon^{(n)}|.$$

Étant donné que $P(X^{(n)} \in A_\varepsilon^{(n)}) \leq 1$, on a donc

$$1 \geq |A_\varepsilon^{(n)}| D^{-n(H_D + \varepsilon)}.$$

□

Compression de données

On peut utiliser le résultat du Théorème 5.2.1 pour compresser des données. Pour ce faire, construisons une application $c : \mathcal{X}^n \rightarrow \mathcal{X}^{\lceil n(H_D + \varepsilon) \rceil + 1}$ de la façon suivante. On se

donne d'abord une application $\tilde{c}^{(n)} : A_\varepsilon^{(n)} \rightarrow \mathcal{X}^{\lceil n(H_D + \varepsilon) \rceil}$. L'inégalité du Théorème 5.2.1 garantit qu'on peut choisir cette application *injective*. On définit ensuite :

$$c^{(n)}(x^{(n)}) = \begin{cases} \tilde{c}^{(n)}(x^{(n)})1 & \text{si } x^{(n)} \in A_\varepsilon^{(n)} \\ 0\dots 0 & \text{si } x^{(n)} \notin A_\varepsilon^{(n)}. \end{cases}$$

La restriction de cette application à $A_\varepsilon^{(n)}$ est injective. Une erreur n'est possible que si $X^{(n)} \notin A_\varepsilon^{(n)}$, et la probabilité de cet événement tend vers 0 en l'infini. Le "décodage" est donc possible avec une erreur aussi petite que l'on souhaite en prenant un n assez grand. Pour une suite de symboles de \mathcal{X} de longueur n , on a maintenant besoin non plus de n symboles de l'alphabet, mais (asymptotiquement) de $(\lceil n(H_D + \varepsilon) \rceil + 1)$, ce qui correspond à un *taux de compression* de

$$\frac{(\lceil n(H_D + \varepsilon) \rceil + 1)}{n}.$$

En choisissant n assez grand et ε assez petit, on peut rendre ce rapport aussi proche que souhaité de H_D . Ce rapport est bien entendu inférieur ou égal à 1 comme le montre l'inégalité fondamentale $H_D \leq 1$ (Théorème 5.1.2).

Le résultat suivant montre qu'on ne peut pas faire mieux.

Théorème 5.2.2 *Supposons qu'il existe $B^{(n)} \subset \mathcal{X}^n$ et un nombre $R > 0$ tels que :*

$$\lim_{n \rightarrow \infty} P(X^{(n)} \in B^{(n)}) = 1$$

$$|B^{(n)}| \leq D^{nR}$$

Alors nécessairement : $R \geq H_D$

Démonstration. Lorsque $x^{(n)} \in A_\varepsilon^{(n)}$, $p(x^{(n)}) \leq D^{-n(H_D - \varepsilon)}$, et donc :

$$\begin{aligned} P(X^{(n)} \in A_\varepsilon^{(n)} \cap B^{(n)}) &\leq D^{-n(H_D - \varepsilon)} |A_\varepsilon^{(n)} \cap B^{(n)}| \\ &\leq D^{-n(H_D - \varepsilon)} |B^{(n)}| \leq D^{-n(H_D - \varepsilon - R)}. \end{aligned}$$

Or par hypothèse $\lim_{n \rightarrow \infty} P(X^{(n)} \in B^{(n)}) = 1$ et (Théorème 5.2.1) $\lim_{n \rightarrow \infty} P(X^{(n)} \in A_\varepsilon^{(n)}) = 1$, d'où :

$$\lim_{n \rightarrow \infty} P(X^{(n)} \in A_\varepsilon^{(n)} \cap B^{(n)}) = 1.$$

Par conséquent :

$$\lim_{n \rightarrow \infty} D^{-n(H_D - \varepsilon - R)} \geq 1$$

ce qui implique que $H_D - \varepsilon - R \leq 0$, c'est-à-dire, $R \geq H_D - \varepsilon$. Comme ε est arbitraire, on a le résultat annoncé. \square

5.3 Codage de source

Dans le procédé de codage qui vient d'être décrit, on se résigne à perdre les suites non typiques. Le dommage semblait mineur parce que, asymptotiquement, la probabilité qu'on rencontre une suite non typique est nulle, les suites non typiques étant par essence rares. Mais il y a des applications où l'on souhaite récupérer intactes *toutes* les suites à l'issue de la compression, y compris et surtout celles qui sont rares (applications bancaires ou militaires par exemple). Le présent paragraphe traite de la *compression de données sans perte d'information*.

Inégalité de Kraft

Pour tout ensemble \mathcal{A} , on note \mathcal{A}^* la collection des chaînes finies d'éléments de \mathcal{A} y compris la chaîne vide \emptyset . Par exemple, si $\mathcal{A} = \{0, 1\}$, les chaînes $y_1 = 0110$, $y_2 = 111$ et $y_3 = 0101010$ sont dans $\{0, 1\}^*$. Concaténer des chaînes de \mathcal{A}^* veut dire qu'on les met bout à bout pour former une chaîne de \mathcal{A}^* . Dans l'exemple, 01101110101010 est obtenue en concaténant y_1 , y_2 et y_3 (on note cette chaîne $y_1 * y_2 * y_3$ ou, plus simplement $y_1 y_2 y_3$), ou bien en concaténant y_1 , \emptyset , y_2 et y_3 . La longueur d'une chaîne de \mathcal{A}^* est le nombre de symboles qu'elle contient (en comptant les répétitions). La chaîne 01101110101010 a pour longueur 14.

Définition 5.3.1 Soit \mathcal{X} un ensemble fini. Un code de \mathcal{X} est une fonction $c : \mathcal{X} \rightarrow \mathcal{A}^*$, où \mathcal{A} est un ensemble fini de cardinal D . On dit que la chaîne $c(x)$ est le mot-code associé au message $x \in \mathcal{X}$. On note $l(c(x))$ la longueur de $c(x)$.

Définition 5.3.2 Le code c est dit *uniquement déchiffrable (UD)* si pour tous entiers $k \geq 1$, $l \geq 1$, et toutes suites $x_1, \dots, x_k, y_1, \dots, y_l \in \mathcal{X}$:

$$c(x_1) \dots c(x_k) = c(y_1) \dots c(y_l) \Rightarrow k = l, x_1 = y_1, \dots, x_k = y_k$$

Définition 5.3.3 Un code c est dit avoir la *propriété du préfixe* si il n'existe aucune paire $x, y \in \mathcal{X}$, $x \neq y$ telle que $c(x)$ soit un préfixe de $c(y)$ ($c(y) = c(x)w$ où $w \in \mathcal{A}^*$, $w \neq \emptyset$). Un tel code sera appelé *code-préfixe*.

Le résultat suivant est une conséquence immédiate de la définition d'un code-préfixe.

Théorème 5.3.1 Un code-préfixe est *uniquement déchiffrable*.

EXEMPLE 5.3.1: Considérons les codes suivants :

1. $\mathcal{X} = \{1, 2, 3, 4\}$, $\mathcal{A} = \{0, 1\}$, $c(1) = 00$, $c(2) = 01$, $c(3) = 10$, $c(4) = 11$
2. $\mathcal{X} = \{1, 2, 3, 4\}$, $\mathcal{A} = \{0, 1\}$, $c(1) = 0$, $c(2) = 1$, $c(3) = 10$, $c(4) = 11$
3. $\mathcal{X} = \{1, 2, 3, 4\}$, $\mathcal{A} = \{0, 1\}$, $c(1) = 0$, $c(2) = 10$, $c(3) = 110$, $c(4) = 111$

Le code 1 et 3 sont UD (ils ont tous les deux la propriété du préfixe), mais pas celui de l'exemple 2 (en effet : $c(2) * c(1) = c(3)$).

Le théorème qui suit nous dit que les longueurs des mots-code d'un code UD vérifient l'inégalité de Kraft, et que, inversement, si des longueurs vérifient cette inégalité, alors il existe un code dont les mots-code ont les longueurs en question.

Théorème 5.3.2 Soit un code $c : \mathcal{X} \rightarrow \mathcal{A}^*$. Notons $D = |\mathcal{A}|$.

1. Si le code est UD, on a $\sum_{x \in \mathcal{X}} D^{-l(c(x))} \leq 1$.
2. Si $(l(x), x \in \mathcal{X})$ est une suite d'entiers telle que $\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1$, alors il existe un code UD tel que $l(c(x)) = l(x)$ pour tout $x \in \mathcal{X}$.

Démonstration. Si c est UD, on définit le code produit d'ordre n , $c^{(n)} : \mathcal{X}^n \rightarrow \mathcal{A}^*$, par

$$c^{(n)}((x_1 \cdots x_n)) = c(x_1) \cdots c(x_n).$$

Ce code est lui aussi UD, et on a (en notant, pour simplifier, $l(c(x)) = l(x)$)

$$\begin{aligned} \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^n &= \sum_{x_1 \in \mathcal{X}} \cdots \sum_{x_n \in \mathcal{X}} D^{l(x_1) + \cdots + l(x_n)} \\ &= \sum_{x^{(n)} \in \mathcal{X}^n} D^{-l(x^{(n)})}, \end{aligned}$$

où $l(x^{(n)}) = l(x_1) + \cdots + l(x_n)$ est la longueur du mot-code $c^{(n)}(x^{(n)})$ où $x^{(n)} = (x_1, \dots, x_n)$. On décompose ensuite la somme précédente suivant chaque longueur k possible ($k \geq 1$ car dans un code UD, il n'y a pas de mot de code de longueur 0) :

$$\begin{aligned} \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^n &= \sum_{k \geq 1} \sum_{\substack{x^{(n)} \in \mathcal{X}^n \\ l(x^{(n)}) = k}} D^{-k} \\ &= \sum_{k \geq 1} \alpha(k) D^{-k} = \sum_{k=1}^{nl_{max}} \alpha(k) D^{-k} \end{aligned}$$

Avec $\alpha(k)$ le nombre de mots de code de $c^{(n)}$ de longueur k et l_{max} est la longueur du plus long mot-code de c . Comme $c^{(n)}$ est uniquement décodable, il y a au plus D^k mots de code de longueur k . Il vient alors :

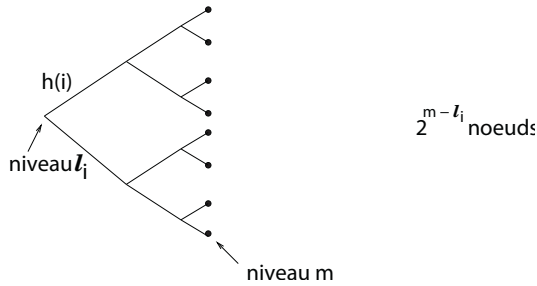
$$\left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^n \leq \sum_{k=1}^{nl_{max}} D^k D^{-k} = nl_{max}$$

ce qui donne :

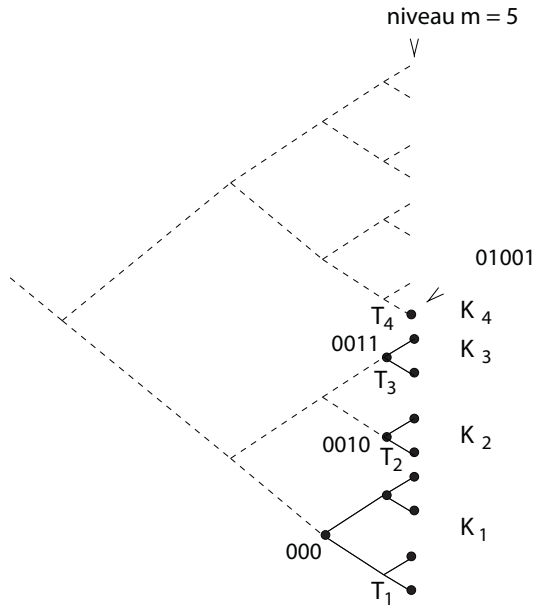
$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq (nl_{max})^{\frac{1}{n}}$$

Le membre de droite de cette inégalité tend vers 1 lorsque n tend vers l'infini, ce qui donne l'inégalité annoncée.

Nous allons démontrer l'assertion 2. On commence par rebaptiser les éléments \mathcal{X} $1, \dots, K$ de sorte que $l(1) \leq l(2) \leq \dots \leq l(K)$. On utilise l'arbre D -aire complet de profondeur $m \geq \max_{x \in \mathcal{X}} l(x)$ dont on partitionne les D^m feuilles (nœuds terminaux) de la manière suivante. On place les $D^{m-l(1)}$ premières feuilles dans le groupe 1, puis on place les $D^{m-l(2)}$ feuilles suivantes dans le groupe 2, et ainsi de suite.



La condition $\sum_{i=1}^K D^{m-l(i)} \leq D^m$ assure que cette construction est possible (pas de "débordement"). On définit ensuite $c(i)$ la représentation binaire du nœud $h(i)$ commun aux feuilles du i -ème groupe. Ce code est un code préfixe, comme on le voit en contemplant la figure qui suit.



□

Le problème du codage de source

Le problème central du codage de source est celui de trouver pour un ensemble fini \mathcal{X} une application $c : \mathcal{X} \rightarrow \mathcal{A}^*$ (appelée *code*), qui minimise la longueur moyenne d'un *mot-code*

$$L(c) = \sum_{x \in \mathcal{X}} p(x) l(c(x)).$$

En énumérant les éléments de \mathcal{X} par $1, \dots, K$, et en notant l_1, \dots, l_K les longueurs de leurs mots-code respectifs, on est conduit au problème de minimisation :

$$\min \sum_{i=1}^K p_i l_i$$

sous la contrainte de Kraft

$$\sum_{i=1}^K D^{-l_i} \leq 1.$$

On notera que les l_i doivent être des entiers. On commencera par étudier le problème en relaxant cette condition, et en supposant les l_i réels non négatifs. Dans ce cas, la contrainte peut alors être remplacée par

$$\sum_{i=1}^K D^{-l_i} = 1$$

car si $\sum_{i=1}^K D^{-l_i} < 1$, on peut diminuer des l_i , et donc diminuer $\sum_{i=1}^K p_i l_i$ tout en respectant la contrainte, et donc améliorer la solution.

Lemme 5.3.1 *La solution optimale du problème ainsi posé est $l_i^* = -\log_D p_i$*

Démonstration. Il suffit d'appliquer l'inégalité de Gibbs avec $q_i = D^{-l_i}$ (qui est bien une probabilité, au vu la contrainte réduite $\sum_{i=1}^K D^{-l_i} = 1$) :

$$-\sum_{i=1}^K p_i \log_D p_i \leq -\sum_{i=1}^K p_i \log_D D^{-l_i} = \sum_{i=1}^K p_i l_i.$$

□

Posons maintenant

$$l_i = \lceil -\log p_i \rceil.$$

En particulier $-\log_D p_i \leq l_i < -\log_D p_i + 1$, d'où l'on tire 2 remarques : Premièrement les nombres entiers ainsi définis satisfont l'inégalité de Kraft. Le Théorème 5.3.2 dit alors qu'il existe un code c uniquement décodable (et même : avec la propriété du préfixe) ayant les mêmes longueurs de mots-code. D'autre part, on a

$$H_D \leq \sum_{i=1}^K p_i l_i < H_D + 1.$$

On code maintenant une suite de variables aléatoires iid X_1, \dots, X_n . Dans ce cas, l'entropie de (X_1, \dots, X_n) est nH_D où H_D est l'entropie d'une quelconque de ces variables aléatoires, disons X_1 . Le résultat précédent montre l'existence d'un code $c^{(n)} : \mathcal{X}^n \rightarrow \mathcal{A}^*$ dont la longueur moyenne $L(c^{(n)})$ satisfait :

$$nH_D \leq L(c^{(n)}) \leq nH_D + 1.$$

Donc le nombre moyen de lettres de l'alphabet \mathcal{A} dont on a besoin par symbole, c'est-à-dire $\frac{L(c)}{n}$, vérifie :

$$H_D \leq \frac{L(c^{(n)})}{n} < H_D + \frac{1}{n}.$$

Le nombre moyen de lettres par symbole tend donc vers H_D lorsque n tend vers l'infini. On ne peut pas faire mieux. En effet, pour tout code UD de longueurs de mots-code l_i , $1 \leq i \leq K$, D^{-l_i} , $1 \leq i \leq K$, définit une sous-probabilité, et donc (inégalité de Gibbs)

$$H_D = - \sum_{i=1}^K p_i \log_D p_i \leq - \sum_{i=1}^K p_i \log_D D^{-l_i} = \sum_{i=1}^K p_i l_i.$$

Le code de Huffman

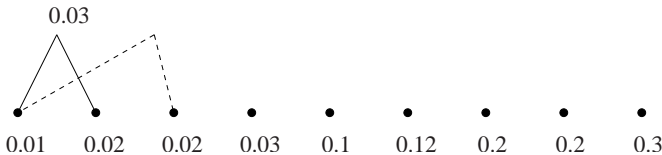
Un code dont la longueur moyenne est minimale peut être construit méthodiquement à l'aide d'un algorithme découvert par *Huffman* (1952). Nous allons montrer comment l'algorithme fonctionne dans un cas particulier puis nous ferons la preuve de son optimalité.

EXEMPLE 5.3.2: La distribution de probabilité est

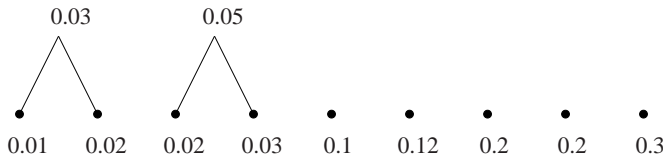
$$p = (0.01, 0.02, 0.02, 0.03, 0.1, 0.12, 0.2, 0.2, 0.3)$$

et on utilise l'alphabet binaire $\{0, 1\}$.

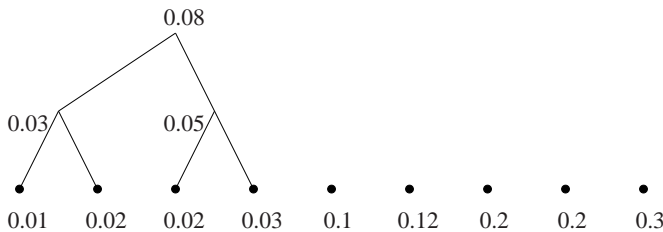
PREMIÈRE ITÉRATION. On commence par associer à chaque probabilité un point. Puis on regroupe les 2 points correspondant aux deux probabilités les plus faibles en un arbre complet à 1 niveau au sommet duquel on inscrit la somme des probabilités. (On a indiqué en pointillé le 2-ème choix possible. Il est équivalent au premier, en ce sens qu'il conduit à un autre code optimal.)



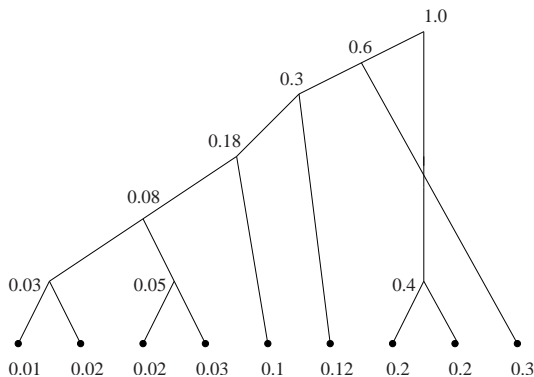
DEUXIÈME ITÉRATION. Les points utilisés sortent du jeu, mais le sommet de l'arbre rentre dans le jeu, porteur de la somme des probabilités des points sortants. Et on recommence. Par exemple avec le premier choix :



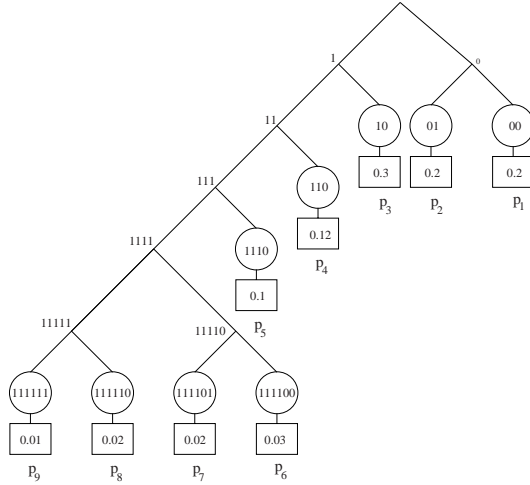
TROISIÈME ITÉRATION. Par exemple, avec le 2-ème choix dans la deuxième itération :



Ainsi de suite jusqu'à la
DERNIÈRE ITÉRATION.



RÉSULTAT FINAL. On réarrange le graphe précédent pour former un arbre binaire sans croisement et à chaque probabilité de la distribution initiale p on associe le code correspondant à sa place dans l'arbre binaire :



La longueur moyenne du code ainsi construit est :

$$\begin{aligned}
 L_0 &= 6 \times (0.01 + 0.02 + 0.02 + 0.03) + 4 \times 0.1 + 3 \times (0.3 + 0.2 + 0.2) \\
 &= 0.48 + 0.4 = 0.36 + 1.4 = 2.64
 \end{aligned}$$

La démonstration de l'optimalité de l'algorithme de Huffman est basée sur le lemme technique suivant :

Lemme 5.3.5 Si $n \geq 3$ et si $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ est une distribution de probabilité telle que

$$\pi_1 \geq \pi_2 \geq \dots \geq \pi_n > 0 ,$$

alors il existe un code optimal c pour π écrit dans l'alphabet binaire $\{0, 1\}$ et tel que

$$c(n) = w * 0 \quad , \quad c(n-1) = w * 1 \quad ,$$

pour au moins une suite w formée de 0 et de 1, et tel que le code c' défini par

$$c'(i) = c(i) \quad (1 \leq i \leq n-2) \quad , \quad h'(n-1) = w$$

est optimal pour la distribution $\pi' = (\pi_1, \pi_2, \dots, \pi_{n-2}, \pi_{n-1} + \pi_n)$.

Démonstration. Soit c un code optimal pour π de longueurs ℓ_1, \dots, ℓ_n . Les probabilités étant ordonnées comme indiqué ci-dessus, on peut supposer que

$$\ell_1 \leq \ell_2 \leq \dots \leq \ell_n .$$

En effet si pour une paire i, j tel que $i < j$ on avait $\ell_i > \ell_j$, le codage obtenu à partir de c en échangeant les mots-codes $c(i)$ et $c(j)$ aurait une longueur moyenne inférieure, tout en conservant la propriété du préfixe.

De plus $\ell_{n-1} = \ell_n$ car s'il n'en était pas ainsi on diminuerait la longueur moyenne tout en conservant la propriété du préfixe en supprimant les $\ell_n - \ell_{n-1}$ derniers digits du mot-code $c(n)$.

Le mot-code $c(n-1)$ s'écrit soit $w * 0$ soit $w * 1$, disons $w * 0$. On peut alors prendre $c(n) = w * 1$. En effet, il n'y a que deux raisons pour nous en empêcher : ou bien le mot $w * 1$ est le mot-code de $c(i)$ pour un $i < n-1$ et on n'aurait plus alors qu'à échanger les mots-codes $c(i)$ et $c(n)$, ou bien le mot $w * 1$ n'est pas un mot-code de c . Dans ce dernier cas on n'a qu'à changer $c(n)$ en $w * 1$ sans changer la longueur moyenne et en conservant la propriété du préfixe (en effet aucun mot-code $c(i)$ pour $i \neq n$ ne peut être préfixe de $w * 1$ sans quoi il serait préfixe de $w * 0 = c(n-1)$, ce qui n'est pas possible puisque c est un code avec la propriété du préfixe).

Considérons maintenant le codage \tilde{c} induit par c sur $\pi' = (\pi_1, \dots, \pi_{n-2}, \pi_{n-1} + \pi_n)$:

$$\begin{cases} \tilde{c}(i) = c(i) & (1 \leq i \leq n-2) \\ \tilde{c}(n-1) = w. \end{cases}$$

La longueur moyenne \tilde{L} de ce codage \tilde{c} est reliée à la longueur moyenne L du codage c par

$$L = \tilde{L} + \pi_{n-1} + \pi_n.$$

Soit maintenant L' la longueur moyenne du codage optimal c' sur π' . A partir du code c' on peut définir un codage \hat{c} sur π par

$$\begin{cases} \hat{c}(i) = c'(i) & (1 \leq i \leq n-2) \\ \hat{c}(n-1) = c'(n-1) * 0 \\ \hat{c}(n) = c'(n-1) * 0. \end{cases}$$

La longueur moyenne \hat{L} de \hat{c} est telle que

$$\hat{L} = L' + \pi_{n-1} + \pi_n.$$

Mais L est la longueur minimale des codages sur π et L' est la longueur minimale des codages sur π' . Donc $L' = \tilde{L}$ et \tilde{c} est un code optimal sur π' . \square

Le lemme ci-dessus justifie les itérations de l'algorithme de Huffman. En effet celles-ci nous ramènent chaque fois à un problème de codage optimal dont la dimension (le nombre de messages) a diminué d'une unité, jusqu'à ce qu'on tombe sur le problème de dimension 2 pour lequel le code optimal est tout trouvé : 0,1.

5.4 Autres interprétations de l'entropie

Le point de vue des questionnaires

On peut se demander pour quelle raison on appelle $H(X)$ la *quantité d'information* contenue dans X . Nous allons voir que $H(X)$ est le nombre (moyen) minimal de questions que l'on doit poser pour découvrir X . Plus précisément, supposons qu'on cherche à identifier un objet parmi K objets distinguables. Pour cela on peut poser toute question que l'on souhaite dont la réponse est binaire : oui ou non. Une stratégie de *questionnaire* consiste en

- (1) un enchaînement de questions, et
- (2) une règle d'arrêt-décision.

Chaque question de l'enchaînement dépend des réponses aux précédentes. On peut donc représenter les questions du questionnaires par des chaînes de symboles binaires. La première question est représentée par la suite vide \emptyset . La question 0110, par exemple, est la 5-ème question posée sachant que les réponses aux 4 questions précédentes sont, dans l'ordre, non, oui, oui, non. La règle d'arrêt-décision choisie est représentée par le choix de K nœuds de l'arbre binaire, N_1, \dots, N_K , avec l'interprétation suivante. Si l'objet j est choisi, le questionnaire suivra le chemin qui va de la racine au nœud N_j où la réponse exacte est alors donnée.

On peut considérer que le mot binaire associé au nœud N_j est le mot-code de l'objet j . Comme le questionnaire doit être admissible, en ce sens qu'il doit conduire à une réponse juste quel que soit l'objet sélectionné, le code binaire ainsi défini a la propriété du préfixe. En effet si pour $i \neq j$ le mot-code N_j était un préfixe de N_i , alors on aboutirait à une erreur si l'objet i était choisi, puisque le questionnaire produirait alors la réponse j . La discussion précédente montre qu'un questionnaire admissible est identifiable à un code-préfixe. Si l'objet à identifier est sélectionné selon la distribution de probabilité p_i , $1 \leq i \leq K$, le nombre moyen de questions à poser est la longueur moyenne du code. On peut donc interpréter la quantité $H_2 = -\sum_{i=1}^K p_i \log_2 p_i$ comme le nombre moyen minimum de questions à réponses binaires (oui ou non) nécessaires pour identifier un objet ainsi choisi au hasard (il faut au besoin regrouper les objets à identifier par groupes de longueur n ; voir la discussion de la Section 5.3).

L'interprétation de Boltzmann

Nous avons donné 3 interprétations de la quantité d'information H :

- (a) *Interprétation du codage* : $H_D(p)$ est la place moyenne minimale qu'occupe un symbole de l'ensemble $M = \{x_1, \dots, x_k\}$ lorsqu'il y a un grand nombre de ces symboles dans les proportions p_1, \dots, p_k et lorsqu'on dispose pour représenter ces symboles d'un alphabet à D lettres.
- (b) *Interprétation des questionnaires* : $H_D(p)$ est le temps moyen minimal d'identification d'un de ces symboles (objets) parmi la population décrite dans (a), lorsqu'on tire les symboles au hasard (donc avec la probabilité p_i pour x_i) et lorsqu'on peut poser toute question qui admet au plus D réponses.

(c) *Interprétation des suites typiques* : $D^{nH_D(p)}$ est le nombre de suites “typiques” de n symboles pris dans un alphabet à D lettres, lorsque ces lettres sont tirées au sort indépendamment les unes des autres selon la distribution p . Ces suites “typiques” ont la propriété suivante : la probabilité de tirer au sort une suite non typique peut être rendue aussi faible que désiré en choisissant n suffisamment grand.

Voici brièvement une quatrième interprétation de $H(p)$, due à Boltzmann. Boltzmann avait fait l'hypothèse que l'entropie au sens thermodynamique d'un système de n particules dans k “micro-états” différents était proportionnel au nombre de “configurations” indistinguables que peut prendre le système. Considérons un ensemble de n particules, chacune d'elles pouvant se trouver dans un micro-état donné parmi E_1, \dots, E_k . Pour rendre la discussion qui va suivre plus concrète, on peut imaginer que les particules en question sont les électrons de n atomes d'hydrogène (rappelons qu'un atome d'hydrogène a un seul électron en orbite autour de son noyau), et que E_1, \dots, E_k sont les niveaux d'énergie de chaque électron permis par la théorie quantique (modèle de Bohr).

On appelle l'ensemble des n particules le *système*. Un *état du système* est un k -tuplet (n_1, \dots, n_k) où $\sum_{i=1}^n n_i = n$, avec l'interprétation suivante : il y a n_i particules dans le micro-état E_i . Si les particules étaient distinguables entre elles, il y aurait $c(n, n_1, \dots, n_k) = n! / n_1! \dots n_k!$ configurations différentes conduisant à l'état (n_1, \dots, n_k) . Nous allons comparer, pour les très grandes valeurs de n , les nombres de configurations correspondant à deux états. Pour cela, on forme le rapport de $c(n_1, \dots, n_k)$ et $c(\tilde{n}_1, \dots, \tilde{n}_k)$ respectivement.

$$\frac{n!}{n_1! \dots n_k!} / \frac{n!}{\tilde{n}_1! \dots \tilde{n}_k!} = \prod_{i=1}^k \frac{\tilde{n}_i!}{n_i!}.$$

Nous allons donner un équivalent asymptotique de ce rapport lorsque n tend vers l'infini de telle façon que

$$\lim_{n \uparrow \infty} \frac{n_i}{n} = p_i \quad , \quad \lim_{n \uparrow \infty} \frac{\tilde{n}_i}{n} = \tilde{p}_i.$$

La quantité équivalente nous sera fournie par la formule de Stirling. On trouve (faire les calculs) que

$$\prod_{i=1}^k \frac{\tilde{n}_i!}{n_i!} \simeq e^{n(-\sum_{i=1}^k p_i \log p_i + \sum_{i=1}^k \tilde{p}_i \log \tilde{p}_i)}.$$

Un état (n_1, \dots, n_k) a donc d'autant plus de configurations que son entropie est grande. L'état maximisant cette quantité est donc le plus typique, celui qu'on a le plus de chance de rencontrer. Le principe de base de la thermodynamique statistique des particules devient alors naturel : si la physique du système ne favorise aucune des configurations qui satisfont aux contraintes macroscopiques, le système se trouvera dans l'état qui maximise l'entropie sous les contraintes macroscopiques.

Prenons par exemple le cas des électrons de valence d'un système de n atomes d'hydrogène et supposons qu'il n'y ait qu'une contrainte macroscopique, à savoir que l'énergie moyenne de électrons de valence est fixée, égale à E :

$$\sum_{i=1}^k p_i E_i = E .$$

L'état du système sera donc celui pour lequel $-\sum_{i=1}^k p_i \log p_i$ est maximum, sous cette contrainte d'énergie moyenne. On verra dans l'Exercice 5.5.4 que pour cet état,

$$p_i = C e^{-\lambda E_i} , \quad (5.1)$$

où C et λ sont choisis de telle sorte que la contrainte physique d'énergie moyenne et la contrainte mathématique $\sum_{i=1}^k p_i = 1$ soient vérifiées.

5.5 Exercices

Exercice 5.5.1.

1. Soit X une variable aléatoire à valeurs dans l'ensemble \mathcal{X} fini, et $Y = \varphi(X)$ où φ est une fonction déterministe bijective. Montrez que

$$H(Y) = H(X) .$$

2. Soit Z est une variable aléatoire à valeurs dans \mathcal{Z} fini, et ψ est une fonction déterministe (pas nécessairement bijective). Montrez que

$$H(Z, \psi(Z)) = H(Z) .$$

Exercice 5.5.2.

Montrez que si un codage a la propriété du préfixe, alors le codage obtenu en mettant les mots-code à l'envers est uniquement déchiffrable. En déduire un code uniquement déchiffrable qui n'a pas la propriété du préfixe.

Exercice 5.5.3.

Trouvez un codage binaire optimal pour la distribution

$$p = (0.01, 0.04, 0.05, 0.07, 0.09, 0.1, 0.14, 0.2, 0.3) .$$

Exercice 5.5.4. *

L'entier k étant fixé, montrez que la distribution (p_1, \dots, p_k) d'information moyenne maximale sous la contrainte $\sum_{i=1}^k p_i E_i = E$ (où les E_i et E sont des nombres réels non négatifs) existe et a la forme

$$p_i = c e^{-\lambda E_i} .$$

Exercice 5.5.5. ENTROPIE CONDITIONNELLE.

Soit X et Y deux variables aléatoires à valeurs dans les ensembles finis \mathcal{X} et \mathcal{Y} respectivement. On note les distributions de probabilité correspondantes p_X , p_Y , et $p_{X,Y}$ la distribution de probabilité de (X, Y) . On définit la fonction $p_{X|Y}$ par

$$\begin{aligned} p_{X|Y}(x, y) &= P(X = x | Y = y) \\ &= \frac{p_{X,Y}(x, y)}{p_Y(y)}. \end{aligned}$$

Cette dernière quantité est la probabilité conditionnelle de $X = x$ sachant $Y = y$. L'entropie conditionnelle de X sachant Y est, par définition, la quantité :

$$H(X|Y) = E[-\log p_{X|Y}(X|Y)].$$

Montrez que

$$H(X, Y) = H(X) + H(Y|X)$$

et symétriquement, $H(Y|X) = H(X, Y) - H(X)$. (La première identité reçoit l'interprétation intuitive suivante en termes de questionnaires : "Le nombre de questions $H(X, Y)$ nécessaires pour identifier (X, Y) est la somme du nombre de questions $H(X)$ nécessaires pour identifier X plus le nombre de questions $H(Y|X)$ nécessaires pour identifier Y sachant quel est X ".)

Exercice 5.5.6.

Démontrez les trois identités suivantes (Voir l'exercice précédent pour les notations) :

$$H(X_1, \dots, X_{n+1}) = H(X_1, \dots, X_n) + H(X_{n+1}|X_1, \dots, X_n)$$

$$H(X_1, \dots, X_n) = H(X_1) + \sum_{i=2}^n H(X_i|X_1, \dots, X_{i-1})$$

$$H(X, Y|Z) = H(Y|Z) + H(X|Y, Z)$$

Exercice 5.5.7. INFORMATION MUTUELLE.

Soit X et Y deux variables aléatoires discrètes, à valeurs dans les ensembles finis \mathcal{X} et \mathcal{Y} respectivement, et de distribution jointe $p_{X,Y}$. Soient p_X et p_Y leurs distributions marginales. L'information mutuelle entre X et Y est, par définition, la quantité

$$I(X; Y) = E \left[\log \frac{p_{X,Y}(X, Y)}{p_X(X)p_Y(Y)} \right].$$

Montrez que

$$\begin{aligned} I(X; Y) &= H(X) + H(Y) - H(X, Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned}$$

Montrez que $I(X; Y) \geq 0$ avec égalité si et seulement si X et Y sont indépendantes. En déduire que $H(X|Y) \leq H(X)$.

Exercice 5.5.8. ENTROPIE D'UNE SUITE STATIONNAIRE.

Soit $\{X_n\}_{n \geq 1}$ une suite de variables aléatoires à valeurs dans \mathcal{X} . Elle est dite *stationnaire* si pour tout $m \geq 1$, le vecteur $(X_{1+k}, \dots, X_{m+k})$ a une distribution de probabilité indépendante de $k \geq 1$. Soit $\{X_n\}_{n \geq 1}$ une suite de variables aléatoires à valeurs dans \mathcal{X} et stationnaire. Montrez que la limite $H = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n)$ existe et que

$$H = \lim_{n \rightarrow \infty} H(X_n | X_1, \dots, X_{n-1}) .$$

Les exercices qui suivent concernent la notion d'entropie dans le cas où les variables ne sont plus discrètes, mais admettent des densités de probabilité.

Exercice 5.5.9. ENTROPIE DIFFÉRENTIELLE.

On se donne un vecteur aléatoire $X \in \mathbb{R}^d$ doté d'une densité de probabilité $f(x)$. On définit l'entropie différentielle de X :

$$\begin{aligned} h(X) &= E[-\log f(X)] \\ &= - \int_{\mathbb{R}^d} f(x) \log(f(x)) dx \end{aligned}$$

(si l'intégrale est bien définie).

- (1) Calculez cette entropie différentielle quand $X \in \mathbb{R}$, $X \sim \mathcal{N}(0, \sigma^2)$.
- (2) Faire de même quand $X \in \mathbb{R}^d$, $X \sim \mathcal{N}(\mu, \Gamma)$ avec Γ matrice strictement positive.

Exercice 5.5.10. Δ -APPROXIMATION

On se donne une variable aléatoire réelle X avec une densité de probabilité f continue. On se donne également un nombre $\Delta > 0$. On appelle delta-approximation de la variable aléatoire réelle X la variable aléatoire X_Δ telle que $X_\Delta = x_i$ pour $x_i \in [i\Delta, (i+1)\Delta)$ tel que :

$$\int_{i\Delta}^{(i+1)\Delta} f(x) dx = f(x_i) \Delta .$$

Montrez que :

$$\lim_{\Delta \downarrow 0} (H(X_\Delta) + \log \Delta) = h(X) .$$

Exercice 5.5.11. DISTANCE DE KULLBACK-LEIBLER.

Soient f et g des densités de probabilité sur \mathbb{R}^d . On définit la quantité $D(f|g)$ par :

$$\begin{aligned} D(f|g) &= \int_{\mathbb{R}^d} f(x) \log \frac{f(x)}{g(x)} dx \quad \text{si } f \ll g \\ &= +\infty \quad \text{autrement} \end{aligned}$$

où $f \ll g$ signifie que $P(g(X) = 0) = 0$ si X admet la densité de probabilité f . Montrez que

$$D(f|g) \geq 0,$$

avec égalité si et seulement si f et g sont identiques.

(2) Soit $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^m$ des vecteurs aléatoires admettant les densités de probabilité marginales f_X , f_Y , et la densité de probabilité jointe $f_{X,Y}$. On note $g = f_X \otimes g_Y$ la densité de probabilité $g(x, y) = f_X(x)g_Y(y)$. On appelle information mutuelle de X et Y la quantité

$$I(X; Y) = D(f_{X,Y} | f_X \otimes f_Y).$$

Montrez que cette quantité est bien définie, et que

$$I(X; Y) \geq 0,$$

avec égalité si et seulement si X et Y sont indépendantes.

(3) Montrez que

$$\begin{aligned} I(X; Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(X|Y) \leq 0 \end{aligned}$$

Exercice 5.5.12. *

(suite des Exercices 5.5.9 et 5.5.11.) Montrez que :

$$\begin{aligned} h(X_1, \dots, X_n) &= \sum_{i=1}^n h(X_i | X_1, \dots, X_{i-1}) \\ &\leq \sum_{i=1}^n h(X_i) \end{aligned}$$

et

- $h(X + c) = h(X)$
- $h(aX) = h(X) + \log |a|$

Exercice 5.5.13. GAUSS MAXIMISE L'ENTROPIE.

Soient $X \in \mathbb{R}^d$ tel que $E[X] = 0$ et $\Gamma = E[XX^T]$. Montrez que :

$$h(X) \leq \frac{1}{2} \log \left[(2\pi e)^d |\Gamma| \right],$$

avec égalité si et seulement si :

$$X \sim \mathcal{N}(0, \Gamma).$$

Chapitre 6

L'espérance comme intégrale

6.1 Résumé de la théorie de l'intégration

En plaçant la théorie des probabilités dans son cadre mathématique, ce chapitre permettra de justifier les manipulations formelles effectuées dans les chapitres précédents. La théorie des probabilités est, techniquement, un domaine particulier de la théorie de la mesure et de l'intégration. Les résultats de cette dernière théorie seront pour la plupart présentés sans démonstration car le but poursuivi est simplement de donner au lecteur une connaissance opérationnelle de la théorie sous-jacente en même temps qu'une motivation pour une étude plus approfondie.

Le symbole

$$\int_X f(x) \mu(dx) \tag{*}$$

(l' *intégrale de Lebesgue* d'une fonction f par rapport à la mesure μ) recouvre une variété d'objets mathématiques, tels que l'intégrale de Lebesgue sur \mathbb{R} ,

$$\int_{\mathbb{R}} f(x) dx$$

(qui généralise l'intégrale de Riemann) et l'intégrale de Lebesgue sur \mathbb{R}^d . Une somme infinie

$$\sum_{n \in \mathbb{Z}} f(n)$$

est aussi une intégrale de Lebesgue (cette fois-ci par rapport à la mesure de comptage sur \mathbb{Z}). L'intégrale de Stieltjes–Lebesgue

$$\int_{\mathbb{R}} f(x) dF(x)$$

par rapport à une fonction F à variation bornée, de même que l'espérance

$$E[Z] = \int_{\Omega} Z(\omega) P(d\omega)$$

d'une variable aléatoire Z , sont également des avatars de l'intégrale de Lebesgue.

L'intégrande f dans (\star) est une fonction à laquelle on impose des conditions très peu restrictives. Par exemple, il suffit qu'elle soit non négative et *mesurable*. Dans le cas particulier où $X = \mathbb{R}$ et μ est la mesure de Lebesgue sur \mathbb{R} — la “longueur” —, cette classe de fonctions est beaucoup plus grande que celle des fonctions non négatives intégrables au sens de Riemann. C'est donc la notion de FONCTION MESURABLE qu'il nous faudra d'abord introduire. Cela débute par une série de définitions.

Tribus

On note $\mathcal{P}(X)$ la collection de tous les sous-ensembles d'un ensemble arbitraire X . On rappelle la définition d'une *tribu* (voir la Définition 1.4.1) :

Définition 6.1.1 Une famille $\mathcal{X} \subseteq \mathcal{P}(X)$ de sous-ensembles de X est appelé *tribu* sur X si :

- (α) $X \in \mathcal{X}$;
- (β) $A \in \mathcal{X} \implies \bar{A} \in \mathcal{X}$;
- (γ) $A_n \in \mathcal{X}$ pour tout $n \in \mathbb{N} \implies \bigcup_{n=0}^{\infty} A_n \in \mathcal{X}$.

On dit alors que (X, \mathcal{X}) est un *espace mesurable*.

Une tribu est donc une collection de sous-ensembles de X qui contient X et qui est stable par complémentation et union dénombrable. On montre facilement qu'elle est aussi stable par intersection dénombrable (Exercice 1.4.1).

$\mathcal{P}(X)$ est la *tribu triviale*, et $\{\Omega, \emptyset\}$ la tribu *grossière*.

Définition 6.1.2 La tribu engendrée par une collection \mathcal{C} non vide de sous-ensembles X est, par définition, la plus petite tribu sur X contenant tous les ensembles dans \mathcal{C} . On la note $\sigma(\mathcal{C})$.

Voici une “construction” théorique de cette tribu. Tout d'abord, on remarque que si $(\mathcal{F}_i, i \in I)$ est une famille de tribus sur X , où I est un index quelconque, alors (Exercice 1.4.3) $\bigcap_{i \in I} \mathcal{F}_i$ est aussi une tribu sur X . Si maintenant $(\mathcal{F}_i, i \in I)$ est la famille des tribus qui contiennent tous les éléments de \mathcal{C} (famille non vide puisqu'elle contient la tribu triviale $\mathcal{P}(X)$), alors $\bigcap_{i \in I} \mathcal{F}_i$ est bien la plus petite tribu contenant tous les ensembles dans \mathcal{C} .

Nous allons définir, de deux manières équivalentes, les *ensembles boréliens* de \mathbb{R}^n . Rappelons qu'un ensemble $O \in \mathbb{R}^n$ est appelé *ouvert* si pour tout $x \in O$, on peut trouver une boule ouverte non vide de centre x et entièrement contenue dans O . Considérons l'ensemble \mathbb{R}^n munie de la topologie euclidienne.

Définition 6.1.3 La tribu borélienne, notée $\mathcal{B}(\mathbb{R}^n)$ ou \mathcal{B}^n , est, par définition, la tribu engendrée par les ouverts.

On démontre que :

Théorème 6.1.1 \mathcal{B}^n est également engendrée par la collection \mathcal{C} de tous les rectangles du type $\prod_{i=1}^n (-\infty, a_i]$, où $a_i \in \mathbb{Q}$ (les rationnels) pour tout $i \in \{1, \dots, n\}$.

Pour $n = 1$ on écrit $\mathcal{B}(\mathbb{R}) = \mathcal{B}$. Pour $I = \prod_{j=-1}^n I_j$, où I_j est un intervalle général de \mathbb{R} (I est alors appelé *rectangle* de \mathbb{R}^n), la tribu borélienne $\mathcal{B}(I)$ sur I est, par définition, celle qui contient tous les ensembles boréliens ($\in \mathcal{B}^n$) contenus dans I . Par définition, $\mathcal{B}(\overline{\mathbb{R}})$, ou $\overline{\mathcal{B}}$, est la tribu sur $\overline{\mathbb{R}}$ engendrée par les intervalles du type $[-\infty, a]$, $a \in \mathbb{R}$.

Fonction mesurable

Définition 6.1.4 Soit (X, \mathcal{X}) et (E, \mathcal{E}) deux espaces mesurables. Une fonction $f : X \rightarrow E$ est dite *fonction mesurable* de (X, \mathcal{X}) dans (E, \mathcal{E}) si

$$f^{-1}(C) \in \mathcal{X} \text{ pour tout } C \in \mathcal{E}. \quad (6.1)$$

Ceci est noté

$$f : (X, \mathcal{X}) \rightarrow (E, \mathcal{E}).$$

Lorsque $f : (X, \mathcal{X}) \rightarrow (\mathbb{R}^k, \mathcal{B}^k)$, on dit que f est une *fonction borélienne* de X dans \mathbb{R}^k . En général, dans une phrase telle que : “ f est une fonction borélienne définie sur X ”, la tribu \mathcal{X} est celle donnée par le contexte.

Une fonction $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$, où (X, \mathcal{X}) est un espace mesurable arbitraire, est appelée *fonction borélienne étendue*, ou simplement *fonction borélienne*. Quant à une fonction $f : (X, \mathcal{X}) \rightarrow (\mathbb{R}, \mathcal{B})$, elle est appelée fonction borélienne *réelle*.

Les résultats énoncés dans le théorème suivant sont utiles pour déterminer si une fonction est mesurable.

Théorème 6.1.2 1. Soit (X, \mathcal{X}) un espace mesurable et soit $n \geq 1$ un entier. Alors $f = (f_1, \dots, f_n)$ est une fonction mesurable de (X, \mathcal{X}) dans $(\mathbb{R}^n, \mathcal{B}^n)$ si et seulement si $\{f_i \leq a_i\} \in \mathcal{X}$ pour tout i , $1 \leq i \leq n$, et pour tous $a_i \in \mathbb{Q}$.

2. Toute fonction continue $f : \mathbb{R}^k \rightarrow \mathbb{R}^m$ est mesurable de $(\mathbb{R}^k, \mathcal{B}^k)$ dans $(\mathbb{R}^m, \mathcal{B}^m)$.

3. Soit (X, \mathcal{X}) , (Y, \mathcal{Y}) et (E, \mathcal{E}) trois espaces mesurables, et soit $\varphi : (X, \mathcal{X}) \rightarrow (Y, \mathcal{Y})$ et $g : (Y, \mathcal{Y}) \rightarrow (E, \mathcal{E})$ des fonctions mesurables. Alors $g \circ \varphi$ est une fonction mesurable de (X, \mathcal{X}) dans (E, \mathcal{E}) .

4. Soit $f, g : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ et soit $\lambda \in \mathbb{R}$. Alors fg , λf , $(f/g)1_{g \neq 0}$ sont des fonctions boréliennes. Il en est de même pour $(f+g)1_C(x)$, où C est l'ensemble où $f+g$ est bien définie (pas une forme indéterminée $\infty - \infty$).

5. Soit $\{f_n\}_{n \geq 1}$ une suite de fonctions mesurables de (X, \mathcal{X}) dans $(\overline{\mathbb{R}}, \overline{\mathcal{B}})$. Alors $\liminf_{n \uparrow \infty} f_n$ et $\limsup_{n \uparrow \infty} f_n$ sont des fonctions boréliennes, et l'ensemble

$$\{\limsup_{n \uparrow \infty} f_n = \liminf_{n \uparrow \infty} f_n\} = \{\exists \lim_{n \uparrow \infty} f_n\}$$

appartient à \mathcal{X} . En particulier, si $\{\exists \lim_{n \uparrow \infty} f_n\} = X$, la fonction $\lim_{n \uparrow \infty} f_n$ est une fonction borélienne.

En résumé : toutes les opérations “usuelles” sur les fonctions mesurables conduisent à des fonctions mesurables.

La clef de la construction de l'intégrale de Lebesgue est le *théorème d'approximation* par les *fonctions boréliennes simples*.

Définition 6.1.5 Une $f : X \rightarrow \mathbb{R}$ de la forme

$$f(x) = \sum_{i=1}^K \alpha_i 1_{A_i}(x) \quad (6.2)$$

où $K \in \mathbb{N}$, et où les $\alpha_i \in \mathbb{R}$ et les $A_i \in \mathcal{X}$, est dite fonction borélienne simple.

Théorème 6.1.3 Soit $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathbb{B}})$ une fonction borélienne non négative. Il existe une suite $\{f_n\}_{n \geq 1}$ non décroissante de fonctions boréliennes simples non négatives qui converge ponctuellement vers f .

Démonstration. On vérifie facilement que la suite définie par

$$f_n(x) = \sum_{k=0}^{n2^n-1} k2^{-n} 1_{A_{k,n}}(x) + n1_{A_n}(x),$$

où $A_{k,n} = \{x \in X : k2^{-n} < f(x) \leq (k+1)2^{-n}\}$ et $A_n = \{x \in X : f(x) > n\}$, a les bonnes propriétés. \square

Mesures

Définition 6.1.6 Soit (X, \mathcal{X}) un espace mesurable et soit $\mu : \mathcal{X} \rightarrow [0, \infty]$ une fonction d'ensembles telle que $\mu(\emptyset) = 0$ et, pour toute famille dénombrable $\{A_n\}_{n \geq 1}$ de sous-ensembles dans \mathcal{X} deux à deux disjoints,

$$\mu\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n). \quad (6.3)$$

La fonction μ est appelée mesure sur (X, \mathcal{X}) , et (X, \mathcal{X}, μ) est appelé espace mesuré.

La propriété (6.3) est la *sigma-additivité*. Les deux propriétés suivantes sont faciles à vérifier (voir (1.4) et (1.5)). La première est la *monotonie* :

$$A \subseteq B \text{ et } A, B \in \mathcal{X} \implies \mu(A) \leq \mu(B).$$

La seconde propriété est la *sous-sigma-additivité* :

$$A_n \in \mathcal{X} \text{ pour tout } n \in \mathbb{N} \implies \mu\left(\bigcup_{n=0}^{\infty} A_n\right) \leq \sum_{n=0}^{\infty} \mu(A_n).$$

Le résultat suivant s'appelle le théorème de *continuité séquentielle de la mesure* :

Théorème 6.1.4 Soit (X, \mathcal{X}, μ) un espace mesuré.

(a) Soit $\{A_n\}_{n \geq 1}$ une suite de sous-ensembles de \mathcal{X} , non-décroissante ($A_n \subseteq A_{n+1}$ pour tout $n \geq 1$). Alors

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \uparrow \infty} \mu(A_n).$$

(b) Soit $\{B_n\}_{n \geq 1}$ une suite de sous-ensembles de \mathcal{X} , non-croissante ($B_{n+1} \subseteq B_n$ pour tout $n \geq 1$), et tels que $\mu(B_{n_0}) < \infty$ pour un $n_0 \in \mathbb{N}_+$. Alors

$$\mu\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n \downarrow \infty} \mu(B_n).$$

Les démonstrations sont analogues à celles du Théorème 1.2.1. Toutefois, noter la légère restriction dans (b) ; voir l'Exercice 6.4.5.

EXEMPLE 6.1.1: LA MESURE DE DIRAC. Soit $a \in X$. La mesure ε_a définie par $\varepsilon_a(C) = 1_C(a)$ est appelée *mesure de Dirac* en $a \in X$. La fonction d'ensembles $\mu : \mathcal{X} \rightarrow [0, \infty]$ définie par

$$\mu(C) = \sum_{i=0}^{\infty} \alpha_i 1_{a_i}(C),$$

où $\alpha_i \in \mathbb{R}_+$ et $a_i \in X$ pour tout $i \in \mathbb{N}$, est une mesure, qu'on notera $\mu = \sum_{i=0}^{\infty} \alpha_i \varepsilon_{a_i}$.

EXEMPLE 6.1.2: MESURE DE COMPTAGE PONDÉRÉE. Soit $\{\alpha_n\}_{n \geq 1}$ une suite de nombres non négatifs. La fonction d'ensembles $\mu : \mathcal{P}(\mathbb{Z}) \rightarrow [0, \infty]$ définie par $\mu(C) = \sum_{n \in C} \alpha_n$ est une mesure sur $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}))$. Lorsque $\alpha_n \equiv 1$, c'est la *mesure de comptage* sur \mathbb{Z} .

EXEMPLE 6.1.3: MESURE DE LEBESGUE. Il existe une et une seule mesure ℓ sur $(\mathbb{R}, \mathcal{B})$ telle que

$$\ell((a, b]) = b - a.$$

Cette mesure est appelée *mesure de Lebesgue* sur \mathbb{R} . L'existence et l'unicité de la mesure de Lebesgue sont intuitives mais pas évidentes. Elles font l'objet d'un théorème (Théorème 6.1.6 ci-dessous).

Soit μ une mesure sur (X, \mathcal{X}) . Si $\mu(X) < \infty$, elle est dite *finie*. Si $\mu(X) = 1$, c'est une *probabilité* (mesure de probabilité). S'il existe une suite $\{K_n\}_{n \geq 1}$ de \mathcal{X} telle que $\mu(K_n) < \infty$ pour tout $n \geq 1$ et $\cup_{n=1}^{\infty} K_n = X$, elle est dite *sigma-finie*. Une mesure μ sur $(\mathbb{R}^n, \mathcal{B}^n)$ telle que $\mu(C) < \infty$ pour tout ensemble borné $C \in \mathcal{B}^n$ est dite *de Radon* (ou *localement finie*).

EXEMPLE 6.1.4: La mesure de Dirac ε_a est une mesure de probabilité. La mesure de comptage ν sur \mathbb{Z} est une mesure sigma-finie. Toute mesure de Radon sur $(\mathbb{R}^n, \mathcal{B}^n)$ est sigma-finie. La mesure de Lebesgue est une mesure de Radon.

Définition 6.1.7 On dit qu'une famille \mathcal{S} de sous ensembles de X est un π -système si elle est stable par intersection : si $A, B \in \mathcal{S}$, alors $A \cap B \in \mathcal{S}$.

Théorème 6.1.5 Soit \mathcal{S} un π -système d'ensembles mesurables de (X, \mathcal{X}) qui engendrent \mathcal{X} . Deux mesures μ_1, μ_2 définies sur cet espace mesurable coïncident si les deux conditions suivantes sont satisfaites :

(a) elles coïncident sur \mathcal{S}

(b) elles sont sigma-finies sur \mathcal{S} (il existe une suite $\{K_n\}_{n \geq 1}$ de \mathcal{S} croissant vers X et telle que $\mu(K_n) < \infty$ pour tout $n \geq 1$).

Fonction de répartition

Définition 6.1.8 Une fonction $F : \mathbb{R} \rightarrow \mathbb{R}$ est appelée fonction de répartition si elle a les propriétés suivantes :

1. F est non décroissante ;
2. F est continue à droite ;
3. F admet en tout point $x \in \mathbb{R}$ une limite à gauche notée $F(x-)$.

(En fait, la propriété 3 est une conséquence de 1 et 2.)

EXEMPLE 6.1.5: Soit μ mesure de Radon sur $(\mathbb{R}, \mathcal{B})$. La fonction

$$F_\mu(t) = \begin{cases} +\mu((0, t]) & \text{si } t \geq 0, \\ -\mu((t, 0]) & \text{si } t < 0 \end{cases}$$

est une fonction de répartition. On a de plus

$$\begin{aligned} F_\mu(b) - F_\mu(a) &= \mu((a, b]), \\ F_\mu(a) - F_\mu(a-) &= \mu(\{a\}). \end{aligned}$$

(La preuve est facile en utilisant la propriété de continuité séquentielle de la mesure ; voir l'Exercice 3.5.15.) La fonction F_μ est appelée fonction de répartition de μ .

Théorème 6.1.6 *Soit $F : \mathbb{R} \rightarrow \mathbb{R}$ une fonction de répartition. Il existe une unique mesure localement finie μ sur $(\mathbb{R}, \mathcal{B})$ telle que $F_\mu = F$.*

Ce résultat est facile à énoncer mais, comme on l'a dit plus haut au sujet de la mesure de Lebesgue, non trivial. Il est typique de la famille de résultats qui répondent à ce type de question : Soit \mathcal{C} une collection de sous-ensembles de X telle que $\mathcal{C} \subseteq \mathcal{X}$, où \mathcal{X} est une tribu sur X . Si on se donne une fonction d'ensembles $u : \mathcal{C} \rightarrow [0, \infty]$, existe-t-il une mesure μ sur (X, \mathcal{X}) telle que $\mu(C) = u(C)$ pour tout $C \in \mathcal{C}$, et si oui, est-elle unique ?

L'unicité est une conséquence du Théorème 6.1.5, car deux mesures de Radon ayant la même fonction de répartition coïncident sur le π -système formé par les intervalles du type $(a, b]$, qui engendre \mathcal{B} .

Presque partout ; ensembles négligeables

La notion d'*ensemble négligeable* a déjà été rencontrée pour les mesures de probabilité.

Définition 6.1.9 *Soit (X, \mathcal{X}, μ) un espace mesuré. Un ensemble μ -négligeable est, par définition, un ensemble contenu dans un ensemble mesurable $N \in \mathcal{X}$ tel que $\mu(N) = 0$. On dit qu'une propriété \mathcal{P} relative aux éléments $x \in X$ est vérifiée μ -presque-partout (μ -p.p.) si l'ensemble $\{x \in X : x \text{ ne vérifie pas } \mathcal{P}\}$ est μ -négligeable.*

Par exemple, si f et g sont deux fonctions boréliennes réelles définies sur X , l'expression

$$f \leq g \quad \mu\text{-p.p.}$$

signifie

$$\mu(\{x : f(x) > g(x)\}) = 0.$$

La démonstration du théorème suivant est immédiate en utilisant la propriété de sous-sigma-additivité de la mesure.

Théorème 6.1.7 *Une union dénombrable d'ensembles μ -négligeables est μ -négligeable.*

EXEMPLE 6.1.6: Tout singleton $\{a\}$, $a \in \mathbb{R}$, est un ensemble borélien de mesure de Lebesgue nulle. En effet, la tribu \mathcal{B} est engendrée par les intervalles $I_a = (-\infty, a]$, $a \in \mathbb{R}$ (Théorème 6.1.1), et donc $\{a\} = \bigcap_{n \geq 1} (I_a - I_{a-1/n})$ est aussi dans \mathcal{B} . Notant ℓ la mesure de Lebesgue, $\ell(I_a - I_{a-1/n}) = 1/n$, et donc $\ell(\{a\}) = \lim_{n \geq 1} \ell(I_a - I_{a-1/n}) = 0$. L'ensemble des rationnels \mathbb{Q} est un ensemble borélien de mesure de Lebesgue nulle car c'est une union dénombrable de singletons de mesure nulle.

Intégrale

Nous pouvons maintenant définir l'intégrale de Lebesgue d'une fonction mesurable $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ par rapport à μ , qu'on notera

$$\int_X f \, d\mu, \quad \int_X f(x) \mu(dx), \quad \text{ou} \quad \mu(f).$$

On définit l'intégrale en trois étapes :

ÉTAPE 1. Pour toute fonction borélienne simple non négative $f : (X, \mathcal{X}) \rightarrow (\mathbb{R}, \mathcal{B})$ de la forme (6.2), on définit l'intégrale de f par rapport à μ , par

$$\int_X f \, d\mu = \sum_{i=1}^K \alpha_i \mu(A_i).$$

ÉTAPE 2. Si $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ est non négative, on pose

$$\int_X f \, d\mu = \lim_{n \uparrow \infty} \int_X f_n \, d\mu, \quad (6.4)$$

où $\{f_n\}_{n \geq 1}$ est une suite non décroissante de fonctions boréliennes simples non négatives telle que $\lim_{n \uparrow \infty} f_n = f$ (Théorème 6.1.3). On démontre que l'intégrale ainsi définie est indépendante du choix de la suite approximante. On notera que la quantité (6.4) est non négative mais qu'elle peut être infinie.

ÉTAPE 3. On démontre que si $f \leq g$, où $f, g : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ sont non négatives, alors

$$\int_X f \, d\mu \leq \int_X g \, d\mu.$$

Par exemple,

$$\int_X f^\pm \, d\mu \leq \int_X |f| \, d\mu,$$

où

$$f^+ = \max(f, 0) \quad \text{et} \quad f^- = \max(-f, 0)$$

(et en particulier $f = f^+ - f^-$ et $f^\pm \leq |f|$). Donc, si

$$\int_X |f| \, d\mu < \infty, \quad (6.5)$$

le membre de droite de

$$\int_X f \, d\mu := \int_X f^+ \, d\mu - \int_X f^- \, d\mu \quad (6.6)$$

a un sens, et cela définit l'intégrale du membre de gauche. De plus, l'intégrale de f par rapport à μ définie de cette manière est finie.

Définition 6.1.10 Une fonction mesurable $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ satisfaisant (6.5) est dite μ -intégrable.

ÉTAPE 3 (suite). L'intégrale peut être définie pour des fonctions *non-intégrables*. Par exemple, pour toute fonction non négative. Plus généralement, si $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ est telle que au moins une des intégrales $\int_X f^+ d\mu$ ou $\int_X f^- d\mu$ est finie, on peut définir l'intégrale comme dans (6.6). Ceci conduit à des formes “fini moins fini”, “fini moins infini” et “infini moins fini” qui ont toutes un sens. Le cas rigoureusement exclu est celui pour lequel $\mu(f^+) = \mu(f^-) = +\infty$.

Nous allons suivre les détails de la construction dans deux exemples très simples.

EXEMPLE 6.1.7: INTÉGRALE PAR RAPPORT À LA MESURE DE COMPTAGE PONDÉRÉE. Toute fonction $f : \mathbb{Z} \rightarrow \mathbb{R}$ est mesurable par rapport à $\mathcal{P}(\mathbb{Z})$ et \mathcal{B} . Avec la mesure μ définie dans l'Exemple 6.1.2, et lorsque, par exemple $f \geq 0$, on a :

$$\mu(f) = \sum_{n=1}^{\infty} \alpha_n f(n).$$

La preuve est facile : il suffit de considérer la suite de fonctions simples qui approximent f :

$$f_n(k) = \sum_{j=-n}^{+n} f(j) 1_{\{j\}}(k)$$

dont l'intégrale est

$$\mu(f_n) = \sum_{j=-n}^{+n} f(j) \mu(\{j\}) = \sum_{j=-n}^{+n} f(j) \alpha_j$$

et de faire tendre n vers l' ∞ .

Lorsque $\alpha_n \equiv 1$, l'intégrale se réduit à une somme de série :

$$\mu(f) = \sum_{n \in \mathbb{Z}} f(n).$$

Dans ce cas, l'intégrabilité équivaut à l'absolue convergence de la série.

EXEMPLE 6.1.8: INTÉGRALE PAR RAPPORT À LA MESURE DE DIRAC. Soit ε_a la mesure de Dirac au point $a \in X$. Alors toute fonction $f : (X, \mathcal{X}) \rightarrow (\mathbb{R}, \mathcal{B})$ est ε_a -intégrable, et

$$\varepsilon_a(f) = f(a).$$

Pour une fonction borélienne simple f de la forme (6.2), on a

$$\varepsilon_a(f) = \sum_{i=1}^k a_i \varepsilon_a(A_i) = \sum_{i=1}^k a_i 1_{A_i}(a) = f(a).$$

Pour une fonction non négative f , et une suite non décroissante quelconque de fonctions boréliennes simples non négatives $\{f_n\}_{n \geq 1}$ convergant vers f , on a

$$\varepsilon_a(f) = \lim_{n \uparrow \infty} \varepsilon_a(f_n) = \lim_{n \uparrow \infty} f_n(a) = f(a).$$

Finalement, pour toute fonction $f : (X, \mathcal{X}) \rightarrow (\mathbb{R}, \mathcal{B})$

$$\varepsilon_a(f) = \varepsilon_a(f^+) - \varepsilon_a(f^-) = f^+(a) - f^-(a) = f(a)$$

est une quantité bien définie.

Propriétés élémentaires de l'intégrale

Rappelons que, pour tout $A \in \mathcal{X}$, par définition de l'intégrale,

$$\int_X 1_A d\mu = \mu(A). \quad (6.7)$$

On notera $\int_A f d\mu$ l'intégrale $\int_X 1_A f d\mu$ dès que cette dernière est bien définie. Voici la liste des propriétés usuelles de l'intégrale.

Théorème 6.1.8 Soit $f, g : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ des fonctions μ -intégrables, et soit $a, b \in \mathbb{R}$. Alors

- (a) $af + bg$ est μ -intégrable et $\mu(af + bg) = a\mu(f) + b\mu(g)$;
- (b) si $f = 0$ μ -p.p., alors $\mu(f) = 0$; si $f = g$ μ -p.p., alors $\mu(f) = \mu(g)$;
- (c) si $f \leq g$ μ -p.p., alors $\mu(f) \leq \mu(g)$;
- (d) $|\mu(f)| \leq \mu(|f|)$;
- (e) si $f \geq 0$ μ -p.p. et $\mu(f) = 0$, alors $f = 0$ μ -p.p. ;
- (f) si $\mu(1_A f) = 0$ pour tout $A \in \mathcal{X}$, alors $f = 0$ μ -p.p.
- (g) si f est μ -intégrable, alors $|f| < \infty$ μ -p.p.

Les propriétés (a) (linéarité) et (c) (monotonie) n'appellent pas de commentaire. La propriété (b) dit que deux fonctions presque partout égales ont la même intégrale, lorsque celle-ci est bien définie. La propriété (d) découle simplement de la définition, puisque $\mu(f) = \mu(f^+) - \mu(f^-)$ et $\mu(|f|) = \mu(f^+) + \mu(f^-)$. Les preuves de (e), (f) et (g) sont laissées en exercice (Exercice 6.4.6).

Soit $f : X \rightarrow \mathbb{C}$ une fonction complexe borélienne (c'est-à-dire $f = f_1 + if_2$, où $f_1, f_2 : (X, \mathcal{X}) \rightarrow (\mathbb{R}, \mathcal{B})$) telle que $\mu(|f|) < \infty$, on définit

$$\int_X f d\mu = \int_X f_1 d\mu + i \int_X f_2 d\mu.$$

L'extension aux fonctions boréliennes complexes des propriétés (a), (b), (d) and (f) du Théorème 6.1.8 est immédiate.

Le résultat suivant montre que le temps passé à calculer des intégrales de Riemann n'est pas du temps perdu.

Théorème 6.1.9 *Soit $f : (\mathbb{R}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B})$ une fonction intégrable au sens de Riemann. Alors elle est intégrable au sens de Lebesgue par rapport à ℓ , et son intégrale de Lebesgue est égale à son intégrale de Riemann.*

EXEMPLE 6.1.9: INTEGRABLE POUR LEBESGUE ET PAS POUR RIEMANN. La réciproque n'est pas vraie : la fonction f définie par $f(x) = 1_{\mathbb{Q}}$ est une fonction borélienne, intégrable au sens de Lebesgue, son intégrale étant égale à zéro car $\{f \neq 0\} = \mathbb{Q}$ a une ℓ -mesure nulle. Cependant, f n'est pas intégrable au sens de Riemann.

EXEMPLE 6.1.10: La fonction $f : (\mathbb{R}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B})$ définie par

$$f(x) = \frac{x}{1+x^2}$$

n'admet pas d'intégrale de Lebesgue, parce que

$$f^+(x) = \frac{x}{1+x^2} 1_{(0,\infty)}(x) \quad \text{et} \quad f^-(x) = -\frac{x}{1+x^2} 1_{(-\infty,0)}(x)$$

ont des intégrales de Lebesgue infinies. Cependant, cette fonction a une intégrale de Riemann *généralisée*

$$\lim_{A \uparrow \infty} \int_{-A}^{+A} \frac{x}{1+x^2} dx = 0.$$

6.2 Les grands théorèmes

Beppo Levi et Lebesgue

On commence par des résultats qui donnent des conditions permettant d'intervertir l'ordre des opérations de limite et d'intégration :

$$\int_X \lim_{n \uparrow \infty} f_n d\mu = \lim_{n \uparrow \infty} \int_X f_n d\mu. \quad (6.8)$$

EXEMPLE 6.2.1: LE CONTRE-EXEMPLE CLASSIQUE. On prend $(X, \mathcal{X}, \mu) = (\mathbb{R}, \mathcal{B}, \ell)$ et

$$f_n(x) = \begin{cases} 0 & \text{si } |x| > \frac{1}{n}, \\ n^2 x + n & \text{si } -\frac{1}{n} \leq x \leq 0, \\ -n^2 x + n & \text{si } 0 \leq x \leq \frac{1}{n}. \end{cases}$$

On a

$$\lim_{n \uparrow \infty} f_n(x) = 0 \quad \text{si } x \neq 0,$$

c'est-à-dire, $\lim_{n \uparrow \infty} f_n = 0$, μ -p.p. Donc $\mu(\lim_{n \uparrow \infty} f_n) = 0$. Cependant, $\mu(f_n) = 1$ pour tout $n \geq 1$.

Le théorème de convergence monotone ci-dessous est aussi appelé *théorème de Beppo Levi*.

Théorème 6.2.1 Soit $f_n : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$, $n \geq 1$, une suite de fonctions telles que

(i) $f_n \geq 0$ μ -p.p. ;

(ii) $f_{n+1} \geq f_n$ μ -p.p.

Alors, il existe une fonction $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ non négative telle que

$$\lim_{n \uparrow \infty} \uparrow f_n = f \quad \mu\text{-p.p.},$$

et l'égalité (6.8) est vraie.

Le théorème de la convergence dominée ci-dessous est aussi appelé *théorème de Lebesgue*.

Théorème 6.2.2 Soit $f_n : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$, $n \geq 1$, une suite de fonctions telles qu'il existe une fonction $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ et une fonction $g : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ μ -intégrable telles que :

(i) $\lim_{n \uparrow \infty} f_n = f$, μ -p.p. ;

(ii) $|f_n| \leq |g|$ μ -p.p. pour tout $n \geq 1$.

Alors l'égalité (6.8) est vraie.

Voici deux applications simples et importantes de ces résultats. Le *théorème de la mesure image* est spécialement utile en théorie des probabilités (voir plus bas).

Mesure image

Définition 6.2.1 Soit (X, \mathcal{X}) et (E, \mathcal{E}) deux espaces mesurables. Soit $h : (X, \mathcal{X}) \rightarrow (E, \mathcal{E})$ une fonction mesurable, et soit μ une mesure sur (X, \mathcal{X}) . On définit la fonction d'ensembles $\mu \circ h^{-1} : \mathcal{E} \rightarrow [0, \infty]$ par la formule

$$(\mu \circ h^{-1})(C) = \mu(h^{-1}(C)). \quad (6.9)$$

Alors, on vérifie (Exercice 6.4.9) que $\mu \circ h^{-1}$ est une mesure sur (E, \mathcal{E}) appelée l'image de μ par h .

Théorème 6.2.3 *Pour toute fonction non négative $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$*

$$\int_X (f \circ h)(x) \mu(dx) = \int_E f(y) (\mu \circ h^{-1})(dy). \quad (6.10)$$

Pour une fonction $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ de signe arbitraire, n'importe laquelle des deux conditions suivantes

(a) $f \circ h$ est μ -intégrable,

(b) f est $\mu \circ h^{-1}$ -intégrable,

implique l'autre, et dans ce cas, l'égalité (6.10) est vraie.

Démonstration. L'égalité (6.10) découle immédiatement de la définition quand f est une fonction simple non négative borélienne. Dans le cas général où f est non négative, on considère une suite de fonctions simples boréliennes non négatives $\{f_n\}_{n \geq 1}$ qui tend vers f . Alors (6.10) découle de la même égalité valable pour les f_n , en laissant $n \uparrow \infty$ et en invoquant le théorème de convergence monotone. Pour les fonctions mesurables de signe arbitraire, on applique (6.10) avec f^+ et f^- , et on fait la différence des égalités obtenues, en remarquant que cela ne conduit pas à une forme indéterminée. \square

Produit d'une mesure par une fonction

Définition 6.2.2 *Soit (X, \mathcal{X}, μ) un espace mesuré et soit $h : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ une fonction mesurable non négative. On définit une fonction d'ensembles $\nu : \mathcal{X} \rightarrow [0, \infty]$ par*

$$\nu(C) = \int_C h(x) \mu(dx). \quad (6.11)$$

Alors, on vérifie (Exercice 6.4.10) que ν est une mesure sur (X, \mathcal{X}) , qu'on appelle produit de μ par la fonction h .

Théorème 6.2.4 *Soit μ , h et ν comme dans la Définition 6.2.2. Pour toute fonction non négative $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$, on a :*

$$\int_X f(x) \nu(dx) = \int_X f(x) h(x) \mu(dx). \quad (6.12)$$

Si $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ est de signe arbitraire, alors l'une quelconque des deux conditions suivantes :

(a) f est ν -intégrable,

(b) fh est μ -intégrable,

entraîne l'autre, et l'égalité (6.12) a lieu.

Démonstration. L'égalité (6.12) découle immédiatement de la définition quand f est une fonction borélienne simple non négative. Dans le cas général où f est non négative, on considère une suite de fonctions simples boréliennes non négatives $\{f_n\}_{n \geq 1}$ qui tend vers f . Alors (6.12) découle de la même égalité valable pour les f_n , en laissant $n \uparrow \infty$ et en invoquant le théorème de convergence monotone. Pour les fonctions mesurables de signe arbitraire, on applique (6.12) avec f^+ et f^- , et on fait la différence des égalités obtenues, en remarquant que cela ne conduit pas à une forme indéterminée. \square

Tonelli et Fubini

On introduit la notion de *mesure produit* et on énonce les *théorèmes de Tonelli et Fubini* concernant la possibilité de changer l'ordre des intégrations.

Soit $(X_1, \mathcal{X}_1, \mu_1)$ et $(X_2, \mathcal{X}_2, \mu_2)$ deux espaces mesurés, où μ_1 et μ_2 sont *sigma-finies*. Sur l'ensemble produit $X = X_1 \times X_2$ on définit la *tribu produit* $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ comme étant la tribu engendrée par les ensembles de la forme $A_1 \times A_2$, où $A_1 \in \mathcal{X}_1$, $A_2 \in \mathcal{X}_2$.

Théorème 6.2.5 *Il existe une unique mesure μ sur $(X_1 \times X_2, \mathcal{X}_1 \times \mathcal{X}_2)$ telle que*

$$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2) \quad (6.13)$$

pour tout $A_1 \in \mathcal{X}_1$, $A_2 \in \mathcal{X}_2$.

La mesure μ ci-dessus est la *mesure produit* de μ_1 et μ_2 , et elle est notée $\mu_1 \otimes \mu_2$, ou, pour simplifier, $\mu_1 \times \mu_2$.

Le résultat ci-dessus s'étend de manière naturelle au produit d'un nombre fini de mesures sigma-finies.

EXEMPLE 6.2.2: MESURE DE LEBESGUE SUR \mathbb{R}^n . L'exemple classique de mesure produit est la mesure de Lebesgue sur $(\mathbb{R}^n, \mathcal{B}^n)$: c'est l'unique mesure ℓ^n sur cet espace mesurable telle que pour tous $A_1, \dots, A_n \in \mathcal{B}$,

$$\ell^n(\Pi_{i=1}^n A_i) = \Pi_{i=1}^n \ell(A_i).$$

Les résultats suivants, qui sont énoncés pour un produit de deux espaces mesurés, ont des extensions évidentes au cas du produit d'un nombre fini de tels espaces.

Théorème 6.2.6 *Soit $(X_1, \mathcal{X}_1, \mu_1)$ et $(X_2, \mathcal{X}_2, \mu_2)$ deux espaces mesurés où μ_1 et μ_2 sont sigma-finies. Soit $(X, \mathcal{X}, \mu) = (X_1 \times X_2, \mathcal{X}_1 \times \mathcal{X}_2, \mu_1 \otimes \mu_2)$.*

(A) Tonelli. Si $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ est non négative, alors, la fonction $x_2 \rightarrow f(x_1, x_2)$ est mesurable par rapport à \mathcal{X}_2 , et

$$x_1 \rightarrow \int_{X_2} f(x_1, x_2) \mu_2(dx_2)$$

est une fonction mesurable par rapport à \mathcal{X}_1 . De plus,

$$\int_X f \, d\mu = \int_{X_1} \left[\int_{X_2} f(x_1, x_2) \, \mu_2(dx_2) \right] \mu_1(dx_1). \quad (6.15)$$

(B) *Fubini*. Si $f : (X, \mathcal{X}) \rightarrow (\overline{\mathbb{R}}, \overline{\mathcal{B}})$ est μ -intégrable, alors, pour μ_1 -presque-tout x_1 , la fonction $x_2 \rightarrow f(x_1, x_2)$ est μ_2 -intégrable et $x_1 \rightarrow \int_{X_2} f(x_1, x_2) \, \mu_2(dx_2)$ est μ_1 -intégrable, et (6.15) est vraie.

On appellera le théorème global ci-dessus le théorème de *Fubini-Tonelli*.

La partie (A) dit qu'on peut intégrer une fonction borélienne non négative de plusieurs variables en intégrant successivement par rapport à chacune des variables, et ceci dans n'importe quel ordre. La partie (B) dit qu'on peut faire de même avec une fonction borélienne de signe arbitraire si cette fonction est μ -intégrable. En général, pour appliquer la partie (B), on utilise la partie (A) avec $f = |f|$ de manière à vérifier que $\int |f| \, d\mu < \infty$.

EXEMPLE 6.2.3: QUAND FUBINI NE S'APPLIQUE PAS. Soit la fonction f définie sur $X_1 \times X_2 = (1, \infty) \times (0, 1)$ par la formule

$$f(x_1, x_2) = e^{-x_1 x_2} - 2e^{-2x_1 x_2}.$$

On a

$$\begin{aligned} \int_{(1, \infty)} f(x_1, x_2) \, dx_1 &= \frac{e^{-x_2} - e^{-2x_2}}{x_2} \\ &= h(x_2) \geq 0, \\ \int_{(0, 1)} f(x_1, x_2) \, dx_2 &= -\frac{e^{-x_1} - e^{-2x_1}}{x_1} \\ &= -h(x_1). \end{aligned}$$

Cependant,

$$\int_0^1 h(x_2) \, dx_2 \neq \int_1^\infty (-h(x_1)) \, dx_1,$$

puisque $h \geq 0$ ℓ -p.p. sur $(0, \infty)$. On voit donc que des intégrations successives peuvent conduire à des résultats différents selon l'ordre des intégrations. En fait, dans cet exemple, $f(x_1, x_2)$ n'est pas intégrable sur $(0, 1) \times (1, \infty)$.

EXEMPLE 6.2.4: Soit $\{f_n\}_{n \in \mathbb{N}}$ une suite de fonctions mesurables de \mathbb{R} dans \mathbb{C} . Le théorème de Fubini appliqué au produit de la mesure de Lebesgue sur \mathbb{R} par la mesure de comptage sur \mathbb{Z} se lit dans ce cas de la façon suivante. Sous la condition

$$\int_{\mathbb{R}} \left(\sum_{n \in \mathbb{Z}} |f_n(t)| \right) dt < \infty,$$

on a pour presque tout $t \in \mathbb{R}$ (par rapport à la mesure de Lebesgue),

$$\sum_{n \in \mathbb{Z}} |f_n(t)| \, dt < \infty$$

et

$$\int_{\mathbb{R}} \left(\sum_{n \in \mathbb{Z}} f_n(t) \right) dt = \sum_{n \in \mathbb{Z}} \left(\int_{\mathbb{R}} f_n(t) dt \right).$$

Si les f_n sont non négatives, la dernière égalité a lieu sans condition.

EXEMPLE 6.2.5: INTÉGRATION PAR PARTIES. Soit μ_1 et μ_2 deux mesures sigma-finies sur $(\mathbb{R}, \mathcal{B})$. Pour tout intervalle $(a, b] \subseteq \mathbb{R}$

$$\mu_1((a, b]) \mu_2((a, b]) = \int_{(a, b]} \mu_1((a, t]) \, \mu_2(dt) + \int_{(a, b]} \mu_2((a, t)) \, \mu_1(dt). \quad (6.16)$$

La preuve consiste à calculer la μ -mesure du carré $(a, b] \times (a, b]$ de deux manières. La première est évidente et donne le premier membre de l'égalité à démontrer. La seconde consiste à observer que $\mu((a, b] \times (a, b]) = \mu(D_1) + \mu(D_2)$, où $D_1 = \{(x, y); a < y \leq b, a < x \leq y\}$ et $D_2 = (a, b] \times (a, b] \setminus D_1$. Alors $\mu(D_1)$ et $\mu(D_2)$ sont calculés grâce au théorème de Tonelli. Par exemple :

$$\mu(D_1) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} 1_{D_1}(x, y) \mu_1(dx) \right) \mu_2(dy)$$

et

$$\int_{\mathbb{R}} 1_{D_1}(x, y) \mu_1(dx) = \int_{\mathbb{R}} 1_{\{a < x \leq y\}} \mu_1(dx) = \mu_1((a, y]).$$

Définition 6.2.3 Soit μ une mesure de Radon sur $(\mathbb{R}, \mathcal{B})$ et soit F_μ sa fonction de répartition. La notation

$$\int_{\mathbb{R}} g(x) F_\mu(dx)$$

est utilisée à la place de $\int_{\mathbb{R}} g(x) \mu(dx)$. On appelle cette intégrale intégrale de Lebesgue–Stieltjes de g par rapport à F_μ .

Avec cette notation, (6.16) devient

$$F_1(b)F_2(b) - F_1(a)F_1(b) = \int_{(a, b]} F_1(x) dF_2(x) + \int_{(a, b]} F_2(x-) dF_1(x),$$

où $F_i := F_{\mu_i}$ ($i = 1, 2$). Ceci est la version de Lebesgue–Stieltjes de la formule d'intégration par parties classique de l'Analyse.

6.3 Probabilité

L'espérance comme intégrale

D'un point de vue formel, la théorie des probabilités n'est qu'un cas particulier de la théorie de la mesure. Cependant, leurs terminologies respectives sont différentes. Nous allons donner la "traduction" de l'une dans l'autre. Rappelons que dans le triplet (Ω, \mathcal{F}, P) , P est une mesure de probabilité, c'est-à-dire une mesure sur l'espace mesurable (Ω, \mathcal{F}) de masse totale $P(\Omega) = 1$.

Le théorème suivant est un cas particulier du Théorème 6.1.5.

Théorème 6.3.1 *Deux mesures de probabilité sur le même espace mesurable (Ω, \mathcal{F}) qui coïncident sur un π -système engendrant \mathcal{F} , sont identiques.*

Définition 6.3.1 *Soit (E, \mathcal{E}) un espace mesurable. Un élément aléatoire à valeurs dans (E, \mathcal{E}) est, par définition, une fonction mesurable $X : (\Omega, \mathcal{F}) \rightarrow (E, \mathcal{E})$.*

Si $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$, ou si $(E, \mathcal{E}) = (\overline{\mathbb{R}}, \overline{\mathcal{B}})$, X est appelé *variable aléatoire* (v.a.) (v.a. réelle si $X \in \mathbb{R}$, v.a. étendue si $X \in \overline{\mathbb{R}}$). Si $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}^n)$, X est appelé *vecteur aléatoire* (de dimension n), et alors $X = (X_1, \dots, X_n)$ où les X_i sont des variables aléatoires réelles. Une *variable aléatoire complexe* est une fonction $X : \Omega \rightarrow \mathbb{C}$ de la forme $X = X_R + iX_I$ où X_R et X_I sont des variables aléatoires réelles.

Définition 6.3.2 *Soit X un élément aléatoire à valeurs dans (E, \mathcal{E}) . La tribu engendrée par X , notée $\sigma(X)$, est, par définition, la tribu sur Ω engendrée par la collection $\mathcal{C} = \{X^{-1}(C); C \in \mathcal{E}\}$.*

Si X est un élément aléatoire à valeurs dans (E, \mathcal{E}) et si g est une fonction mesurable de (E, \mathcal{E}) dans $(\overline{\mathbb{R}}, \overline{\mathcal{B}})$, alors $g(X)$ est (Propriété (3) après la Définition 6.1.4) une variable aléatoire.

Puisque la variable aléatoire X est une fonction mesurable, on peut définir, sous certaines conditions, son intégrale par rapport à la mesure de probabilité P , qu'on appelle *espérance* de X :

$$E[X] = \int_{\Omega} X(\omega) P(d\omega).$$

Rappelons simplement les étapes de la construction de l'intégrale de Lebesgue dans le cas où la mesure μ est la mesure de probabilité P . D'abord, si $A \in \mathcal{F}$,

$$E[1_A] = P(A).$$

Plus généralement, si X est une variable aléatoire simple non négative :

$$X(\omega) = \sum_{i=1}^K \alpha_i 1_{A_i}(\omega),$$

où $\alpha_i \in \mathbb{R}$, $A_i \in \mathcal{F}$, alors

$$E[X] = \sum_{i=1}^N \alpha_i P(A_i).$$

Pour une variable aléatoire réelle non négative X , l'espérance est toujours définie par

$$E[X] = \lim_{n \uparrow \infty} E[X_n],$$

où $\{X_n\}_{n \geq 1}$ est une suite non-décroissante de fonctions boréliennes simples non négatives qui converge vers X . Cette définition ne dépend pas du choix de la suite approximante ayant les propriétés mentionnées. Si X est de signe arbitraire, l'espérance est définie par $E[X] = E[X^+] - E[X^-]$ si $E[X^+]$ et $E[X^-]$ ne sont pas tous deux infinis. Si $E[X^+]$ et $E[X^-]$ sont infinis, l'espérance n'est pas définie. Si $E[|X|] < \infty$, X est dite *intégrable*, et dans ce cas, $E[X]$ est un nombre fini.

Les propriétés de base de l'espérance sont la *linéarité* et la *monotonie* : si X_1 et X_2 sont des variables aléatoires pour lesquelles l'espérance est bien définie, alors, pour tous $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$E[\lambda_1 X_1 + \lambda_2 X_2] = \lambda_1 E[X_1] + \lambda_2 E[X_2],$$

dès que chacun des membres a un sens (pas de forme $\infty - \infty$). Aussi, lorsque $X_1 \leq X_2$, P-p.s., alors

$$E[X_1] \leq E[X_2].$$

Finalement, si $E[X]$ est bien défini, alors

$$|E[X]| \leq E[|X|].$$

Markov et Jensen. Nous allons énoncer deux résultats, à savoir l'inégalité de Markov et l'inégalité de Jensen, qui ont déjà été démontrés dans le cas de variables discrètes. Nous n'aurons pas besoin de présenter de preuves dans le cas général, puisque la preuve donnée dans le cas des variables discrètes n'utilise que les propriétés de linéarité et de monotonie de l'espérance et le fait que l'espérance d'une indicatrice d'événement est la probabilité de cet événement, propriétés qui sont générales.

Théorème 6.3.2 *Soit une variable aléatoire réelle non négative Z et un nombre réel $a > 0$. On a l'inégalité de Markov*

$$P(Z \geq a) \leq \frac{E[Z]}{a}.$$

En particulier, si X une variable aléatoire réelle de carré intégrable, pour tout $\varepsilon > 0$, on a l'inégalité de Chebyshev

$$P(|X - m_X| \geq \varepsilon) \leq \frac{\sigma_X^2}{\varepsilon^2}.$$

Démonstration. Voir le Théorème 2.2.2. □

Théorème 6.3.3 *Soit φ une fonction convexe définie sur un intervalle $I \subset \mathbb{R}$ contenant toutes les valeurs possibles d'une variable aléatoire réelle X . Alors si X et $\varphi(X)$ sont intégrables,*

$$\varphi(E[X]) \leq E[\varphi(X)] .$$

Démonstration. Voir le Théorème 2.2.3. □

Nous allons énoncer à nouveau, cette fois-ci en termes d'espérance, les théorèmes qui donnent des conditions générales garantissant que l'ordre des opérations de limite et d'espérance peuvent être intervertis :

$$E \left[\lim_{n \uparrow \infty} X_n \right] = \lim_{n \uparrow \infty} E[X_n] . \quad (6.17)$$

Tout d'abord, le théorème de CONVERGENCE MONOTONE (Théorème 6.2.1) :

Théorème 6.3.4 *Soit $\{X_n\}_{n \geq 1}$ une suite de variables aléatoires telle que pour tout $n \geq 1$,*

$$0 \leq X_n \leq X_{n+1}, \text{ P-p.s.}$$

Alors (6.17) est vraie.

Puis le théorème de CONVERGENCE DOMINÉE (Théorème 6.2.2).

Théorème 6.3.5 *Soit $\{X_n\}_{n \geq 1}$ une suite de variables aléatoires telle que pour tout n en dehors d'un ensemble P-négligeable \mathcal{N} : la limite $\lim_{n \uparrow \infty} X_n(\omega)$ existe, et pour tout $n \geq 1$*

$$|X_n| \leq Y ,$$

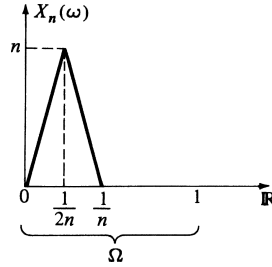
où Y est une variable aléatoire intégrable. Alors (6.17) est vraie.

EXEMPLE 6.3.1: Voici l'exemple classique montrant que (6.17) peut être mise en défaut si on n'impose pas de conditions adéquates. Pour la suite définie par la figure suivante, pour lequel P est la mesure de Lebesgue sur $\Omega = [0, 1]$, on a

$$E[\lim_{n \uparrow \infty} X_n] = E[0] = 0$$

et

$$\lim_{n \uparrow \infty} E[X_n] = \lim_{n \uparrow \infty} 1 = 1 .$$



Les deux exemples suivants justifient les calculs des premiers chapitres dans lesquels on s'était permis sans justification d'intervertir les opérations de somme infinie et d'espérance.

EXEMPLE 6.3.2: Soit $\{S_n\}_{n \geq 1}$ une suite de variables aléatoires non négatives. Alors

$$E \left[\sum_{n=1}^{\infty} S_n \right] = \sum_{n=1}^{\infty} E[S_n]. \quad (6.18)$$

Il suffit d'appliquer le théorème de convergence monotone, avec $X_n = \sum_{k=1}^n S_k$ et $X = \sum_{n=1}^{\infty} S_n$.

EXEMPLE 6.3.3: Soit $\{S_n\}_{n \geq 1}$ une suite de variables aléatoires réelles telle que $\sum_{n \geq 1} E[|S_n|] < \infty$. Alors (6.18) a lieu. Il suffit d'appliquer le théorème de convergence dominée avec $X_n = \sum_{k=1}^n S_k$, $X = \sum_{n=1}^{\infty} S_n$ et $Y = \sum_{k=1}^n |S_k|$. (Par le résultat de l'Exemple 6.3.2, $E[Y] = \sum_{k=1}^n E[|S_k|] < \infty$.)

Distribution d'un élément aléatoire

Définition 6.3.3 Soit X un élément aléatoire à valeurs dans (E, \mathcal{E}) . Sa distribution est, par définition, la mesure de probabilité Q_X sur (E, \mathcal{E}) , image de la mesure de probabilité P par l'application X de (Ω, \mathcal{F}) dans (E, \mathcal{E}) : pour tout $C \in \mathcal{E}$,

$$Q_X(C) = P(X \in C).$$

Si E est un ensemble dénombrable, muni de sa tribu "naturelle", la tribu triviale $\mathcal{P}(E)$, la distribution Q_X est entièrement déterminée par la donnée $\{Q(\{a\})\}_{a \in E}$, où

$$Q(\{a\}) = P(X = a).$$

Le Théorème 6.2.3 se lit, dans le contexte de la théorie des probabilités :

Théorème 6.3.6 Soit X un élément aléatoire à valeurs dans (E, \mathcal{E}) et de distribution de probabilité Q_X . Soit g est une fonction mesurable de (E, \mathcal{E}) dans $(\overline{\mathbb{R}}, \overline{\mathcal{B}})$. On a la formule du changement de variable :

$$E[g(X)] = \int_E g(x) Q_X(dx),$$

dès que l'un des deux membres de cette égalité a un sens (auquel cas l'autre membre est également bien défini).

Dans le cas particulier où $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$, prenant $C = (-\infty, x]$, on a

$$Q_X((-\infty, x]) = P(X \leq x) = F_X(x),$$

où F_X est la fonction de répartition de X , et

$$E[g(X)] = \int_{\mathbb{R}} g(x) Q(dx) = \int_{\mathbb{R}} g(x) dF(x),$$

par définition de l'intégrale de Stieltjes-Lebesgue du troisième membre.

Dans le cas particulier où $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}^n)$, et où le vecteur aléatoire X admet une densité de probabilité f_X , c'est-à-dire si sa distribution Q_X est le produit de la mesure de Lebesgue sur $(\mathbb{R}^n, \mathcal{B}^n)$ par la fonction f_X , le Théorème 6.2.4 nous dit que

$$E[g(X)] = \int_{\mathbb{R}^n} g(x) f_X(x) dx.$$

Dans le cas particulier où E est dénombrable,

$$E[g(X)] = \sum_{a \in E} g(a) P(X = a).$$

Indépendance

On rappelle les définitions concernant l'indépendance.

Définition 6.3.4 Deux événements A et B sont dits indépendants si

$$P(A \cap B) = P(A)P(B).$$

Plus généralement, une famille $\{A_i\}_{i \in I}$ d'événements, où I est un ensemble d'indices arbitraire, est dite indépendante si pour tout sous-ensemble $J \in I$ fini,

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j).$$

Définition 6.3.5 Deux éléments aléatoires $X : (\Omega, \mathcal{F}) \rightarrow (E, \mathcal{E})$ et $Y : (\Omega, \mathcal{F}) \rightarrow (G, \mathcal{G})$ sont dits indépendants si pour tous $C \in \mathcal{E}$, $D \in \mathcal{G}$

$$P(\{X \in C\} \cap \{Y \in D\}) = P(X \in C)P(Y \in D).$$

Plus généralement, une famille $\{X_i\}_{i \in I}$, où I est un ensemble d'indices arbitraire, de variable aléatoires $X_i : (\Omega, \mathcal{F}) \rightarrow (E_i, \mathcal{E}_i)$, $i \in I$, est dite indépendante si pour tout sous-ensemble $J \in I$ fini,

$$P\left(\bigcap_{j \in J} \{X_j \in C_j\}\right) = \prod_{j \in J} P(X_j \in C_j)$$

pour tout $C_j \in \mathcal{E}_j$ ($j \in J$).

Théorème 6.3.7 Si les éléments aléatoires $X : (\Omega, \mathcal{F}) \rightarrow (E, \mathcal{E})$ et $Y : (\Omega, \mathcal{F}) \rightarrow (G, \mathcal{G})$ sont indépendants, alors il en est de même de $\varphi(X)$ et $\psi(Y)$, où $\varphi : (E, \mathcal{E}) \rightarrow (E', \mathcal{E}')$, $\psi : (G, \mathcal{G}) \rightarrow (G', \mathcal{G}')$.

Démonstration. Pour tout $C' \in \mathcal{E}'$, $D' \in \mathcal{G}'$, les ensembles $C = \varphi^{-1}(C')$ et $D = \psi^{-1}(D')$ sont dans \mathcal{E} and \mathcal{G} respectivement, puisque φ et ψ sont mesurables. On a

$$\begin{aligned} P(\varphi(X) \in C', \psi(Y) \in D') &= P(X \in C, Y \in D) \\ &= P(X \in C) P(Y \in D) \\ &= P(\varphi(X) \in C') P(\psi(Y) \in D'). \end{aligned}$$

□

Le résultat ci-dessus, énoncé pour deux éléments aléatoires, s'étend au cas d'un nombre fini quelconque d'éléments aléatoires. Ce résultat a été appliqué à maintes reprises dans les premiers chapitres, sans justification à cause de son contenu intuitif.

EXEMPLE 6.3.4: Si $X = (X_1, \dots, X_n)$ et $Y = (Y_1, \dots, Y_n)$ sont des vecteurs réels indépendants, alors les variables $X_1^2 + \dots + X_n^2$ et $Y_1^2 + \dots + Y_n^2$ sont indépendantes.

Théorème 6.3.8 Soit (Ω, \mathcal{F}, P) un espace de probabilité sur lequel sont données deux variables aléatoires réelles X et Y . Pour que ces deux variables aléatoires soient indépendantes, il faut et il suffit que pour tout $a, b \in \mathbb{R}$, $P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b)$.

Démonstration. En effet, les distributions de probabilité $Q_{(X,Y)}$ et $Q_X \times Q_Y$ coïncident sur le π -système formé par les rectangles du type $(-\infty, a] \times (-\infty, b]$, et qui engendrent la tribu $\mathcal{B}(\mathbb{R}^2)$. Elles sont donc identiques, d'après le Théorème 6.3.1. □

L'indépendance de deux variables aléatoires X et Y est équivalente à la factorisation de leur distribution jointe :

$$Q_{(X,Y)} = Q_X \times Q_Y ,$$

où $Q_{(X,Y)}$, Q_X , et Q_Y sont les distributions de (X, Y) , X , et Y , respectivement. En effet, pour tous ensembles de la forme $C \times D$, où $C \in \mathcal{E}$, $D \in \mathcal{G}$,

$$\begin{aligned} Q_{(X,Y)}(C \times D) &= P((X, Y) \in C \times D) \\ &= P(X \in C, Y \in D) \\ &= P(X \in C)P(Y \in D) \\ &= Q_X(C)Q_Y(D), \end{aligned}$$

et ceci implique que $Q_{(X,Y)}$ est la mesure produit de Q_X et Q_Y .

En particulier, le théorème de Fubini–Tonelli donne, dans un cas très général, le *théorème du produit des espérances* que nous avons vu dans les cas particuliers des variables discrètes et des vecteurs avec une densité de probabilité.

Théorème 6.3.9 *Soit X et Y deux éléments aléatoires à valeurs dans (E, \mathcal{E}) et (G, \mathcal{G}) respectivement. Alors, pour tout $g : (E, \mathcal{E}) \rightarrow (\mathbb{R}, \mathcal{B})$ et tout $h : (G, \mathcal{G}) \rightarrow (\mathbb{R}, \mathcal{B})$ tels que $E[|g(X)|] < \infty$ et $E[|h(Y)|] < \infty$, ou tels que $g \geq 0$ et $h \geq 0$, on a la formule produit pour les espérances*

$$E[g(X)h(Y)] = E[g(X)] E[h(Y)] .$$

Démonstration. Si g et h sont non négatives, en utilisant le théorème de Tonelli :

$$\begin{aligned} E[g(X)h(Y)] &= \int_{\mathbb{R}} \int_{\mathbb{R}} g(x)h(y)Q_{(X,Y)}(dx \times dy) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} g(x)h(y)Q_X(dx)Q_Y(dy) \\ &= \left(\int_{\mathbb{R}} g(x)Q_X(dx) \right) \left(\int_{\mathbb{R}} h(y)Q_Y(dy) \right) \\ &= E[g(X)]E[h(Y)]. \end{aligned}$$

Dans le cas général,

$$E[|g(X)||h(Y)|] = E[|g(X)|] E[|h(Y)|] ,$$

et donc chacun des membres est fini dès que l'autre l'est. Dans ce cas les espérances du type $E[g^+(X)]$ ou $E[h^-(Y)]$ sont finies, et les calculs qui suivent sont valables, ne

recéant aucune forme indéterminée $+\infty - \infty$:

$$\begin{aligned}
 E[g(X)h(Y)] &= E[(g^+(X) - g^-(X))(h^+(Y) - h^-(Y))] \\
 &= E[g^+(X)h^+(Y)] - E[g^+(X)h^-(Y)] - E[g^-(X)h^+(Y)] + E[g^-(X)h^-(Y)] \\
 &= E[g^+(X)]E[h^+(Y)] - E[g^+(X)]E[h^-(Y)] - E[g^-(X)]E[h^+(Y)] \\
 &\quad + E[g^-(X)]E[h^-(Y)] \\
 &= E[g(X)]E[h(Y)].
 \end{aligned}$$

□

Une extension de la théorie élémentaire de l'espérance conditionnelle

Soit X et Y deux vecteurs aléatoires de dimensions p et n respectivement, admettant une densité de probabilité jointe $f_{X,Y}(x,y)$. Soit $g : \mathbb{R}^p \times \mathbb{R}^n \rightarrow \mathbb{R}_+$ une fonction borélienne non négative (*resp.*, telle que $g(X,Y)$ est intégrable). On a vu (Théorème 4.1.6) que pour toute fonction mesurable non négative (*resp.*, bornée) $v : \mathbb{R}^n \rightarrow \mathbb{R}$ on a $E^Y[g(X,Y)v(Y)] = E^Y[g(X,Y)]v(Y)$, et donc, en prenant les espérances des deux membres de cette égalité :

$$E[E^Y[g(X,Y)]v(Y)] = E[g(X,Y)v(Y)]. \quad (6.19)$$

Inspiré par ce résultat, nous allons définir l'espérance conditionnelle d'une variable aléatoire Z , pas nécessairement de la forme $g(X,Y)$, prenant ses valeurs dans $\overline{\mathbb{R}}$.

Définition 6.3.6 Soit Z et Y comme ci-dessus, où, de plus, Z est non négative, (*resp.* intégrable). L'espérance conditionnelle $E^Y[Z]$ est, par définition la variable de la forme $\psi(Y)$ où $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ est mesurable non négative (*resp.* intégrable) et telle que

$$E[\psi(Y)v(Y)] = E[Zv(Y)] \quad (6.20)$$

pour toute fonction mesurable bornée non négative $v : \mathbb{R}^n \rightarrow \mathbb{R}$.

Théorème 6.3.10 Dans la situation de la définition ci-dessus, l'espérance conditionnelle existe et est essentiellement unique. (Par "essentiellement unique", on entend la chose suivante : Si ψ_1 et ψ_2 remplissent les conditions, alors $\psi_1(Y) = \psi_2(Y)$ P -presque sûrement.)

Démonstration. On omet la preuve d'existence. En pratique, il suffit de trouver "une" fonction ψ par construction. L'unicité (que nous allons démontrer) garantit alors que c'est "essentiellement" "la" fonction ψ . Soit donc deux fonctions candidates ψ_1 et ψ_2 . En particulier, $E[\psi_1(Y)v(Y)] = E[\psi_2(Y)v(Y)] (= E[Zv(Y)])$, ou $E[(\psi_1(Y) - \psi_2(Y))v(Y)] = 0$, pour toute fonction mesurable bornée non négative $v : \mathbb{R}^n \rightarrow \mathbb{R}$. Prenons $v(Y) = 1_{\{\psi_1(Y) - \psi_2(Y) > 0\}}$ pour obtenir

$$E[(\psi_1(Y) - \psi_2(Y))1_{\{\psi_1(Y) - \psi_2(Y) > 0\}}] = 0.$$

La variable non négative $(\psi_1(Y) - \psi_2(Y))1_{\{\psi_1(Y) - \psi_2(Y) > 0\}}$ a une espérance nulle. Elle est donc presque sûrement nulle. En d'autres termes, $\psi_1(Y) - \psi_2(Y) \leq 0$, P -p.s. En échangeant les rôles de ψ_1 et ψ_2 , on obtient que $\psi_1(Y) - \psi_2(Y) \geq 0$, P -p.s. Donc, finalement, $\psi_1(Y) - \psi_2(Y) = 0$ P -p.s. \square

Voici deux exemples qui montrent comment cette nouvelle théorie se connecte à la théorie élémentaire.

EXEMPLE 6.3.5: Soit Y une variable à valeurs entières. On veut démontrer que

$$E^Y[Z] = \sum_{n=1}^{\infty} \frac{E[Z1_{\{Y=n\}}]}{P(Y=n)} 1_{\{Y=n\}}, \quad (6.21)$$

où, par convention, $\frac{E[Z1_{\{Y=n\}}]}{P(Y=n)} = 0$ si $P(Y=n) = 0$.

Démonstration. On doit vérifier (6.20) pour toute fonction mesurable $v : \mathbb{R} \rightarrow \mathbb{R}$ non négative et bornée. Le membre de droite est égal à

$$\begin{aligned} & E \left[\left(\sum_{n \geq 1} \frac{E[Z1_{\{Y=n\}}]}{P(Y=n)} 1_{\{Y=n\}} \right) \left(\sum_{k \geq 1} v(k) 1_{\{Y=k\}} \right) \right] \\ &= E \left[\sum_{n \geq 1} \frac{E[Z1_{\{Y=n\}}]}{P(Y=n)} v(n) 1_{\{Y=n\}} \right] \\ &= \sum_{n \geq 1} \frac{E[Z1_{\{Y=n\}}]}{P(Y=n)} v(n) E[1_{\{Y=n\}}] \\ &= \sum_{n \geq 1} \frac{E[Z1_{\{Y=n\}}]}{P(Y=n)} v(n) P(Y=n) \\ &= \sum_{n \geq 1} E[Z1_{\{Y=n\}}] v(n) \\ &= \sum_{n \geq 1} E[Z1_{\{Y=n\}} v(n)] \\ &= E[Z \left(\sum_{n \geq 1} v(n) 1_{\{Y=n\}} \right)] = E[Zv(Y)] \end{aligned}$$

\square

EXEMPLE 6.3.6: Soit X et Y deux vecteurs aléatoires de dimensions p and n respectivement, de densité de probabilité jointe $f_{X,Y}(x,y)$. On rappelle la définition de la densité de probabilité conditionnelle $f_X^{Y=y}(x)$:

$$f_X^{Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

avec la convention $f_X^{Y=y}(x) = 0$ quand $f_Y(y) = 0$. Soit $g : \mathbb{R}^{p+n} \rightarrow \mathbb{R}$ une fonction mesurable telle que $Z = g(X, Y)$ soit intégrable. Une version de l'espérance conditionnelle de Z sachant Y est $\psi(Y)$, où

$$\psi(y) = \int_{\mathbb{R}^p} g(x, y) f_X^{Y=y}(x) dx.$$

Démonstration. On vérifie que $\psi(Y)$ est intégrable. On a en effet :

$$|\psi(y)| \leq \int_{\mathbb{R}^p} |g(x, y)| f_Z^{Y=y}(x) dx,$$

et donc

$$\begin{aligned} E[|\psi(Y)|] &= \int_{\mathbb{R}^n} |\psi(y)| f_Y(y) dy \leq \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^p} |g(x, y)| f_Z^{Y=y}(x) dx \right) f_Y(y) dy \\ &= \int_{\mathbb{R}^p} \int_{\mathbb{R}^n} |g(x, y)| f_X^{Y=y}(x) f_Y(y) dx dy \\ &= \int_{\mathbb{R}^p} \int_{\mathbb{R}^n} |g(x, y)| f_{X,Y}(x, y) dx dy \\ &= E[|g(X, Y)|] = E[|Z|] < \infty. \end{aligned}$$

On vérifie maintenant (6.20), où v est mesurable, non négative et bornée. Le membre de droite est

$$\begin{aligned} E[\psi(Y)v(Y)] &= \int_{\mathbb{R}^n} \psi(y)v(y)f_Y(y)dy \\ &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^p} g(x, y)f_X^{Y=y}(x)dx \right) v(y)f_Y(y)dy \\ &= \int_{\mathbb{R}^p} \int_{\mathbb{R}^n} g(x, y)v(y)f_X^{Y=y}(x)f_Z(y) dx dy \\ &= \int_{\mathbb{R}^p} \int_{\mathbb{R}^n} g(x, y)v(y)f_{X,Y}(x, y) dx dy \\ &= E[g(X, Y)v(Y)] = E[Zv(Y)]. \end{aligned}$$

□

Nous allons revoir les propriétés élémentaires de l'espérance conditionnelle dans ce cadre plus général.

Théorème 6.3.11 *Soit Y un vecteur aléatoire, et soit Z, Z_1, Z_2 des variables intégrables (resp. non négatives), $\lambda_1, \lambda_2 \in \mathbb{R}$ (resp. $\in \mathbb{R}_+$).*

Règle 1. (*linéarité*)

$$E^Y[\lambda_1 Z_1 + \lambda_2 Z_2] = \lambda_1 E^Y[Z_1] + \lambda_2 E^Y[Z_2].$$

Règle 2. Si Z est indépendante de Y , alors

$$E^Y[Z] = E[Z].$$

Règle 3. Si Z est une fonction mesurable de Y ,

$$E^Y[Z] = Z.$$

Règle 4. (*monotonie*) Si $Z_1 \leq Z_2$ P -p.s., alors

$$E^Y[Z_1] \leq E^X[Z_2], P\text{-p.s.}$$

En particulier, si Z est une variable aléatoire non négative, $E^Y[Z] \geq 0$, P -p.s..

Règle 5. (*conditionnements successifs*). Supposons que $Y = (Y_1, Y_2)$, où Y_1 et Y_2 sont des vecteurs aléatoires de dimensions respectives m_1 et m_2 . Alors :

$$E^{Y_1}[E^{(Y_1, Y_2)}[Z]] = E^{Y_1}[Z].$$

Démonstration. On fait le cas intégrable, le cas non négatif est similaire.

Règle 1 : $\lambda_1 E^Y[X_1] + \lambda_2 E^Y[X_2]$ est bien une fonction mesurable de Y , qui est de plus intégrable. De plus, pour toute fonction borélienne $v : \mathbb{R}^n \rightarrow \mathbb{R}$ non négative et bornée

$$E[(\lambda_1 E^Y[X_1] + \lambda_2 E^Y[X_2])v(Y)] = E[(\lambda_1 Y_1 + \lambda_2 Y_2)v(Y)],$$

ce qui découle immédiatement de la définition de $E^Y[X_i]$, selon laquelle $E[E^Y[X_i]v(Y)] = E[Y_i v(Y)]$, $i = 1, 2$.

Règle 2 : La constante $E[Z]$ est bien de la forme $\psi(Y)$. D'autre part, pour toute fonction borélienne $v : \mathbb{R}^n \rightarrow \mathbb{R}$ non négative et bornée, $E[E[Z]v(Y)] = E[Zv(Y)]$, puisque Z et Y sont indépendantes, et en particulier $E[Zv(Y)] = E[Z]E[v(Y)] = E[E[Z]v(Y)]$.

Règle 3 : Y est bien de la forme $\psi(Y)$, et ce qui reste à vérifier est une tautologie : $E[Yv(Y)] = E[Yv(Y)]$.

Règle 4 : Pour toute fonction borélienne $v : \mathbb{R}^n \rightarrow \mathbb{R}$ non négative et bornée,

$$E[E^Y[Y_1]v(Y)] = E[Y_1 v(Y)] \leq E[Y_2 v(Y)] = E[E^Y[Y_2]v(Y)]$$

Donc

$$E[(E^Y[Y_2] - E^Y[Y_1])v(Y)] \geq 0$$

Si on prend $v(Y) = 1_{\{E^Y[Y_2] < E^Y[Y_1]\}}$, on obtient

$$0 \leq E[(E^Y[Y_2] - E^Y[Y_1])1_{\{E^Y[Y_2] < E^Y[Y_1]\}}] \geq 0,$$

et donc

$$E[(E^Y[Y_2] - E^Y[Y_1])1_{\{E^Y[Y_2] < E^Y[Y_1]\}}] \geq 0,$$

ce qui donne le résultat annoncé, d'après (e) du Théorème 6.1.8.

Règle 5 : $E^{Y_1}[E^Y[Z]]$ est une fonction mesurable de Y_1 et donc, *a fortiori*, de Y . Il reste à vérifier que

$$E[E^{Y_1}[E^Y[Z]]v_1(Y_1)] = E[Zv_1(Y_1)],$$

pour toute fonction borélienne $v_1 : \mathbb{R}^{m_1} \rightarrow \mathbb{R}$ non négative et bornée. On a

$$E[[E^Y[Z]v_1(Y_1)] = E[Zv_1(Y_1)].$$

En effet, $v_1(Y_1)$ est bien de la forme $v(Y)$. De plus,

$$E[E^{Y_1}[E^Y[Z]]v(Y)] = E[[E^Y[Z]v(Y)],$$

par définition de $E^{Y_1}[E^Y[Z]]$. □

Théorème 6.3.12 Soit Z une variable aléatoire de la forme $Z = c(Y)X$, où c est une fonction mesurable bornée (resp. non négative), et Z est une variable aléatoire intégrable (resp. non négative). Alors

$$E^Y[c(Y)Z] = c(Y)E^Y[Z]$$

Démonstration. On fait la démonstration dans le cas intégrable, le cas non négatif étant similaire. On observe d'abord que $c(Y)E^Y[Z]$ est de la forme $\psi(Y)$, où ψ est mesurable. Comme $c(Y)$ est bornée et $E^Y[Z]$ intégrable, $\psi(Y)$ est intégrable. Il reste à prouver que pour toute fonction borélienne $v : \mathbb{R}^n \rightarrow \mathbb{R}$ non négative et bornée,

$$E[c(Y)Zv(Y)] = E[c(Y)E^Y[Z]v(Y)].$$

Mais comme $c(Y)v(Y)$ est bornée, on a, par définition de $E^Y[Z]$,

$$E[E^Y[Z]c(Y)v(Y)] = E[Zc(Y)v(Y)].$$

□

6.4 Exercices

Exercice 6.4.1. MESURABILITÉ PAR RAPPORT À LA TRIBU GROSSIÈRE.

Montrez qu'une fonction mesurable par rapport à la tribu grossière (celle dont les deux seuls éléments sont l'ensemble tout entier et l'ensemble vide) est une constante.

Exercice 6.4.2. *

Soit X et E des ensembles quelconques, $f : X \rightarrow E$ une fonction de X dans E , \mathcal{C} une famille non vide arbitraire de sous-ensembles de E . Démontrez que

$$\sigma(f^{-1}(\mathcal{C})) = f^{-1}(\sigma(\mathcal{C})).$$

Exercice 6.4.3.

Soit $f : X \rightarrow E$ une fonction. Soit \mathcal{E} une tribu sur E . Quelle est la plus petite tribu \mathcal{X} sur X telle que f soit mesurable par rapport à \mathcal{X} et \mathcal{E} ?

Exercice 6.4.4.

Soit $f : X \rightarrow E$ une fonction. Est-il vrai que si $|f|$ est mesurable par rapport à \mathcal{X} et \mathcal{E} , il en est de même de f ?

Exercice 6.4.5.

Dans l'énoncé de la propriété de continuité séquentielle pour les suites décroissantes (Partie (a) du Théorème 6.1.4), montrez à l'aide d'un contre-exemple la nécessité de la condition $\mu(B_{n_0}) < \infty$ pour un n_0 .

Exercice 6.4.6.

Démontrez les propriétés (e), (f) et (g) du Théorème 6.1.8.

Exercice 6.4.7. LEMME DE SCHEFFÉ.

Soit f et f_n , $n \geq 1$ des fonctions mesurables de (X, \mathcal{X}) dans $(\mathbb{R}, \mathcal{B})$, non négatives et μ -intégrables, et telles que $\lim_{n \uparrow \infty} f_n = f$ μ -p.p. et $\lim_{n \uparrow \infty} \int_X f_n d\mu = \int_X f d\mu$. Montrez que $\lim_{n \uparrow \infty} \int_X |f_n - f| d\mu = 0$. (Indication : $|a - b| = a + b - \inf(a, b)$.)

Exercice 6.4.8.

Montrez que pour tout $a, b \in \mathbb{R}$,

$$\int_{\mathbb{R}_+} \frac{t e^{-at}}{1 - e^{-bt}} dt = \sum_{n=0}^{+\infty} \frac{1}{(a + nb)^2}.$$

Exercice 6.4.9.

Démontrez que la fonction d'ensembles $\mu \circ h^{-1}$ de la Définition 6.2.1 est bien une mesure.

Exercice 6.4.10.

Démontrez que la fonction d'ensembles ν de la Définition 6.2.2 est bien une mesure.

Exercice 6.4.11.

Démontrez que si deux fonctions continues $f, g : \mathbb{R} \rightarrow \mathbb{R}$ sont ℓ -presque partout égales, elles sont en fait partout égales.

Exercice 6.4.12.

Démontrez l'unicité de la mesure produit dans le Théorème 6.2.5.

Exercice 6.4.13.

Énoncez le théorème de Tonelli–Fubini pour le produit de deux mesures de comptage sur \mathbb{Z} .

Exercice 6.4.14. LE CARRELAGE.

On dit qu'un rectangle $[a, b] \times [c, d]$ de \mathbb{R}^2 a la propriété (A) si *au moins un* de ses côtés a une longueur égale à un entier. On vous donne un rectangle Δ qui est la réunion d'un nombre fini de rectangles disjoints (à part leurs frontières qui peuvent se chevaucher) ayant la propriété (A). Montrez que Δ a la propriété (A). (Indication : $\int_a^b e^{2i\pi x} dx \dots$)

Exercice 6.4.15. TRANSFORMÉE DE FOURIER.

Soit $f : (\mathbb{R}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B})$ une fonction intégrable par rapport à la mesure de Lebesgue. Montrez que pour tout $\nu \in \mathbb{R}$,

$$\hat{f}(\nu) = \int_{\mathbb{R}} f(t) e^{-2i\pi \nu t} dt$$

est bien défini, et que la fonction \hat{f} est continue et bornée uniformément. (\hat{f} est la *transformée de Fourier* de f).

Exercice 6.4.16. CONVOLUTION.

Soit $f, g : (\mathbb{R}, \mathcal{B}) \rightarrow (\mathbb{R}, \mathcal{B})$ deux fonctions intégrables par rapport à la mesure de Lebesgue, et soit \hat{f}, \hat{g} leurs transformées de Fourier (Exercice 6.4.15).

1. Montrez que

$$\int_{\mathbb{R}} \int_{\mathbb{R}} |f(t-s)g(s)| dt ds < \infty.$$

2. Dédurre de cela que pour presque-tout (pour la mesure de Lebesgue) $t \in \mathbb{R}$, la fonction $s \mapsto f(t-s)g(s)$ est ℓ -intégrable, et donc, que la convolution $f * g$, où

$$(f * g)(t) = \int_{\mathbb{R}} f(t-s)g(s) ds$$

est presque-partout bien définie.

3. Pour tout t tel que la dernière intégrale n'est pas bien définie, on pose $(f * g)(t) = 0$. Montrez que $f * g$ est ℓ -intégrable et que sa transformée de Fourier est $\widehat{f * g} = \hat{f}\hat{g}$.

Exercice 6.4.17. CONTRE-EXEMPLE FUBINI.

Soit $f : [0, 1]^2 \rightarrow \mathbb{R}$ définie par

$$f(x, y) = \frac{x^2 - y^2}{(x^2 + y^2)^2} 1_{\{(x, y) \neq (0, 0)\}}$$

Calculez $\int_{[0, 1]} \left(\int_{[0, 1]} f(x, y) dx \right) dy$ et $\int_{[0, 1]} \left(\int_{[0, 1]} f(x, y) dy \right) dx$. Est-ce que f est intégrable (relativement à la mesure de Lebesgue) sur $[0, 1]^2$?

Exercice 6.4.18.

Démontrez le Théorème 3.1.2.

Exercice 6.4.19. TRANSFORMÉE DE LAPLACE.

Soit X une variable aléatoire non négative. Démontrez que

$$\lim_{\theta \uparrow \infty; \theta > 0} E \left[e^{-\theta X} \right] = P(X = 0).$$

Exercice 6.4.20. FORMULE DU TÉLÉSCOPE.

Démontrez que pour toute variable aléatoire non négative X , l'application $(x, \omega) \rightarrow 1_{\{X(\omega) > x\}}$ est mesurable par rapport à $(\mathbb{R}_+ \times \Omega, \mathcal{B}(\mathbb{R}_+) \times \mathcal{F})$ et $(\mathbb{R}, \mathcal{B})$. Démontrez la formule du télescope :

$$E[X] = \int_0^\infty [1 - F(x)] dx.$$

Exercice 6.4.21. *

Soit X une variable aléatoire non négative, et soit $G : \mathbb{R}_+ \rightarrow \mathbb{C}$ une fonction primitive de $g : \mathbb{R}_+ \rightarrow \mathbb{C}$, en ce sens que : Pour tout $x \geq 0$,

$$G(x) = G(0) + \int_0^x g(u) du.$$

(a) Soit X une variable aléatoire non négative de moyenne finie μ et telle que $E[G(X)] < \infty$. Montrez que

$$E[G(X)] = G(0) + \int_0^\infty g(x) P(X > x) dx.$$

(b) Soit X comme dans (a), et soit \bar{X} une variable aléatoire non négative de densité de probabilité

$$\mu^{-1} P(X \geq x).$$

Montrez que

$$E \left[e^{iuX} \right] = \frac{E \left[e^{iuX} \right] - 1}{i\mu u}.$$

Exercice 6.4.22.

Soit X un vecteur aléatoire de \mathbb{R}^d de densité de probabilité f . Montrez que $P(f(X) = 0) = 0$.

Exercice 6.4.23.

Soit X une variable réelle de densité de probabilité f_X . Soit $h : \mathbb{R} \rightarrow \mathbb{R}$ une fonction mesurable telle que $h(X)$ est intégrable. Démontrez que

$$E[h(X)|X^2] = h(\sqrt{X^2}) \frac{f_X(\sqrt{X^2})}{f_X(\sqrt{X^2}) + f_X(-\sqrt{X^2})} + h(-\sqrt{X^2}) \frac{f_X(-\sqrt{X^2})}{f_X(\sqrt{X^2}) + f_X(-\sqrt{X^2})}.$$

Exercice 6.4.24. *

Soit X une variable aléatoire réelle intégrable. Posons $X^+ = \max(X, 0)$, $X^- = \max(-X, 0)$. Montrez que

$$E[X|X^+] = X^+ - \frac{E[X^-]}{P(X^+ = 0)} 1_{\{X^+ = 0\}}.$$

Chapitre 7

Suites de variables aléatoires

Ce chapitre est consacré aux diverses notions de convergence d'une suite de variables aléatoires, principalement la convergence presque-sûre et la convergence en distribution, mais aussi la convergence en probabilité et la convergence en moyenne quadratique. Le lecteur y trouvera les deux piliers de la théorie des probabilités et de la statistique : la loi forte des grands nombres et le théorème de la limite gaussienne.

7.1 Convergence presque-sûre

La loi forte des grands nombres

La loi forte des grands nombres est essentielle à la théorie des probabilités, car elle justifie *a posteriori* son axiomatique en la reliant à la notion intuitive de fréquence et donne une explication rationnelle à la présence d'un certain ordre macroscopique dans le chaos microscopique.

Voici la version due à Kolmogorov (1933) :

Théorème 7.1.1 *Soit $(X_n, n \geq 1)$ une suite de variables aléatoires définies sur le même espace de probabilité (Ω, \mathcal{F}, P) , indépendantes et identiquement distribuées (IID), et intégrables. Alors la moyenne empirique $S_n(\omega)/n = (X_1(\omega) + \dots + X_n(\omega))/n$ converge vers $E[X_1]$ pour tout $\omega \notin \mathcal{N}$, où \mathcal{N} est un ensemble de probabilité nulle.*

On note ceci

$$\lim_{n \uparrow \infty} \frac{\sum_{n=1}^{\infty} X_n}{n} = E[X_1], \quad \text{P-p.s.} \quad (\star)$$

En particulier, si $(X_n, n \geq 1)$ est une suite quelconque de variables aléatoires IID et A un sous-ensemble de \mathbb{R} , la suite $(1_A(X_n), n \geq 1)$ est IID, de moyenne $P(A)$, et donc

$$\lim_{n \uparrow \infty} \frac{\sum_{n=1}^{\infty} 1_A(X_n)}{n} = P(X_1 \in A), \quad \text{P-p.s.}$$

C'est la loi forte des grands nombres d'Émile Borel (1909) : la probabilité est la fréquence empirique asymptotique.

Lemme de Borel–Cantelli

Nous aurons besoin d'un certain nombre de définitions et de préliminaires techniques qui font intervenir les suites d'événements.

Soit une suite $(A_n, n \geq 1)$ de sous-ensembles de Ω . Formons pour chaque $n \geq 1$, le sous-ensemble

$$B_n = \bigcup_{p=n}^{\infty} A_p.$$

Lorsque n croît, B_n décroît, et on peut donc parler de la limite (décroissante) de la suite $(B_n, n \geq 1)$, égale, par définition, à $\bigcap_{n=1}^{\infty} B_n$. C'est cette limite qu'on appelle la limite supérieure de la suite $(A_n, n \geq 1)$:

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{p=n}^{\infty} A_p.$$

L'interprétation de cet ensemble est la suivante :

Théorème 7.1.2 *Pour que $\omega \in \limsup_n A_n$, il faut et il suffit qu'il existe une infinité d'indices p tels que $\omega \in A_p$.*

Démonstration. Exercice 7.4.1. □

Dire que l'événement $\limsup_n A_n$ est réalisé par l'épreuve ω équivaut à dire que les événements A_n se réalisent une infinité de fois pour l'épreuve ω . Dans la littérature probabiliste de langue anglaise, on utilise pour cet ensemble l'écriture suggestive $\{A_n \text{ i.o.}\}$ (“i.o.” est une abréviation de “infinitely often”).

Le *lemme de Borel–Cantelli* est un des résultats les plus importants de la théorie des probabilités. Il donne une condition suffisante pour que, presque sûrement, l'événement A_n ne se produise qu'un nombre fini de fois.

Théorème 7.1.3 *Soit (Ω, \mathcal{F}, P) un espace de probabilité et $(A_n, n \geq 1)$ une suite quelconque d'événements de \mathcal{F} . On a l'implication :*

$$\sum_{n=1}^{\infty} P(A_n) < \infty \Rightarrow P(\limsup_n A_n) = 0.$$

Démonstration. La propriété de continuité séquentielle de la probabilité donne

$$P(\limsup_n A_n) = P\left(\bigcap_{n=1}^{\infty} \bigcup_{p=n}^{\infty} A_p\right) = \lim P\left(\bigcup_{p=n}^{\infty} A_p\right).$$

De plus (sous-sigma-additivité de la probabilité) :

$$P\left(\bigcup_{p=n}^{\infty} A_p\right) \leq \sum_{p=n}^{\infty} P(A_p).$$

En tant que reste d'une série convergente, le membre de droite de l'inégalité précédente tend vers 0, et donc $P(\limsup_n A_n) = 0$. \square

Voici maintenant une *reciproque partielle du lemme de Borel–Cantelli*. Elle ne s'applique que pour une suite d'événements indépendants.

Théorème 7.1.4 *Soit (Ω, \mathcal{F}, P) un espace de probabilité et $(A_n, n \geq 1)$ une suite d'événements de \mathcal{F} indépendants. On a l'implication :*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \Rightarrow P(\limsup_n A_n) = 1.$$

Démonstration. Le produit infini

$$\prod_{k=n}^{\infty} (1 - P(A_k))$$

est nul puisque la série $\sum_{k=n}^{\infty} P(A_k)$ diverge. On a donc, en utilisant l'indépendance des événements A_k ,

$$\prod_{k=n}^{\infty} (1 - P(A_k)) = \prod_{k=n}^{\infty} P(\bar{A}_k) = P\left(\bigcap_{k=n}^{\infty} \bar{A}_k\right) = 0,$$

ou encore, en passant au complémentaire,

$$P\left(\bigcup_{k=n}^{\infty} A_k\right) = 1.$$

On a donc (continuité séquentielle de la probabilité),

$$\lim_{n \uparrow \infty} P\left(\bigcup_{k=n}^{\infty} A_k\right) = P\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right) = 1.$$

\square

Les théorèmes 7.1.3 et 7.1.4 donnent :

Théorème 7.1.5 *Pour une suite $(A_n, n \geq 1)$ d'événements indépendants, $\limsup_n A_n$ est un événement de probabilité 0 ou 1 selon que la somme $\sum_{n=1}^{\infty} P(A_n)$ est finie ou infinie.*

Convergence presque sûre

Définition 7.1.1 Soit $(X_n, n \geq 1)$ une suite de variables aléatoires définies sur le même espace de probabilité (Ω, \mathcal{F}, P) et soit X une autre variable aléatoire sur ce même espace. On dit que la suite $(X_n, n \geq 1)$ converge presque sûrement vers X si et seulement s'il existe un événement \mathcal{N} de probabilité nulle tel que

$$\omega \notin \mathcal{N} \Rightarrow \lim_{n \uparrow \infty} X_n(\omega) = X(\omega) .$$

On note ceci de différentes manières :

$$X_n \xrightarrow{\text{p.s.}} X, \quad X_n \longrightarrow X, \text{ p.s.}, \quad \lim_{n \uparrow \infty} X_n = X, \text{ p.s.}.$$

Voici un critère de convergence presque-sûre :

Théorème 7.1.6 La suite de variables aléatoires $(X_n, n \geq 1)$ converge presque sûrement vers la variable aléatoire réelle X si et seulement si :

$$\text{pour tout } \varepsilon > 0, \quad P(\limsup_{n \uparrow \infty} \{|X_n - X| \geq \varepsilon\}) = 0.$$

Démonstration. La condition est nécessaire. En effet, soit \mathcal{N} l'ensemble négligeable de la définition Définition 7.1.1. Pour tout ω hors de \mathcal{N} , le nombre d'indices n tels que $|X_n(\omega) - X(\omega)| \geq \varepsilon$ est fini. De l'interprétation de l'ensemble "lim sup" donné par le Théorème 7.1.2, on déduit que $\limsup_n \{|X_n - X| \geq \varepsilon\} \subseteq \mathcal{N}$.

La condition est suffisante. Introduisons les événements $\mathcal{N}_k = \limsup_n \{|X_n - X| \geq \frac{1}{k}\}$. Par hypothèse, $P(\mathcal{N}_k) = 0$ pour tout $k \geq 1$. Si on définit $\mathcal{N} = \cup_{k=1}^{\infty} \mathcal{N}_k$, on a bien $P(\mathcal{N}) = 0$ (sous sigma-additivité de la fonction probabilité). Reste à montrer que si $\omega \notin \mathcal{N}$ alors $\lim X_n(\omega) = X(\omega)$. On encore, de façon équivalente : si $\omega \notin \mathcal{N}$, pour tout $\varepsilon > 0$, il n'existe qu'un nombre fini de n tels que $|X_n(\omega) - X(\omega)| \geq \varepsilon$. En effet, choisissons $k \geq 1$ tel que $\varepsilon > \frac{1}{k}$. Comme $\omega \notin \mathcal{N}$ et que $\mathcal{N}_k \subset \mathcal{N}$, alors $\omega \notin \mathcal{N}_k$. D'après la définition de \mathcal{N}_k et l'interprétation de lim sup, ω est donc tel que $|X_n(\omega) - X(\omega)| \geq \frac{1}{k}$ n'a lieu que pour un nombre fini de n . A fortiori $|X_n(\omega) - X(\omega)| \geq \varepsilon$ n'a lieu que pour un nombre fini de n . \square

Le lemme de Borel–Cantelli est à la base de la très utile condition suffisante de convergence presque sûre suivante.

Théorème 7.1.7 Pour que la suite $(X_n, n \geq 1)$ de variables aléatoires converge presque-sûrement vers la variable aléatoire réelle X , il suffit que

$$\sum_{n=0}^{\infty} P(|X_n - X| \geq \varepsilon_n) < \infty ,$$

où $(\varepsilon_k, k \geq 1)$ est une suite de nombres réels positifs tendant vers 0.

Démonstration. De l'hypothèse et du lemme de Borel–Cantelli, on déduit qu'il existe un ensemble \mathcal{N} de probabilité nulle et tel que pour tout $\omega \notin \mathcal{N}$, $|X_n(\omega) - X(\omega)| \geq \varepsilon_n$ seulement pour un nombre fini d'indices. Donc, à partir d'un certain rang $N(\omega)$, $|X_n(\omega) - X(\omega)| < \varepsilon_n$. \square

EXEMPLE 7.1.1: LA LOI FORTE DE BOREL. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes définies sur l'espace de probabilité (Ω, \mathcal{F}, P) , à valeurs dans $\{0, 1\}$ et telles que

$$P(X_n = 1) = p.$$

Alors

$$\frac{S_n}{n} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{p.s.} p.$$

(Ici, le nombre p figure la variable aléatoire qui prend la seule valeur p .)

Démonstration. D'après le Théorème 7.1.6, il suffit de démontrer que pour tout $\varepsilon > 0$,

$$P\left(\limsup_n \left\{\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right\}\right) = 0.$$

Mais ceci, d'après le lemme de Borel-Cantelli, est une conséquence de

$$\sum_{n=1}^{\infty} P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) < \infty.$$

Cette dernière convergence découle de la majoration

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{1}{4\varepsilon^4} \frac{1}{n^2},$$

objet de l'Exercice 7.4.5. \square

EXEMPLE 7.1.2: RÉCRÉATION : LES SUITES DE BOREL. On considère une suite de Bernoulli $(X_n, n \geq 1)$, c'est-à-dire une suite de variables aléatoires indépendantes et identiquement distribuées, prenant les valeurs 0 ou 1 avec $P(X_n = 1) = p$. On appelle "cache" toute suite strictement croissante d'entiers positifs $(n_i, 1 \leq i \leq k)$. Soit $(\varepsilon_i, 1 \leq i \leq k)$ une suite de 0 et de 1. L'ensemble $(n_i, \varepsilon_i, 1 \leq i \leq k)$ est appelé un *motif*. On définit pour $n \geq 1$ la variable aléatoire

$$Y_n = \begin{cases} 1 & \text{si } X_{n+n_i} = \varepsilon_i \text{ pour tout } i \quad (1 \leq i \leq k) \\ 0 & \text{autrement.} \end{cases}$$

On a :

$$\frac{Y_1 + \dots + Y_n}{n} \xrightarrow{p.s.} p^h q^{k-h} \quad \text{où } h = \sum_{i=1}^k \varepsilon_i. \quad (\star)$$

La démonstration sera faite, pour simplifier les notations, dans le cas où $\varepsilon_i = 1$, $1 \leq i \leq k$. Définissons pour tout i tel que $1 \leq i \leq n_k$, $Y_n^{(i)} = Y_{i+n \times n_k}$. Pour i fixé la suite $(Y_n^{(i)}, n \geq 1)$ est IID et $P(Y_1^{(i)} = 1) = P(X_{i+n_1} = 1, \dots, X_{i+n_k} = 1) = p^k$. On a donc d'après la loi des grands nombres $\lim_{N \uparrow \infty} (1/N) \sum_{n=1}^N Y_n^{(i)} = p^k$. Ceci étant vrai pour tout i ($1 \leq i \leq n_k$), le résultat annoncé en découle.

Comme la suite de Bernoulli avec $p = \frac{1}{2}$, qui est la suite aléatoire par excellence, satisfait à (\star) , avec $p = \frac{1}{2}$, pour tous les motifs possibles, Borel a eu l'idée de définir une suite *déterministe* de 0 et de 1, soit $(x_n, n \geq 1)$, comme *parfaitement aléatoire*, si pour tout motif,

$$\lim_{n \uparrow \infty} \frac{y_1 + \dots + y_n}{n} = \frac{1}{2^k},$$

où les y_n sont définis comme les Y_n . Cette définition, qui a l'air raisonnable, intuitivement, n'est cependant pas satisfaisante. En effet, on peut montrer que la *suite de Champernowne* :

$$0110111001011101111000 \dots,$$

qui est l'écriture de la suite des nombres entiers en binaire (en commençant par 0), est aléatoire au sens de Borel!

Loi des grands nombres pour les suites non-corrélées

La loi forte des grands nombres qui suit est plus simple à démontrer que celle de Kolmogorov, mais elle requiert l'existence des seconds moments. D'un autre côté, elle n'exige pas l'indépendance, seulement la non-corrélation.

Théorème 7.1.8 *Soit $(X_n, n \geq 1)$ une suite de variables aléatoires intégrables identiquement distribuées de variance finie, et non-corrélées. On a alors :*

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{p.s.} E[X_1].$$

Démonstration. Ce sont encore l'inégalité de Markov et le lemme de Borel–Cantelli qui vont nous servir. On suppose $E[X_1] = 0$, ce qui ne restreint pas la généralité de la démonstration.

L'idée générale est la suivante : pour un entier positif m on considère les variables aléatoires

$$S_{m^2} = X_1 + X_2 + \dots + X_{m^2}$$

(où $S_n = X_1 + X_2 + \dots + X_n$) et

$$Z_m = \sup_{1 \leq k \leq 2m+1} (|X_{m^2+1} + \dots + X_{m^2+k}|).$$

On démontrera bientôt que, lorsque m tend vers l'infini,

$$\frac{S_{m^2}}{m^2} \xrightarrow{p.s.} 0 \tag{7.1}$$

et

$$\frac{Z_m}{m^2} \xrightarrow{p.s.} 0 . \quad (7.2)$$

Cela admis, on utilise la majoration suivante :

$$\left| \frac{S_n}{n} \right| \leq \left| \frac{S_{m(n)^2}}{m(n)^2} \right| + \left| \frac{Z_{m(n)}}{m(n)^2} \right| ,$$

où $m(n)$ est l'entier m tel que

$$m^2 < n \leq (m+1)^2 .$$

L'entier $m(n)$ tend vers l'infini avec n . Le résultat annoncé découle de cette observation et de (7.1) et (7.2).

Reste à démontrer les convergences annoncées. Commençons par (7.1). Pour un $\varepsilon > 0$ quelconque, l'inégalité de Tchebychev donne

$$P(|S_{m^2}/m^2| \geq \varepsilon) \leq \frac{\text{Var}(S_{m^2})}{m^4 \varepsilon^2} = \frac{m^2 \sigma^2}{m^4 \varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \frac{1}{m^2} .$$

Donc, en prenant $\varepsilon = m^{-\frac{1}{4}}$

$$P(|S_{m^2}/m^2| \geq m^{-\frac{1}{4}}) \leq \frac{\sigma^2}{m^{\frac{3}{2}}} .$$

Le Théorème 7.1.7 permet de conclure.

Pour (7.2), notons d'abord que

$$P\left(\left|\frac{Z_m}{m^2}\right| \geq \varepsilon\right) = P(|Z_m| \geq m^2 \varepsilon) \leq P\left(\bigcup_{k=1}^{2m+1} \{|\xi_k| \geq m^2 \varepsilon\}\right)$$

où $|\xi_k| = |X_{m^2+1} + \dots + X_{m^2+k}|$ (la dépendance en m de cette variable aléatoire n'est pas mentionnée, pour les alléger les notations). Mais d'après la sous-additivité de la fonction probabilité,

$$P\left(\bigcup_{k=1}^{2m+1} \{|\xi_k| \geq m^2 \varepsilon\}\right) \leq \sum_{k=1}^{2m+1} P(|\xi_k| \geq m^2 \varepsilon) .$$

L'inégalité de Tchebychev nous donne d'autre part

$$P(|\xi_k| \geq m^2 \varepsilon) \leq \frac{\text{Var}(\xi_k)}{m^4 \varepsilon^2} \leq \frac{(2m+1)\sigma^2}{m^4 \varepsilon^2} .$$

puisque ξ_k est la somme de moins de $2m+1$ variables aléatoires X_i différentes. En combinant les trois dernières inégalités, on voit que

$$P\left(\left|\frac{Z_m}{m^2}\right| \geq \varepsilon\right) \leq \frac{\sigma^2 (2m+1)^2}{\varepsilon^2 m^4} \leq C \frac{\sigma^2}{\varepsilon^2} \frac{1}{m^2} ,$$

pour une constante $C < \infty$. Le reste suit comme dans la démonstration de (7.1). \square

Grandes déviations à la loi des grands nombres

Voici les *inégalités de Chernoff* qui nous serviront à produire une nouvelle version de la loi forte des grands nombres. Cette version est plus restrictive que celle de Kolmogorov et même que celle du Théorème 7.1.8, l'intérêt résidant surtout dans l'estimation assez fine des grands écarts à la loi des grands nombres que permettent les inégalités de Chernoff.

Théorème 7.1.9 *Soit X une variable aléatoire réelle telle que pour tout $t \in \mathbb{R}$:*

$$\phi(t) = E[e^{tX}] < \infty. \quad (7.3)$$

On a pour tout $a \geq 0$,

$$P(X \geq a) \leq e^{-h(a)} \quad (a \geq 0) \quad (7.4)$$

et

$$P(X \leq -a) \leq e^{-h(-a)} \quad (a \geq 0), \quad (7.5)$$

où la fonction h est la transformée de Crámer de X , définie par :

$$h(a) = \begin{cases} \sup_{t \geq 0} (at - \log \phi(t)) & \text{si } a \geq 0 \\ \sup_{t \leq 0} (-at - \log \phi(t)) & \text{si } a \leq 0. \end{cases} \quad (7.6)$$

Démonstration. Fixons $t \geq 0$ et $a \geq 0$. De la majoration

$$e^{t(X-a)} \geq 1_{\{X \geq a\}},$$

on tire la majoration

$$E[e^{t(X-a)}] \geq P(X \geq a).$$

Or

$$E[e^{t(X-a)}] = E[e^{tX}]e^{-ta} = e^{-(at-\psi(t))}.$$

On a donc, pour $t \geq 0$, $a \geq 0$:

$$P(X \geq a) \leq e^{-(at-\psi(t))},$$

et donc

$$P(X \geq a) \leq e^{-\sup_{t \geq 0} (at-\psi(t))} = e^{-h(a)}.$$

Pour démontrer (7.5) on procède de manière analogue. On fixe $a \geq 0$ et $t \leq 0$. On a alors la majoration

$$\begin{aligned} E[e^{t(X+a)}] &\geq E[1_{\{X \leq -a\}}] \\ &= P(X \leq -a). \end{aligned}$$

D'où

$$P(X \leq -a) \leq E[e^{t(X+a)}] = e^{at+\psi(t)} = e^{-(-at-\psi(t))},$$

et comme t est un nombre négatif arbitraire,

$$P(X \leq -a) \leq e^{-\sup_{t \leq 0} (-at - \psi(t))} = e^{-h(-a)} .$$

□

La transformée de Cr mer a une propri t  remarquable qui en fait tout l'int r t.

Th or me 7.1.10 *Si $(X_n, n \geq 1)$ est une suite de variables al atoires ind pendantes et de m me loi que la variable al atoire X du Th or me 7.1.9, alors la transform e de Cramer $h_n(a)$ de la moyenne empirique $\overline{X}_n = \frac{X_1 + \dots + X_n}{n}$ et la transform e de Cramer $h(a)$ de X sont reli es par la relation*

$$h_n(a) = n h(a) .$$

D monstration. Si on appelle $\phi_n(t)$ la fonction g n ratrice de \overline{X}_n , alors en utilisant l'hypoth se d'ind pendance

$$\phi_n(t) = E[e^{t \frac{X_1 + \dots + X_n}{n}}] = E[e^{(\frac{t}{n} X_1 + \dots + \frac{t}{n} X_n)}] = \prod_{i=1}^n E[e^{\frac{t}{n} X_i}] .$$

On a donc

$$\phi_n(t) = \left\{ \phi\left(\frac{t}{n}\right) \right\}^n$$

et

$$\psi_n(t) = n \psi\left(\frac{t}{n}\right) ,$$

o  $\psi_n(t) = \log \phi_n(t)$. Pour $a \geq 0$

$$h_n(a) = \sup_{t \geq 0} \left(at - n \psi\left(\frac{t}{n}\right) \right) = n \sup_{t \geq 0} \left(a \frac{t}{n} - \psi\left(\frac{t}{n}\right) \right) = n h(a) ,$$

et la m me relation vaut pour $a \leq 0$, avec la m me d monstration. □

Les in galit s de Chernoff (7.4) et (7.5) appliqu es   $\frac{X_1 + \dots + X_n}{n}$ conduisent aux majorations :

Th or me 7.1.11 *Sous les conditions du Th or me 7.1.10, on a la majoration*

$$P\left(\frac{X_1 + \dots + X_n}{n} \geq a\right) \leq e^{-n h(a)}$$

et

$$P\left(\frac{X_1 + \dots + X_n}{n} \leq -a\right) \leq e^{-n h(-a)} ,$$

o  $a > 0$.

En particulier, pour tout $\varepsilon > 0$

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n}\right| \geq \varepsilon\right) \leq e^{-n h(\varepsilon)} + e^{-n h(-\varepsilon)}.$$

C'est cette majoration qu'on appelle le théorème des *grandes déviations à la loi des grands nombres*. Puisque $h(\varepsilon)$ et $h(-\varepsilon)$ sont strictement positifs

$$\sum_{n=1}^{\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n}\right| \geq \varepsilon\right) < \infty.$$

Le Théorème 7.1.6 nous permet alors de conclure que $\lim_{n \uparrow \infty} \frac{X_1 + \cdots + X_n}{n} = 0$, p.s., ce qui donne une autre preuve, sous la condition (7.3), de la loi forte des grands nombres.

Échange des opérations de limite et d'espérance

Nous allons énoncer à nouveau (pour le lecteur qui aurait sauté le chapitre 6) deux théorèmes extrêmement utiles donnant des conditions pour que

$$E\left[\lim_{n \uparrow \infty} X_n\right] = \lim_{n \uparrow \infty} E[X_n]. \quad (7.7)$$

Nous commencerons par le théorème de la *convergence monotone*, dit de *Beppo Levi*.

Théorème 7.1.12 Soit $(X_n, n \geq 1)$ une suite de variables aléatoires telles que pour tout ω hors d'un ensemble \mathcal{N} de probabilité nulle et pour tout $n \geq 1$,

$$0 \leq X_n(\omega) \leq X_{n+1}(\omega).$$

Alors (7.7) est vraie.

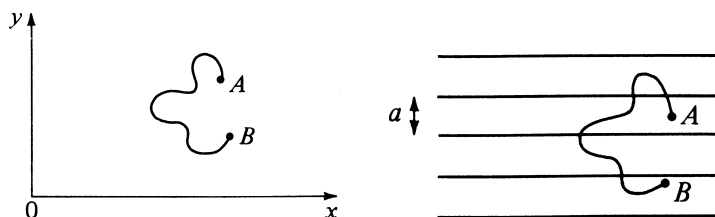
Voici maintenant le théorème de *convergence dominée*, dit de *Lebesgue*.

Théorème 7.1.13 Soit $(X_n, n \geq 1)$ une suite de variables aléatoires telles que pour tout ω hors d'un ensemble \mathcal{N} de probabilité nulle, $\lim_{n \uparrow \infty} X_n(\omega)$ existe, et pour tout $n \geq 1$,

$$|X_n(\omega)| \leq Y(\omega),$$

où Y est une variable aléatoire intégrable. Alors (7.7) est vraie.

EXEMPLE 7.1.3: RÉCRÉATION : L'AIGUILLE DE BUFFON. Une courbe \mathcal{C} rigide (d'extrémités A et B) est jetée au hasard sur un réseau de droites parallèles équidistantes de a . Quel est le nombre moyen $E[N(\mathcal{C})]$ de points d'intersection de la courbe et du réseau ?

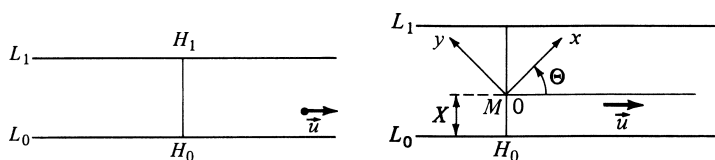


Buffon avait considéré un cas particulier de ce problème et calculé la probabilité pour qu'un segment (l'aiguille de Buffon) de longueur $\ell < a$ touche une droite du réseau. Il s'agit bien d'un cas particulier, car si on appelle N le nombre d'intersections de la courbe avec le réseau, cette variable aléatoire ne peut prendre, lorsque l'aiguille est de longueur ℓ strictement inférieure à la période du réseau, que deux valeurs, 0 ou 1, et donc

$$E[N] = 0 \times P(N = 0) + 1 \times P(N = 1) = P(N = 1) ,$$

la probabilité que l'aiguille touche le réseau.

Avant de résoudre le problème général, il faut préciser ce qu'on entend par "jeter une courbe au hasard". Voici le modèle proposé : La courbe \mathcal{C} étant solidaire d'un plan muni d'un repère cartésien xOy , le positionnement de ce plan sur le réseau est décrit de la façon suivante.



Un point H_0 est fixé sur la droite L_0 et on place au hasard l'origine 0 sur la perpendiculaire à L_0 élevée à partir de H_0 , entre les droites L_0 et L_1 de la manière suivante :

$$P(X \in [c, d]) = \frac{d - c}{a} ,$$

pour tout sous-intervalle $[c, d]$ de $[0, a]$, où X est la distance OH_0 . Puis on choisit l'orientation de Ox par rapport à une direction fixe (disons parallèle aux droites du réseau et dirigée de gauche à droite), selon

$$P(\Theta \in [\theta_1, \theta_2]) = \frac{\theta_2 - \theta_1}{2\pi}$$

pour tout intervalle $[\theta_1, \theta_2]$ de $[0, 2\pi]$, où Θ est l'angle de Ox avec la direction \vec{u} . Comme le réseau est périodique, le choix des 2 lignes L_0 et L_1 au lieu de 2 lignes successives quelconques L_n et L_{n-1} n'influe pas sur le nombre moyen des intersections. De même le déplacement de H_0 sur L_0 ne changera pas ce nombre moyen. On peut donc raisonnablement considérer que le positionnement de la courbe \mathcal{C} décrit ci-dessus est fait au hasard.

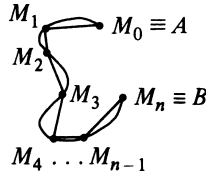
Soit \mathcal{C}_n la ligne brisée $M_0M_1 \dots M_n$. On a :

$$N(\mathcal{C}_n) = N(M_0M_1) + \dots + N(M_{n-1}M_n)$$

Nous prouverons très bientôt que le nombre moyen d'intersections d'un segment avec le réseau est proportionnel à la longueur du segment. On a donc :

$$E[N(\mathcal{C}_n)] = k\ell(\mathcal{C}_n)$$

où $\ell(\mathcal{C}_n)$ est la longueur de \mathcal{C}_n et k est un coefficient de proportionnalité à déterminer.



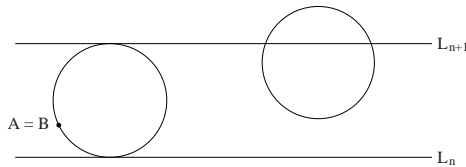
Si la courbe \mathcal{C} est rectifiable, on peut l'approximer par une suite de lignes brisées du type \mathcal{C}_n , où $M_0 = A$ et $M_n = B$, et telles que

$$\lim_{n \uparrow \infty} \ell(\mathcal{C}_n) = \ell(\mathcal{C}) .$$

On a donc en passant à la limite,

$$E[N(\mathcal{C})] = k \times \ell(\mathcal{C}) .$$

(Pour justifier ce passage à la limite, on peut invoquer le théorème de Beppo Levi. On observe en effet que on peut choisir les lignes brisées \mathcal{C}_n de façon que $N(\mathcal{C}_n)$ croisse avec n , et admette pour limite, dans le cas rectifiable, $N(\mathcal{C})$.) Pour calculer k , considérons la courbe formée d'un cercle de diamètre a . Dans ce cas particulier, le nombre d'intersection est *toujours* égal à 2, et donc $E[N(\mathcal{C})] = 2$.



On déduit donc que

$$k = \frac{E[N(\mathcal{C})]}{\ell(\mathcal{C})} = \frac{2}{\pi a} .$$

Finalement :

$$E[N(\mathcal{C})] = \frac{2}{\pi a} \ell(\mathcal{C}) , \tag{7.8}$$

et en particulier la probabilité pour qu'une aiguille de longueur $\ell < a$ touche le réseau est $\frac{2\ell}{\pi a}$.

Pour compléter l'argument, il reste à prouver directement que dans le cas d'une aiguille de longueur ℓ quelconque le nombre moyen d'intersections est proportionnel à ℓ . L'espérance du nombre d'intersections d'une telle aiguille avec le réseau est une fonction de ℓ , soit $f(\ell)$. Si on met deux aiguilles de longueur ℓ_1 et ℓ_2 bout à bout de façon à n'en former qu'une seule de longueur $\ell = \ell_1 + \ell_2$, il apparaît alors que

$$f(\ell_1 + \ell_2) = f(\ell_1) + f(\ell_2) .$$

La solution de cette équation fonctionnelle est $f(\ell) = k\ell$.

Théorème 7.1.14 *Soit X une variable aléatoire réelle de fonction caractéristique ψ , et telle que $E[|X|^n] < \infty$ pour un $n \geq 1$. Alors, pour tout entier $r \leq n$, la r -ième dérivée de ψ , notée $\psi^{(r)}$, existe et est donnée par la formule*

$$\psi^{(r)}(u) = i^r E[X^r e^{iuX}] . \quad (7.9)$$

En particulier, $i^r E[X^r] = \psi^{(r)}(0)$. De plus

$$\psi(u) = \sum_{r=0}^n \frac{(iu)^r}{r!} E[X^r] + \frac{(iu)^n}{n!} \varepsilon_n(u) , \quad (7.10)$$

où $\lim_{n \uparrow \infty} \varepsilon_n(u) = 0$ et $|\varepsilon_n(u)| \leq 3E[|X|^n]$.

Démonstration. Pour tout réel $a \geq 0$ et tout entier $r \leq n$, $a^r \leq 1 + a^n$ (en effet, si $a \leq 1$, alors $a^r \leq 1$, et si $a \geq 1$, alors $a^r \leq a^n$). En particulier,

$$\begin{aligned} E[|X|^r] &\leq E[1 + |X|^n] \\ &= 1 + E[|X|^n] < \infty . \end{aligned}$$

Supposons que pour $r < n$,

$$\psi^{(r)}(u) = i^r E[X^r e^{iuX}] .$$

Alors,

$$\begin{aligned} \frac{\psi^{(r)}(u+h) - \psi^{(r)}(u)}{h} &= i^r E \left[X^r \frac{e^{i(u+h)X} - e^{iuX}}{h} \right] \\ &= i^r E \left[X^r e^{iuX} \frac{e^{ihX} - 1}{h} \right] . \end{aligned}$$

La quantité sous l'espérance tend vers $iX^{r+1}e^{iuX}$ lorsque $h \rightarrow 0$, et de plus, elle est bornée en valeur absolue par une fonction intégrable, puisque

$$\left| X^r e^{iuX} \frac{e^{ihX} - 1}{h} \right| \leq \left| X^r \frac{e^{ihX} - 1}{h} \right| \leq |X|^{r+1} .$$

(Pour la dernière inégalité, on utilise le fait que $|e^{ia} - 1| = |2 \sin \frac{a}{2}| \leq |a|$.) Donc, par convergence dominée,

$$\begin{aligned}\psi^{(r+1)}(u) &= \lim_{h \rightarrow 0} \frac{\psi(u+h) - \psi(u)}{h} \\ &= i^r E \left[\lim_{h \rightarrow 0} X^r e^{iuX} \frac{e^{ihX} - 1}{h} \right] \\ &= i^{r+1} E [X^{r+1} e^{iuX}].\end{aligned}$$

L'égalité (7.9) suit, car l'hypothèse d'induction est triviale pour $r = 0$.

Preuve de (7.10). La formule de Taylor donne, pour $y \in \mathbb{R}$,

$$e^{iy} = \cos y + i \sin y = \sum_{k=0}^{n-1} \frac{(iy)^k}{k!} + \frac{(iy)^n}{n!} (\cos(\alpha_1 y) + i \sin(\alpha_2 y))$$

où $\alpha_1 = \alpha_1(y)$ et $\alpha_2 = \alpha_2(y) \in [-1, +1]$. Donc

$$e^{iuX} = \sum_{k=0}^{n-1} \frac{(iuX)^k}{k!} + \frac{(iuX)^n}{n!} (\cos(\theta_1 uX) + i \sin(\theta_2 uX))$$

où $\theta_1 = \theta_1(\omega), \theta_2 = \theta_2(\omega) \in [-1, +1]$, et

$$E[e^{iuX}] = \sum_{k=0}^{n-1} \frac{(iu)^k}{k!} E[X^k] + \frac{(iu)^n}{n!} (E[X^n] + \varepsilon_n(u)),$$

avec

$$\varepsilon_n(u) = E[X^n (\cos \theta_1 uX + i \sin \theta_2 uX - 1)].$$

La variable $X^n (\cos \theta_1 uX + i \sin \theta_2 uX - 1)$ est bornée en valeur absolue par la variable aléatoire intégrable $3|X|^n$ et tend vers 0 lorsque $u \rightarrow 0$. On a donc, par convergence dominée, $\lim_{u \rightarrow 0} \varepsilon_n(u) = 0$. \square

7.2 Convergence en distribution

Le théorème de la loi gaussienne limite

Supposons qu'on ait à mesurer expérimentalement une grandeur dont la valeur (inconnue) est m . S'il n'y avait pas d'erreurs expérimentales, il suffirait d'une seule mesure. Mais les mesures sont toujours entachées d'erreurs et une expérience se traduit par un résultat aléatoire X . On suppose que la moyenne m est la valeur cherchée. Pour approcher la vraie valeur m on fait n expériences indépendantes et identiques ce qui va donner lieu à n variables aléatoires X_1, \dots, X_n indépendantes et identiquement distribuées, de moyenne m et de variance σ^2 supposée finie, et on va déclarer que la valeur cherchée est la *moyenne empirique* pour un échantillon de taille n , soit $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$.

La loi forte des grands nombres nous rassure sur une telle évaluation de m si n est assez grand. Cependant, on sait que la moyenne empirique n'est pas égale à la moyenne et on aimerait avoir une idée de combien elle s'en éloigne. Des résultats tels que l'inégalité de Tchebychev, ou une inégalité plus sophistiquée telle que l'inégalité de Chernoff (voir plus haut), répondent à ces questions en donnant des bornes en probabilité de l'écart entre la moyenne empirique et l'espérance mathématique.

Le *théorème de la loi gaussienne limite* quant à lui permet une évaluation asymptotique de la *loi* de l'écart standardisé $(\bar{X}_n - m)/\sigma\sqrt{n}$ (dans ce contexte, "standardisé" veut dire qu'on a ramené la moyenne à 0 et la variance à 1). Plus précisément

$$\lim_{n \uparrow \infty} P \left(\frac{1}{\sigma\sqrt{n}} \sum_{j=1}^n (X_j - m) \leq x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy . \quad (7.11)$$

Il existe un cas où la démonstration de (7.11) est triviale : c'est quand chacune des variables aléatoires X_n est une gaussienne. Comme les X_n sont indépendantes, la somme $X_1 + \dots + X_n$ est aussi une gaussienne dont la moyenne est la somme des moyennes, nm , et la variance est la somme des variances, $n\sigma^2$. La moyenne empirique standardisée est donc une gaussienne standard. Dans ce cas, *pour tout* n ,

$$P \left(\frac{\sum_{j=1}^n (X_j - m)}{\sigma\sqrt{n}} \leq x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}y^2} dy .$$

La notion de base est ici celle de *convergence en distribution*.

Définition 7.2.1 Soit $(X_n, n \geq 1)$ une suite de variables aléatoires réelles, et X une variable aléatoire réelle. On dit que la suite $(X_n, n \geq 1)$ converge en distribution, ou en loi, vers X si

$$\lim_{n \uparrow \infty} P(X_n \leq x) = P(X \leq x) \text{ pour tout } x \in \mathbb{R} \text{ tel que } P(X = x) = 0 . \quad (7.12)$$

On note ceci

$$X_n \xrightarrow{\mathcal{D}} X .$$

EXEMPLE 7.2.1: Si on avait imposé la convergence pour tout $x \in \mathbb{R}$, on n'aurait pas pu dire par exemple que $X_n \xrightarrow{\mathcal{D}} X$ lorsque les X sont des variables aléatoires réelles telles que

$$P \left(X_n = a + \frac{1}{n} \right) = 1, \quad P(X = a) = 1 .$$

En effet, au point de discontinuité a , on a $P(X_n \leq a) = 0$, et cette quantité ne tend donc pas vers $P(X \leq a) = 1$.

Notons que la définition de la convergence en distribution ne fait intervenir que les fonctions de répartition des variables aléatoires en cause. La notion de base est celle de *convergence faible des fonctions de répartition* :

Définition 7.2.2 *On dit que la suite $(F_n, n \geq 1)$ de fonctions de répartition converge faiblement vers la fonction de répartition F si et seulement si*

$$\lim_{n \uparrow \infty} F_n(x) = F(x) \text{ pour tout } x \in \mathbb{R} \text{ tel que } F(x) = F(x-) . \quad (7.13)$$

Il se peut très bien que, dans la définition de la convergence en distribution, les $(X_n, n \geq 1)$ et X ne soient pas définies sur le même espace. On peut même se passer de définir X , pourvu qu'on ait une fonction de répartition. On dira par exemple que la suite $(X_n, n \geq 1)$ de variables aléatoires converge en distribution vers F , où F est une fonction de répartition sur \mathbb{R} , si

$$\lim_{n \uparrow \infty} P(X_n \leq x) = F(x) \text{ pour tout } x \in \mathbb{R} \text{ tel que } F(x) - F(x-) = 0 . \quad (7.14)$$

On note ceci

$$X_n \xrightarrow{\mathcal{D}} F .$$

Cette notation n'est pas très orthodoxe, mais nous l'utiliserons souvent. Par exemple (7.11) se lit, avec un abus de notation supplémentaire :

$$\frac{1}{\sigma\sqrt{n}} \sum_{j=1}^n (X_j - m) \rightarrow \mathcal{N}(0, 1) .$$

EXEMPLE 7.2.2: Cet exemple très simple est destiné à faire prendre conscience au lecteur, si nécessaire, de la différence entre la convergence presque sûre et la convergence en distribution. En effet, si on considère une suite IID, disons de variables à valeurs dans $\{0, 1\}$, de distribution commune $P(X_n = 0) = P(X_n = 1) = \frac{1}{2}$, il est clair que cette suite converge en distribution, et qu'elle ne converge pas presque sûrement !

Le critère des fonctions caractéristiques

Nous admettrons sans démonstration le résultat fondamental suivant.

Théorème 7.2.1 *Soit $(X_n, n \geq 1)$ une suite de variables aléatoires réelles et soit X une variable aléatoire réelle. Les trois propriétés suivantes sont équivalentes :*

(i) $X_n \xrightarrow{\mathcal{D}} X$

(ii) Pour toute fonction $f : \mathbb{R} \rightarrow \mathbb{R}$ continue et bornée

$$\lim_{n \uparrow \infty} E[f(X_n)] = E[f(X)] .$$

(iii) Pour tout $u \in \mathbb{R}$,

$$\lim_{n \uparrow \infty} E[e^{iuX_n}] = E[e^{iuX}] .$$

L'équivalence de (iii) et de (i) s'appelle le critère de la fonction caractéristique. Il est dû à Paul Lévy, qui a donné une forme plus fine de (iii) :

Théorème 7.2.2 Soit $(X_n, n \geq 1)$ une suite de variables aléatoires réelles, telle que pour tout $u \in \mathbb{R}$,

$$\lim_{n \uparrow \infty} E[e^{iuX_n}] = \phi(u) ,$$

où ϕ est continue en 0. Alors, ϕ est la fonction caractéristique d'une variable aléatoire réelle X , et $X_n \xrightarrow{\mathcal{D}} X$.

Ce résultat permet d'identifier la limite de fonctions caractéristiques, lorsqu'elle est continue à l'origine (une condition nécessaire pour une fonction caractéristique), comme une fonction caractéristique.

Sur ces bases, nous sommes maintenant en mesure de démontrer le théorème de la limite centrale.

Théorème 7.2.3 Soit $(X_n, n \geq 1)$ une suite IID de variables aléatoires de carré intégrable, de moyenne m et de variance σ^2 . Pour tout $x \in \mathbb{R}$ on a (7.11).

Démonstration. On supposera, sans perte de généralité que $m = 0$. Au vu du Théorème 7.2.1, il suffit de démontrer que

$$\lim_{n \uparrow \infty} \phi_n(u) = e^{-\sigma^2 u^2 / 2},$$

où, en utilisant l'indépendance des X_n ,

$$\begin{aligned} \phi_n(u) &= E \left[\exp \left\{ iu \frac{\sum_{j=1}^n X_j}{\sqrt{n}} \right\} \right] \\ &= \prod_{j=1}^n E \left[\exp \left\{ i \frac{u}{\sqrt{n}} X_j \right\} \right] \\ &= \psi \left(\frac{u}{\sqrt{n}} \right)^n, \end{aligned}$$

où ψ est la fonction caractéristique de X_1 . Du développement de Taylor de ψ en 0 (Théorème 7.1.14)

$$\psi(u) = 1 + \frac{\psi''(0)}{2!} u^2 + o(u^2),$$

on déduit que pour $u \in \mathbb{R}$ fixé,

$$\psi \left(\frac{u}{\sqrt{n}} \right) = 1 - \frac{1}{n} \frac{\sigma^2 u^2}{2} + o \left(\frac{1}{n} \right),$$

et donc

$$\lim_{n \uparrow \infty} \log \{ \phi_n(u) \} = \lim_{n \uparrow \infty} n \left(\log \left\{ 1 - \frac{\sigma^2 u^2}{2n} + o \left(\frac{1}{n} \right) \right\} \right) = -\frac{1}{2} \sigma^2 u^2.$$

Le Théorème 7.2.1 permet de conclure. □

Un exemple simple de test statistique

Voici une des utilisations du théorème de la limite gaussienne. On a fait n expériences indépendantes et identiques qui donnent n copies indépendantes d'une variable X de variance σ^2 finie et de moyenne m . On veut se faire une idée de la répartition de la moyenne empirique $\frac{X_1 + \dots + X_n}{n}$ autour de la valeur moyenne m . On fait l'approximation, justifiée par le théorème de la limite gaussienne que $(\frac{X_1 + \dots + X_n}{n} - m) \times \frac{\sqrt{n}}{\sigma}$ est une gaussienne standard. On a donc

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - m\right| \geq a\right) \sim 1 - \int_{\frac{a}{\sigma}\sqrt{n}}^{\frac{a}{\sigma}\sqrt{n}} e^{-\frac{y^2}{2}} dy.$$

Supposons que la moyenne m soit inconnue. On l'approximera naturellement par la moyenne empirique $\frac{X_1 + \dots + X_n}{n}$.

On se fixe un *intervalle de confiance* $[-a, +a]$ et une "tolérance" $\varepsilon \in (0, 1)$. On veut que la probabilité que la valeur absolue de l'erreur commise soit supérieure à a soit inférieure à ε :

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - m\right| \geq a\right) \leq \varepsilon.$$

D'après le théorème de la limite gaussienne, le premier membre de cette inégalité peut être évalué, si n est grand, par $1 - \int_{\frac{a}{\sigma}\sqrt{n}}^{\frac{a}{\sigma}\sqrt{n}} e^{-\frac{y^2}{2}} dy$. On devra donc prendre un nombre d'expériences n tel que

$$\varepsilon \geq 1 - \int_{\frac{a}{\sigma}\sqrt{n}}^{\frac{a}{\sigma}\sqrt{n}} e^{-\frac{y^2}{2}} dy. \quad (7.15)$$

Pour appliquer cette procédure, il faut avoir connaissance de la variance σ^2 . A défaut, on prendra une borne supérieure σ_0 de σ , et on déterminera n par

$$\varepsilon \geq 1 - \int_{\frac{a}{\sigma_0}\sqrt{n}}^{\frac{a}{\sigma_0}\sqrt{n}} e^{-\frac{y^2}{2}} dy,$$

ce qui entraîne a fortiori l'inégalité (7.15).

Cas des vecteurs aléatoires

Pour les vecteurs aléatoires, la définition de la convergence en distribution à l'aide des fonctions de répartition n'est pas pratique. On a cependant un résultat analogue au Théorème 7.2.1.

Théorème 7.2.4 Soit $(X_n, n \geq 1)$ une suite de vecteurs aléatoires de \mathbb{R}^d et soit X un vecteur aléatoire de \mathbb{R}^d . Les deux propriétés suivantes sont équivalentes

(i) Pour toute fonction $f : \mathbb{R}^d \rightarrow \mathbb{R}$ continue et bornée

$$\lim_{n \uparrow \infty} E[f(X_n)] = E[f(X)] .$$

(ii) Pour tout $u \in \mathbb{R}^d$,

$$\lim_{n \uparrow \infty} E[e^{iu^T X_n}] = E[e^{iu^T X}] .$$

On dit alors que $(X_n, n \geq 1)$ converge en distribution vers X .

Le lecteur imaginera facilement la version du Théorème 7.2.2 applicable au cas des vecteurs aléatoires.

7.3 Autres types de convergence

Convergence en probabilité

Définition 7.3.1 Soit $(X_n, n \geq 1)$ une suite de variables aléatoires réelles définies sur le même espace de probabilité (Ω, \mathcal{F}, P) et soit X une autre variable aléatoire sur ce même espace. On dit que la suite $(X_n, n \geq 1)$ converge en probabilité vers X si, pour tout $\varepsilon > 0$,

$$\lim_{n \uparrow \infty} P(|X_n - X| \geq \varepsilon) = 0. \quad (7.16)$$

La convergence presque-sûre est plus forte que la convergence en probabilité, “mais pas beaucoup plus”. Les énoncés précis sont donnés dans les deux théorèmes qui suivent.

Théorème 7.3.1 La convergence presque-sûre entraîne la convergence en probabilité.

Démonstration. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires qui converge presque sûrement vers la variable aléatoire X . Posons, pour $\varepsilon > 0$ fixé,

$$A_n = \{|X_n - X| \geq \varepsilon\} .$$

On a

$$\lim_{n \uparrow \infty} P(A_n) \leq \lim_{n \uparrow \infty} P\left(\bigcup_{m=n}^{\infty} A_m\right) = P(\limsup_{n \uparrow \infty} A_n).$$

Cette dernière quantité est nulle (Théorème 7.1.6). □

Théorème 7.3.2 Soit $(X_n, n \geq 1)$ une suite de variables aléatoires qui converge en probabilité vers X . On peut en extraire une sous-suite $(X_{n_k}, k \geq 1)$ qui converge presque-sûrement vers X .

Démonstration. On a pour tout $\varepsilon > 0$

$$\lim_{n \uparrow \infty} P(|X_n - X| \geq \varepsilon) = 0 .$$

En particulier, il existe un entier n_1 tel que

$$P(|X_{n_1} - X| \geq 1) \leq \frac{1}{2} ,$$

et un entier $n_2 > n_1$ tel que

$$P\left(|X_{n_2} - X| \geq \frac{1}{2}\right) \leq \frac{1}{2^2} ,$$

ainsi de suite, jusqu'à obtenir une suite strictement croissante $(n_k, k \geq 1)$ telle que pour tout $k \geq 1$

$$P\left(|X_{n_k} - X| \geq \frac{1}{k}\right) < \frac{1}{2^k} .$$

En particulier,

$$\sum_{k=1}^{\infty} P\left(|X_{n_k} - X| \geq \frac{1}{k}\right) < \infty ,$$

d'où le résultat d'après le Théorème 7.1.7. □

L'Exercice 7.4.3 est recommandé, car il donne un exemple dans lequel on a la convergence en probabilité mais pas la convergence presque-sûre.

Pour compléter le tableau hiérarchique des convergences, mentionnons la *convergence en moyenne quadratique* :

Définition 7.3.2 Soit $(X_n, n \geq 1)$ une suite de variables aléatoires de carré intégrable définies sur le même espace de probabilité (Ω, \mathcal{F}, P) et soit X une autre variable aléatoire de carré intégrable sur ce même espace. On dit que X_n converge en moyenne quadratique vers X si

$$\lim_{n \uparrow \infty} E[|X_n - X|^2] = 0 .$$

On note ceci $X_n \xrightarrow{m.q.} X$.

Définition 7.3.3 Soit $(X_n, n \geq 1)$ une suite de vecteurs aléatoires de \mathbb{R}^m de carré intégrable (chaque coordonnée est de carré intégrable) définis sur le même espace de probabilité (Ω, \mathcal{F}, P) et soit X un autre vecteur aléatoire de \mathbb{R}^m de carré intégrable sur ce même espace. On dit que la suite de vecteurs aléatoires $(X_n, n \geq 1)$ converge en moyenne quadratique vers le vecteur aléatoire X si pour tout $j, 1 \leq j \leq m$,

$$X_{n,j} \xrightarrow{m.q.} X_j ,$$

où $X_n = (X_{n,1}, \dots, X_{n,m})$ et $X = (X_1, \dots, X_m)$.

Théorème 7.3.3 Soit $(X_n, n \geq 1)$ et $(Y_n, n \geq 1)$ deux suites de variables aléatoires réelles définies sur le même espace de probabilité qui convergent en moyenne quadratique vers X et Y respectivement. On a

$$\lim_{n,m \uparrow \infty} E[X_n Y_m] = E[XY] .$$

En particulier, $\lim_{n \uparrow \infty} E[X_n] = E[X]$ (faire $Y_n = 1$), et $\lim_{n,m \uparrow \infty} \text{cov}(X_n, Y_m) = \text{cov}(X, Y)$.

Démonstration. Exercice 7.4.10. □

Théorème 7.3.4 La convergence en moyenne quadratique d'une suite de variables aléatoires entraîne la convergence en probabilité vers la même limite.

Démonstration. Cette implication découle immédiatement de l'inégalité de Markov

$$P(|X_n - X| \geq \varepsilon) \leq \frac{E[|X_n - X|^2]}{\varepsilon^2} .$$

□

Théorème 7.3.5 La convergence en probabilité d'une suite de variables aléatoires entraîne la convergence en distribution vers la même limite.

Démonstration. Étant donné $\varepsilon > 0$ et $\delta > 0$, pour n suffisamment grand,

$$P(|Z_n - Z| \geq \varepsilon) \leq \delta .$$

Supposons qu'on puisse montrer que

$$P(Z \leq x - \varepsilon) - \delta \leq P(Z_n \leq x) \leq P(Z \leq x + \varepsilon) + \delta . \quad (7.17)$$

Si x est un point de continuité de $x \rightarrow P(Z \leq x)$, ε peut être choisi de telle sorte que

$$\begin{aligned} P(Z \leq x - \varepsilon) &\geq P(Z \leq x) - \delta , \\ P(Z \leq x + \varepsilon) &\leq P(Z \leq x) + \delta , \end{aligned}$$

et donc, pour un δ arbitraire, et pour tout n suffisamment grand,

$$P(Z \leq x) - 2\delta \leq P(Z_n \leq x) \leq P(Z \leq x) + 2\delta .$$

Reaste à prouver (7.17). On commence par l'inégalité de droite. (Celle de gauche utilise les mêmes arguments.) Il suffit d'écrire

$$\begin{aligned} P(Z_n \leq x) &= P(Z_n \leq x, |Z_n - Z| < \varepsilon) + P(Z_n \leq x, |Z_n - Z| \geq \varepsilon) \\ &\leq P(Z \leq x + \varepsilon) + P(|Z_n - Z| \geq \varepsilon) \\ &\leq P(Z \leq x + \varepsilon) + \delta . \end{aligned}$$

□

Théorème 7.3.6 Soit $(X_n, n \geq 1)$ une suite de vecteurs aléatoires gaussiens de \mathbb{R}^m qui convergent en moyenne quadratique vers le vecteur X . Alors X est nécessairement un vecteur gaussien.

Démonstration. Pour tout $a \in \mathbb{R}^m$, $a^T X_n \xrightarrow{m.q.} a^T X$ et donc (Théorème 7.3.5) $a^T X_n \xrightarrow{\mathcal{D}} a^T X$. En particulier, d'après le critère des fonctions caractéristiques, pour tout $u \in \mathbb{R}^m$

$$\lim_{n \uparrow \infty} \phi_{X_n}(u) = \phi_X(u) .$$

Or

$$\phi_{X_n}(u) = e^{iu^T E[X_n] - \frac{1}{2} u^T \text{cov}(X_n, X_n) u} .$$

Mais (Théorème 7.3.3)

$$E[X] = \lim_{n \uparrow \infty} E[X_n] \quad , \quad \text{cov}(X, X) = \lim_{n \uparrow \infty} \text{cov}(X_n, X_n) .$$

On a donc

$$\phi_X(u) = e^{iu^T E[X] - \frac{1}{2} u^T \text{cov}(X, X) u} ,$$

ce qui est la fonction caractéristique d'un vecteur gaussien. □

7.4 Exercices

Exercice 7.4.1.

Démontrez le Théorème 7.1.2.

Exercice 7.4.2.

Soit une suite quelconque $(A_n, n \geq 1)$ de sous-ensembles de Ω . Formons pour chaque $n \geq 1$, le sous-ensemble

$$C_n = \bigcap_{p=n}^{\infty} A_p$$

(aussi noté $\inf_{n \geq p} A_p$). La suite $(C_n, n \geq 1)$ est croissante, et on peut donc définir $\lim \uparrow C_n = \cup_{n=1}^{\infty} C_n$. C'est cette limite qu'on appelle la limite inférieure de la suite $(A_n, n \geq 1)$ et que l'on dénote $\liminf_n A_n$. Donc, *par définition*,

$$\liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{p=n}^{\infty} A_p .$$

Démontrez que pour que $\omega \in \liminf_n A_n$ il faut et il suffit qu'il existe un indice $N = N(\omega)$ tel que pour tous les indices p supérieurs à N , $\omega \in A_p$.

Exercice 7.4.3.

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes à valeurs dans $\{0, 1\}$. On pose

$$p_n = P(X_n = 1) .$$

Montrez qu'une condition nécessaire et suffisante pour que $X_n \xrightarrow{Pr} 0$ est

$$\lim_{n \uparrow \infty} p_n = 0 .$$

Montrez qu'une condition nécessaire et suffisante pour que $X_n \xrightarrow{p.s.} 0$ est

$$\sum_{n=1}^{\infty} p_n < \infty .$$

Exercice 7.4.4.

Démontrez le lemme de Borel–Cantelli de manière différente de celle du cours, en considérant la somme $\sum_{n=1}^{\infty} 1_{A_n}$.

Exercice 7.4.5.

Prouvez la majoration

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{1}{4\varepsilon^4} \frac{1}{n^2} \quad (\varepsilon > 0, n \geq 1),$$

où $\frac{S_n}{n}$ est la moyenne empirique $\frac{X_1 + \dots + X_n}{n}$ relative à une suite IID $(X_n, n \geq 1)$ de variables aléatoires de Bernoulli de moyenne p .

Exercice 7.4.6.

On considère une suite $(X_n, n \geq 1)$ de variables aléatoires réelles *identiquement distribuées et indépendantes*, avec

$$P(X_n = 0) = P(X_n = 1) = \frac{1}{2} .$$

On se donne une suite $(a_n, n \geq 1)$ d'entiers satisfaisant à :

$$\log_2 n \leq a_n < (\log_2 n) + 1 .$$

On regarde $(X_n, n \geq 1)$ par blocs successifs, le n -ème bloc, de longueur a_{n+1} , étant

$$X_{a_1 + \dots + a_n + 1}, X_{a_1 + \dots + a_n + 2}, \dots, X_{a_1 + \dots + a_n + a_{n+1}} .$$

Quelle est la probabilité pour qu'on rencontre une infinité de blocs qui ne contiennent que des 1 ?

Exercice 7.4.7. CARACTÉRISATION DE LA LOI GAUSSIENNE.

Soit F une fonction de répartition sur \mathbb{R} , de moyenne nulle et de variance $\sigma^2 < \infty$. On suppose que si X_1 et X_2 sont deux variables aléatoires indépendantes de répartition commune F , alors $\frac{X_1 + X_2}{\sqrt{2}}$ admet aussi la répartition F .

Montrez que F est la fonction de répartition d'une gaussienne centrée de variance σ^2 .

Exercice 7.4.8.

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires poissonniennes qui converge en loi vers une variable aléatoire X réelle. Montrez que X est nécessairement poissonnienne.

Exercice 7.4.9.

Soit $F(x)$ une fonction de répartition de probabilité sur \mathbb{R} telle que

$$\begin{aligned} F(x) &= 0 & \text{si} & & x < a \\ 0 < F(x) < 1 & \text{si} & & a < x < b \\ F(x) &= 1 & \text{si} & & x \geq b, \end{aligned}$$

où $a, b \in \mathbb{R}$. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires IID, de fonction de répartition commune $F(x)$. On note

$$\begin{aligned} Y_n &= \inf(X_1, \dots, X_n) \\ Z_n &= \sup(X_1, \dots, X_n). \end{aligned}$$

A. Montrez que $Y_n \xrightarrow{\mathcal{D}} a$ et que $Z_n \xrightarrow{\mathcal{D}} b$.

B. Montrez que si les X_n sont uniformément distribuées sur $[0, 1]$, $nY_n \xrightarrow{\mathcal{D}} \mathcal{E}(1)$, la distribution exponentielle de paramètre 1.

Exercice 7.4.10. CONTINUITÉ DU PRODUIT SCALAIRE.

Soit $(X_n, n \geq 1)$ et $(Y_n, n \geq 1)$ deux suites de variables aléatoires réelles qui convergent en moyenne quadratique vers X et Y respectivement. Montrez que

$$\lim_{n \uparrow \infty} E[X_n Y_n] = E[XY].$$

Exercice 7.4.11. CAUCHY SOUS TOUS LES ANGLES, 1.

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires IID de Cauchy. On pose $S_n = X_1 + \dots + X_n$. Est-ce que $(S_n/\sqrt{n}, n \geq 1)$ converge en probabilité ? en moyenne quadratique ? presque sûrement ? en distribution ?

Exercice 7.4.12. * CAUCHY SOUS TOUS LES ANGLES, 2.

Est-ce que $(S_n/n^2, n \geq 1)$ converge en probabilité ? en moyenne quadratique ? en distribution ?

Exercice 7.4.13. CAUCHY SOUS TOUS LES ANGLES, 3.

Est-ce que $(S_n/n, n \geq 1)$ converge en distribution ? presque sûrement ?

Exercice 7.4.14. * SOMME DE SOMMES DE GAUSSIENNES.

Soit $\{\varepsilon_n\}_{n \geq 1}$ une suite IID de variables gaussiennes standard. On pose $X_n = \varepsilon_1 + \dots + \varepsilon_n$. Montrez que

$$\frac{X_1 + \dots + X_n}{n\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

pour un certain σ à calculer.

Exercice 7.4.15. * CONVERGENCE EN LOI DE VARIABLES GAUSSIENNES.

Soit $\{U_n\}_{n \geq 0}$ une suite IID de variables gaussiennes standard. On définit la suite $\{X_n\}_{n \geq 0}$ par

$$X_0 = U_0, \quad X_{n+1} = aX_n + U_{n+1} \quad (n \geq 0).$$

(i) Montrez que la famille $\{X_n\}_{n \geq 0}$ est gaussienne, et que si $a < 1$, $X_n \sim \mathcal{N}(0, \sigma^2)$ pour un certain σ à calculer.

(ii) Montrez que si $a > 1$, $(X_n/a^n, n \geq 1)$ converge en moyenne quadratique. Cette dernière suite converge-t-elle en loi ?

Chapitre 8

Chaînes de Markov

8.1 La matrice de transition

Une suite de variables aléatoires $\{X_n\}_{n \geq 0}$ à valeurs dans l'espace dénombrable E est appelé *processus stochastique* (à temps discret) (à valeurs dans E). L'ensemble E est l'*espace d'état*, dont les éléments seront notés i, j, k, \dots . Lorsque $X_n = i$, le processus est dit *être dans*, ou *visiter*, l'état i au temps n .

Les suites de variables aléatoires IID sont bien des processus stochastiques, mais du point de vue de la modélisation, elles ne sont pas satisfaisantes, ne prenant pas en compte la dynamique des systèmes en évolution, du fait de l'indépendance. Pour introduire cette dynamique, il faut tenir compte de l'influence du passé, ce que font les chaînes de Markov, à la façon des équations de récurrence dans les systèmes déterministes. En fait, les chaînes de Markov sont des processus stochastiques dont l'évolution est régie par une équation de récurrence du type $X_{n+1} = f(X_n, Z_{n+1})$, où $\{Z_n\}_{n \geq 1}$ est une suite IID indépendante de la valeur initiale X_0 (voir plus bas). Cette structure extrêmement simple suffit à générer une grande variété de comportements. C'est pour cela que les chaînes de Markov trouvent des applications dans beaucoup de domaines comme, par exemple, la biologie, la physique, la sociologie, la recherche opérationnelle et les sciences de l'ingénieur, où elles donnent des réponses qualitatives aussi bien que quantitatives aux problèmes posés.

Voici la définition :

Définition 8.1.1 Si pour tout entier $n \geq 0$ et tous états $i_0, i_1, \dots, i_{n-1}, i, j \in E$,

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i), \quad (8.1)$$

le processus $\{X_n\}_{n \geq 0}$ est appelé chaîne de Markov. Celle-ci est dite homogène si le second membre de (8.1) ne dépend pas de n .

On abrègera “chaîne de Markov homogène” par “CMH”.

La propriété (8.1) est la *propriété de Markov*.

La matrice $\mathbf{P} = \{p_{ij}\}_{i,j \in E}$, où

$$p_{ij} = P(X_{n+1} = j \mid X_n = i)$$

est la *probabilité de transition* de i vers j , est appelée *matrice de transition* de la chaîne. Comme ses éléments sont des probabilités et puisqu’une transition a nécessairement lieu d’un état vers un autre état, on a

$$p_{ij} \geq 0, \text{ et } \sum_{k \in E} p_{ik} = 1$$

pour tous états i, j . Une matrice \mathbf{P} indexée par E et satisfaisant les propriétés ci-dessus est appelée *matrice stochastique*.

L’espace d’état peut être infini et par conséquent une telle matrice n’est pas du type étudié en algèbre linéaire. Toutefois, l’addition et la multiplication sont définies par les mêmes règles formelles. Par exemple, avec $A = \{a_{ij}\}_{i,j \in E}$ et $B = \{b_{ij}\}_{i,j \in E}$, le produit $C = AB$ est la matrice $\{c_{ij}\}_{i,j \in E}$, où $c_{ij} = \sum_{k \in E} a_{ik} b_{kj}$. La notation $x = \{x_i\}_{i \in E}$ représente formellement un vecteur *colonne* et x^T est donc un vecteur *ligne*, le transposé de x . Par exemple, $y = \{y_i\}_{i \in E}$ donné par $y^T = x^T A$ est défini par $y_i = \sum_{k \in E} x_k a_{ki}$. Semblablement, $z = \{z_i\}_{i \in E}$ donné par $z = Ax$ est défini par $z_i = \sum_{k \in E} a_{ik} z_k$.

Une matrice de transition \mathbf{P} est parfois représentée par son *graphe de transition* G , un graphe dont les nœuds sont les états de E et qui a une arête orientée de i vers j si et seulement si $p_{ij} > 0$, auquel cas cette arête est ornée de l’étiquette p_{ij} .

La propriété de Markov (8.1) s’étend facilement (voir l’Exercice 8.5.2) comme suit

$$\begin{aligned} P(X_{n+1} = j_1, \dots, X_{n+k} = j_k \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\ = P(X_{n+1} = j_1, \dots, X_{n+k} = j_k \mid X_n = i) \end{aligned} \quad (8.2)$$

pour tous $i_0, \dots, i_{n-1}, i, j_1, \dots, j_k$. En notant

$$A = \{X_{n+1} = j_1, \dots, X_{n+k} = j_k\}, \quad B = \{X_0 = i_0, \dots, X_{n-1} = i_{n-1}\},$$

la dernière égalité se lit $P(A \mid X_n = i, B) = P(A \mid X_n = i)$, et celle-ci se lit à son tour

$$P(A \cap B \mid X_n = i) = P(A \mid X_n = i)P(B \mid X_n = i).$$

Donc, A et B sont conditionnellement indépendants sachant $X_n = i$. Tel est le contenu du :

Théorème 8.1.1 *Pour tout $i \in E$, $n \geq 1$, le futur au temps n et le passé au temps n sont conditionnellement indépendants étant donné l’état présent $X_n = i$.*

Voir cependant l’Exercice 8.5.1.

Le Théorème 8.1.1 montre en particulier que la propriété de Markov est indépendante de la direction du temps.

La distribution d'une CMH

La distribution au temps n de la chaîne est le vecteur $\nu_n = \{\nu_n(i)\}_{i \in E}$, où

$$\nu_n(i) = P(X_n = i).$$

La règle des causes totales donne $\nu_{n+1}(j) = \sum_{i \in E} \nu_n(i) p_{ij}$, c'est-à-dire, sous forme matricielle, $\nu_{n+1}^T = \nu_n^T \mathbf{P}$. En itérant cette égalité, on obtient pour tout $n \geq 1$:

$$\nu_n^T = \nu_0^T \mathbf{P}^n. \quad (8.3)$$

La matrice \mathbf{P}^n est appelée *matrice de transition en n étapes* car son terme général n'est autre que

$$p_{ij}(n) = P(X_{n+m} = j \mid X_m = i).$$

En effet, d'après la règle de Bayes séquentielle et en tenant compte de la propriété de Markov, on trouve pour le second membre de la dernière égalité

$$\sum_{i_1, \dots, i_{n-1} \in E} p_{ii_1} p_{i_1 i_2} \cdots p_{i_{n-1} j},$$

qui est bien le terme général de la n -ème puissance de la matrice \mathbf{P} .

Par convention, \mathbf{P}^0 est la matrice identité ($p_{i,j}(0) = 1_{\{i=j\}}$).

La variable X_0 est l'*état initial*, et sa distribution de probabilité ν_0 est la *distribution initiale*. La règle de Bayes séquentielle donne :

$$\begin{aligned} P(X_0 = i_0, X_1 = i_1, \dots, X_k = i_k) \\ = P(X_0 = i_0) P(X_1 = i_1 \mid X_0 = i_0) \cdots P(X_k = i_k \mid X_{k-1} = i_{k-1}, \dots, X_0 = i_0), \end{aligned}$$

et donc, dans le cas d'une CMH,

$$P(X_0 = i_0, X_1 = i_1, \dots, X_k = i_k) = \nu_0(i_0) p_{i_0 i_1} \cdots p_{i_{k-1} i_k}. \quad (8.4)$$

La donnée de (8.4) pour tout $k \geq 0$ et tous états i_0, i_1, \dots, i_k constitue la *distribution de probabilité* de la CMH. Donc :

Théorème 8.1.2 *La distribution de probabilité d'une CMH est entièrement déterminée par sa distribution initiale et sa matrice de transition.*

Notation. On abrègera $P(A \mid X_0 = i)$ en $P_i(A)$. Pour toute probabilité μ sur E , on notera $P_\mu(A) = \sum_{i \in E} \mu(i) P(A \mid X_0 = i)$. C'est la probabilité qui régit l'évolution de la chaîne lorsque la distribution initiale est μ .

Réurrences markoviennes

Nombre de CMH sont décrites par une équation de récurrence contrôlée par un “bruit blanc”. Plus précisément,

Théorème 8.1.3 *Soit $\{Z_n\}_{n \geq 1}$ une suite IID de variables aléatoires à valeurs dans un espace arbitraire G . Soit E un espace dénombrable, et soit une fonction $f : E \times G \rightarrow E$. Soit X_0 une variable aléatoire à valeurs dans E , indépendante de la suite $\{Z_n\}_{n \geq 1}$. L'équation de récurrence*

$$X_{n+1} = f(X_n, Z_{n+1}) \quad (8.5)$$

définit alors une CMH.

Démonstration. L'itération de (8.5) montre que pour tout $n \geq 1$, il existe une fonction g_n telle que $X_n = g_n(X_0, Z_1, \dots, Z_n)$, et donc $P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(f(i, Z_{n+1}) = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(f(i, Z_{n+1}) = j)$, puisque l'événement $\{X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i\}$ est exprimable en termes de X_0, Z_1, \dots, Z_n et est donc indépendant de Z_{n+1} . Semblablement, $P(X_{n+1} = j \mid X_n = i) = P(f(i, Z_{n+1}) = j)$. On a donc une chaîne de Markov, qui est de plus homogène, puisque le second membre de la dernière égalité ne dépend pas de n . Explicitement :

$$p_{ij} = P(f(i, Z_1) = j). \quad (8.6)$$

□

EXEMPLE 8.1.1: MARCHE ALÉATOIRE SUR \mathbb{Z} , TAKE 1. Soit X_0 une variable à valeurs dans \mathbb{Z} . Soit $\{Z_n\}_{n \geq 1}$ une suite de variables IID indépendante de X_0 , prenant les valeurs $+1$ ou -1 , et de distribution de probabilité commune

$$P(Z_n = +1) = p,$$

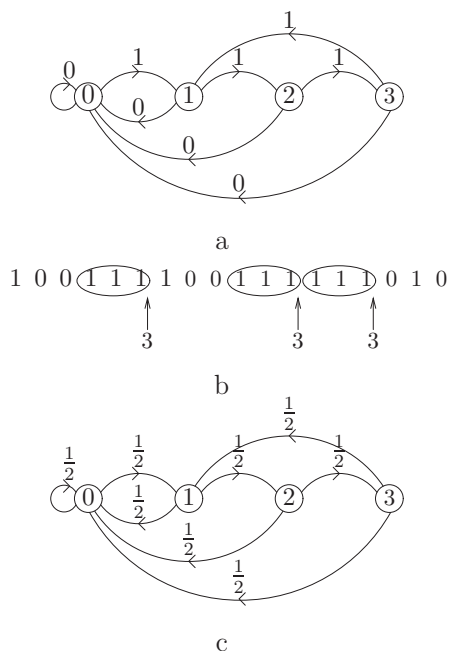
où $0 < p < 1$. Le processus $\{X_n\}_{n \geq 1}$ défini par

$$X_{n+1} = X_n + Z_{n+1}$$

est, au vu du Théorème 8.1.3, une CMH, appelée *marche aléatoire* sur \mathbb{Z} .

EXEMPLE 8.1.2: AUTOMATES STOCHASTIQUES. Un *automate fini* (E, \mathcal{A}, f) peut lire les lettres d'un *alphabet fini* \mathcal{A} écrites sur un ruban infini. Cet automate se trouve à un instant donné dans un quelconque des *états* d'un ensemble fini E , et son évolution est régie par une fonction $f : E \times \mathcal{A} \rightarrow E$ comme suit. Quand l'automate est dans l'état $i \in E$ et lit la lettre $a \in \mathcal{A}$, il passe de l'état i à l'état $j = f(i, a)$ et lit sur le ruban la prochaine lettre à droite de celle qu'il vient de lire.

Un automate est représenté par son graphe de transition G ayant pour nœuds les états de E . Il y a une arête orientée du nœud (état) i au nœud j si et seulement s'il existe $a \in \mathcal{A}$



Un automate stochastique

tel que $j = f(i, a)$, et cette arête reçoit alors l'étiquette a . Si $j = f(i, a_1) = f(i, a_2)$ pour $a_1 \neq a_2$, il y a alors 2 arêtes orientées de i vers j avec les étiquettes a_1 et a_2 , ou, plus économiquement, une seule arête orientée avec l'étiquette (a_1, a_2) . Plus généralement, une arête orientée donnée peut avoir des étiquettes multiples à tous ordres.

Considérons à titre d'exemple l'automate avec l'alphabet $\mathcal{A} = \{0, 1\}$ correspondant au graphe de transition de la Figure a. L'automate, initialisé dans l'état 0, lit la suite de la Figure b de gauche à droite, passant successivement dans les états (y compris l'état initial 0)

0 1 0 0 1 2 3 1 0 0 1 2 3 1 2 3 0 1 0 .

Il apparaît que l'automate est dans l'état 3 après avoir lu trois 1 successifs qui ne chevauchent pas un bloc de trois 1 consécutifs précédemment enregistré (voir Figure b), et seulement dans cette circonstance.

Si la suite de lettres lues par l'automate est $\{Z_n\}_{n \geq 1}$, la suite des états $\{X_n\}_{n \geq 0}$ est alors donnée par l'équation de récurrence $X_{n+1} = f(X_n, Z_{n+1})$ et donc, si $\{Z_n\}_{n \geq 1}$ est IID et indépendante de l'état initial X_0 , alors $\{X_n\}_{n \geq 1}$ est, au vu du Théorème 8.1.3, une CMH. Son graphe de transition est représenté dans la Figure c.

Toutes les CMH ne reçoivent pas une description "naturelle" du type de celle du Théorème 8.1.3. Cependant une telle description est toujours possible dans le sens suivant :

Théorème 8.1.4 À toute matrice stochastique \mathbf{P} sur E , on peut associer une CMH $\{X_n\}_{n \geq 0}$ avec cette matrice pour matrice de transition, et admettant une représentation comme dans le Théorème 8.1.3.

Démonstration. On identifie E à \mathbb{N} . On pose

$$X_{n+1} = j \text{ si } \sum_{k=0}^{j-1} p_{X_n k} < Z_{n+1} \leq \sum_{k=0}^j p_{X_n k},$$

où $\{Z_n\}_{n \geq 1}$ est IID, uniformément distribuée sur $(0, 1]$. On peut appliquer le Théorème 8.1.3, et on vérifie que cette CMH a la matrice de transition requise (Formule (8.6)). \square

Cette représentation est utile pour simuler de *petites* (avec un petit espace d'état) CMH. Elle a également un intérêt théorique. En particulier, lorsqu'on cherche à démontrer une propriété qui concerne la distribution d'une CMH, on peut toujours supposer que celle-ci admet une représentation comme dans le Théorème 8.1.3. Cette remarque sera utile à plusieurs reprises.

Toutes les CMH ne reçoivent pas de manière "naturelle" une description du type donné dans le Théorème 8.1.3. Une légère modification de ce théorème en élargit considérablement la portée.

Théorème 8.1.5 Considérons la situation du Théorème 8.1.3, sauf en ce qui concerne la distribution de X_0, Z_1, Z_2, \dots . On suppose à la place que F est dénombrable et que pour tout $n \geq 0$, Z_{n+1} est conditionnellement indépendant de $Z_n, \dots, Z_1, X_{n-1}, \dots, X_0$ sachant X_n , c'est-à-dire : pour tout $k, k_1, \dots, k_n \in F, i_0, i_1, \dots, i_{n-1}, i \in E$,

$$\begin{aligned} P(Z_{n+1} = k \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0, Z_n = k_n, \dots, Z_1 = k_1) \\ = P(Z_{n+1} = k \mid X_n = i), \end{aligned}$$

où la dernière quantité est indépendante de n . Alors, $\{X_n\}_{n \geq 0}$ est une CMH dont les probabilités de transition sont données par la formule :

$$p_{ij} = P(f(i, Z_1) = j \mid X_0 = i).$$

Démonstration. Exercice 8.5.3. \square

EXEMPLE 8.1.3: L'URNE D'EHRENFEST, TAKE 1. Ce modèle simplifié de diffusion à travers une membrane poreuse fut proposé en 1907 par les physiciens autrichiens Tattiana et Paul Ehrenfest pour décrire en termes de physique statistique les échanges de chaleur entre deux systèmes à différentes températures, et ainsi de mieux comprendre le phénomène d'irréversibilité thermodynamique (Exemple 8.3.1).

Il y a N particules qui peuvent se trouver dans le compartiment A ou le compartiment B . Supposons qu'au temps $n \geq 0$, $X_n = i$ particules sont dans A . On choisit alors une particule au hasard, et cette particule est alors transférée du compartiment où elle se trouvait dans l'autre. Le prochain état du système, X_{n+1} , est donc, soit $i-1$ (la particule

déplacée fut trouvée dans le compartiment A) avec la probabilité $\frac{i}{N}$, soit $i + 1$ (elle fut trouvée dans B) avec la probabilité $\frac{N-i}{N}$.

Ce modèle relève du Théorème 8.1.5. Pour tout $n \geq 0$,

$$X_{n+1} = X_n + Z_{n+1},$$

où $Z_n \in \{-1, +1\}$ et $P(Z_{n+1} = -1 \mid X_n = i) = \frac{i}{N}$. Les termes non nuls de la matrice de transition sont donc

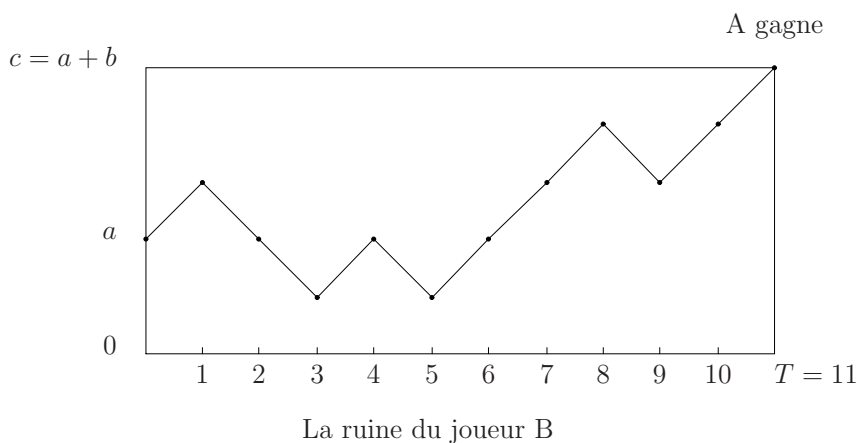
$$p_{i,i+1} = \frac{N-i}{N}, \quad p_{i,i-1} = \frac{i}{N}.$$

Analyse à un pas

Certaines fonctionnelles de CMH, en particulier les probabilités d'absorption par un ensemble d'états A fermé ($\sum_{j \in A} p_{ij} = 1$ pour tout $i \in A$) et les temps moyens avant absorption, peuvent être évalués grâce à la méthode appelée *analyse à un pas*. Cette technique, qui est le moteur de la plupart des calculs en théorie des chaînes de Markov, va être illustrée par les exemple suivants.

EXEMPLE 8.1.4: LA RUINE DU JOUEUR, TAKE 1. Deux joueurs A et B jouent à pile ou face, où la probabilité de face est $p \in (0, 1)$, et les tirages successifs sont IID. Si on note X_n la fortune du joueur A au temps n , alors $X_{n+1} = X_n + Z_{n+1}$, où $Z_{n+1} = +1$ (resp., -1) avec probabilité p (resp., $q = 1 - p$), et $\{Z_n\}_{n \geq 1}$ est IID. En d'autres mots, A parie 1 euro sur face à chaque lancer, et B parie la même chose sur pile. Le processus $\{X_n\}_{n \geq 1}$ est donc la marche aléatoire de l'Exemple 8.1.1. Les fortunes initiales de A et B sont respectivement a et b . Le jeu se termine dès qu'un des deux joueurs est ruiné.

La durée du jeu est T , le premier temps n auquel $X_n = 0$ ou c , et la probabilité que A gagne est $u(a) = P(X_T = c \mid X_0 = a)$.



Au lieu de calculer seulement $u(a)$, l'analyse à un pas calcule

$$u(i) = P(X_T = c \mid X_0 = i)$$

pour tout i , $0 \leq i \leq c$, et pour ce faire, elle génère une équation de récurrence pour les $u(i)$ en décomposant l'événement " A gagne" selon les éventualités du premier lancer, et en utilisant la règle des causes totales. Si $X_0 = i$, $1 \leq i \leq c - 1$, alors $X_1 = i + 1$ (resp., $X_1 = i - 1$) avec probabilité p (resp., q), et la probabilité de ruine de B sachant que A a une fortune initiale $i + 1$ (resp., $i - 1$) est $u(i + 1)$ (resp., $u(i - 1)$). Donc, pour tout i , $1 \leq i \leq c - 1$,

$$u(i) = pu(i + 1) + qu(i - 1),$$

avec les conditions aux frontières $u(0) = 0$, $u(c) = 1$.

L'équation caractéristique associée à cette équation de récurrence linéaire est $pr^2 - r + q = 0$. Elle a deux racines distinctes, $r_1 = 1$ et $r_2 = \frac{q}{p}$, si $p \neq q$, et une racine double, $r_1 = 1$, si $p = q = \frac{1}{2}$. La solution générale est donc $u(i) = \lambda r_1^i + \mu r_2^i = \lambda + \mu \left(\frac{q}{p}\right)^i$ quand $p \neq q$, et $u(i) = \lambda r_1^i + \mu i r_1^i = \lambda + \mu i$ lorsque $p = q = \frac{1}{2}$. Tenant compte des conditions aux frontières, on peut calculer λ et μ . Le résultat est, pour $p \neq q$,

$$u(i) = \frac{1 - \left(\frac{q}{p}\right)^i}{1 - \left(\frac{q}{p}\right)^c},$$

et pour $p = q = \frac{1}{2}$,

$$u(i) = \frac{i}{c}.$$

Dans le cas $p = q = \frac{1}{2}$, la probabilité $v(i)$ que B gagne quand la fortune initiale de B est $c - i$ est obtenue en remplaçant i par $c - i$ dans l'expression de $u(i)$. Ce qui donne $v(i) = \frac{c-i}{c} = 1 - \frac{i}{c}$. On vérifie que $u(i) + v(i) = 1$, ce qui signifie en particulier que la probabilité pour que le jeu ne s'éternise pas est 1. Le lecteur est invité à vérifier qu'il en est de même pour $p \neq q$.

EXEMPLE 8.1.5: LA RUINE DU JOUEUR, TAKE 2. (suite de l'Exemple 8.1.4) La durée $m(i) = E[T \mid X_0 = i]$ du jeu quand la fortune initiale du joueur A est i vérifie l'équation de récurrence

$$m(i) = 1 + pm(i + 1) + qm(i - 1)$$

lorsque $1 \leq i \leq c - 1$. En effet, la pièce doit être lancée au moins une fois, et avec la probabilité p (resp., q) la fortune de A sera $i + 1$ (resp., $i - 1$), et donc $m(i + 1)$ (resp., $m(i - 1)$) lancers supplémentaires seront en moyenne nécessaires pour finir le jeu. Les conditions aux frontières sont

$$m(0) = 0, m(c) = 0.$$

Réécrivons l'équation de récurrence sous la forme $-1 = p(m(i+1) - m(i)) - q(m(i) - m(i-1))$. Notant

$$y_i = m(i) - m(i-1),$$

on a, pour $1 \leq i \leq c-1$,

$$-1 = py_{i+1} - qy_i$$

et

$$m(i) = y_1 + y_2 + \cdots + y_i.$$

Nous allons résoudre cette équation de récurrence avec $p = q = \frac{1}{2}$. Pour cela, on la réécrit

$$\begin{aligned} -1 &= \frac{1}{2}y_2 - \frac{1}{2}y_1, \\ -1 &= \frac{1}{2}y_3 - \frac{1}{2}y_2, \\ &\vdots \\ -1 &= \frac{1}{2}y_i - \frac{1}{2}y_{i-1}, \end{aligned}$$

et donc, en sommant,

$$-(i-1) = \frac{1}{2}y_i - \frac{1}{2}y_1,$$

c'est-à-dire, pour $1 \leq i \leq c$,

$$y_i = y_1 - 2(i-1).$$

Reportant cette expression dans $m(i) = y_1 + y_2 + \cdots + y_i$, et observant que $y_1 = m(1)$, on obtient

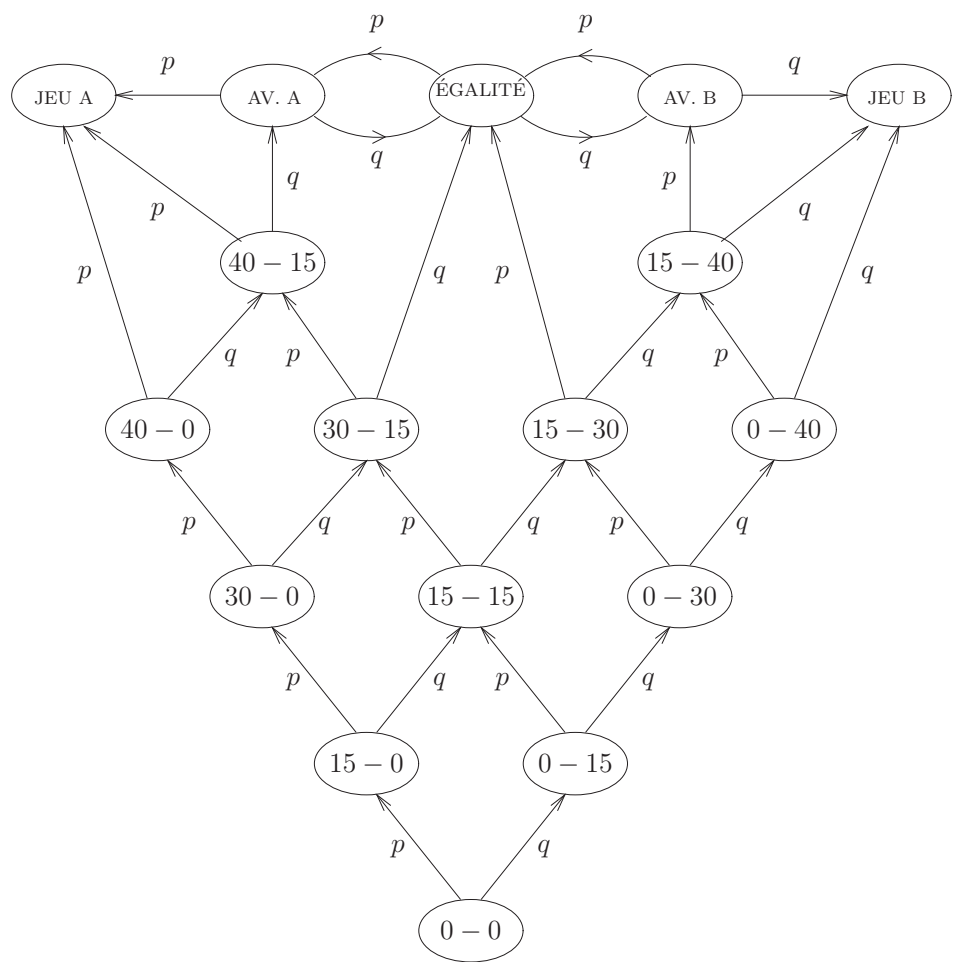
$$m(i) = im(1) - 2[1 + 2 + \cdots + (i-1)] = im(1) - i(i-1).$$

La condition $m(c) = 0$ donne $cm(1) = c(c-1)$ et donc, finalement,

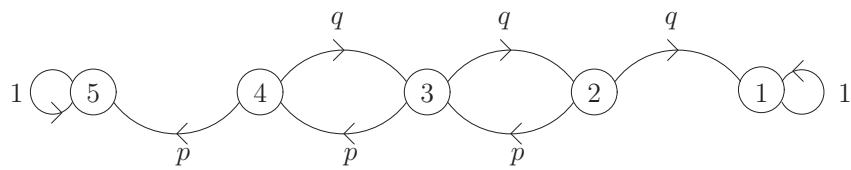
$$m(i) = i(c-i).$$

EXEMPLE 8.1.6: RÉCRÉATION : TENNIS. Au tennis, si on ignore les “tie-breaks”, un “jeu” peut être modélisé par une CMH dont le graphe de transition est donné par la figure ci-dessous, où p est la probabilité que le serveur A gagne le point, et $q = 1 - p$. On veut calculer la probabilité pour que B remporte le jeu.

On voit sur la figure qu'un jeu comporte deux étapes : on atteint d'abord un des états supérieurs du graphe et ensuite, on reste dans les états supérieurs jusqu'à l'absorption en “jeu pour A ” ou “jeu pour B .” En modifiant les noms des états supérieurs, le graphe de la CMH de la deuxième étape est donné par la Figure b.



a



b

Tennis : un jeu sans la règle du *tie-break*

L'analyse à un pas donne pour b_i , la probabilité pour que B gagne sachant que la deuxième étape débute en i ,

$$b_1 = 1, \quad b_5 = 0$$

et

$$\begin{aligned} b_2 &= q + pb_3, \\ b_3 &= qb_2 + pb_4, \\ b_4 &= qb_3. \end{aligned}$$

pour $p \neq q$,

$$(b_1, b_2, b_3, b_4, b_5) = \left(1, q \frac{1-pq}{1-2pq}, \frac{q^2}{1-2pq}, \frac{q^3}{1-2pq}, 0 \right).$$

Si on débute en 0-0, B gagne avec la probabilité $\sum_{i=1}^5 p(i)b_i$, où $p(i)$ est la probabilité que la deuxième étape débute en i . Une simple énumération des chemins allant de 0-0 à l'état supérieur 1 donne $p(1) = q^4 + q^3pq + q^2pq^2 + qpq^3 + pq^4$, c'est-à-dire,

$$p(1) = q^4(1 + 4p).$$

Le lecteur pourra terminer les calculs.

Distribution stationnaire

Définition 8.1.2 Une distribution de probabilité π satisfaisant l'équation de balance globale

$$\pi^T = \pi^T \mathbf{P} \tag{8.7}$$

est appelée distribution stationnaire de la matrice de transition \mathbf{P} , ou de la CMH.

L'équation de balance globale dit que pour tout état i ,

$$\pi(i) = \sum_{j \in E} \pi(j)p_{ji}.$$

L'itération de (8.7) donne $\pi^T = \pi^T \mathbf{P}^n$ pour tout $n \geq 0$, et donc, au vu de (8.3), si la distribution initiale $\nu = \pi$, alors $\nu_n = \pi$ pour tout $n \geq 0$: si une chaîne a pour distribution initiale la distribution stationnaire, elle garde cette distribution pour toujours. Mais il y a plus. En effet, dans ce cas,

$$\begin{aligned} P(X_n = i_0, X_{n+1} = i_1, \dots, X_{n+k} = i_k) &= P(X_n = i_0)p_{i_0 i_1} \dots p_{i_{k-1} i_k} \\ &= \pi(i_0)p_{i_0 i_1} \dots p_{i_{k-1} i_k} \end{aligned}$$

ne dépend plus n . C'est dans ce sens qu'on dit que la chaîne est *stationnaire*. On dit aussi qu'elle se trouve en *régime stationnaire*, ou en *équilibre*. En résumé :

Théorème 8.1.6 *Si la distribution initiale d'une CMH est la distribution stationnaire, la CMH est stationnaire.*

L'équation de balance globale $\pi^T \mathbf{P} = \pi^T$, et la relation $\pi^T \mathbf{1} = 1$ (où $\mathbf{1}$ est un vecteur colonne dont toutes les coordonnées sont égales à 1) qui exprime que π est une probabilité, donnent, lorsque E est fini, $|E| + 1$ équations avec $|E|$ variables. Une de ces $|E|$ équations est superflue sous la contrainte $\pi^T \mathbf{1} = 1$. En effet, si on fait la somme des équations de $\pi^T \mathbf{P} = \pi^T$ on obtient $\pi^T \mathbf{P} \mathbf{1} = \pi^T \mathbf{1}$, c'est-à-dire, $\pi^T \mathbf{1} = 1$.

EXEMPLE 8.1.7: CHAÎNE À 2 ÉTATS. L'espace d'état est $E = \{1, 2\}$ et la matrice de transition est

$$\mathbf{P} = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix},$$

où $\alpha, \beta \in (0, 1)$. Les équations de balance globale sont

$$\begin{aligned} \pi(1) &= \pi(1)(1 - \alpha) + \pi(2)\beta, \\ \pi(2) &= \pi(1)\alpha + \pi(2)(1 - \beta). \end{aligned}$$

Ce système dépendant se réduit à une seule équation $\pi(1)\alpha = \pi(2)\beta$, à laquelle il faut ajouter $\pi(1) + \pi(2) = 1$ qui exprime que π est une probabilité. On obtient

$$\pi(1) = \frac{\beta}{\alpha + \beta}, \quad \pi(2) = \frac{\alpha}{\alpha + \beta}.$$

EXEMPLE 8.1.8: L'URNE D'EHRENFEST, TAKE 2. (suite de l'Exemple 8.1.3) Les équations de balance globale sont pour i , $1 \leq i \leq N - 1$,

$$\pi(i) = \pi(i - 1) \left(1 - \frac{i - 1}{N} \right) + \pi(i + 1) \frac{i + 1}{N}$$

et pour les états extrêmes,

$$\pi(0) = \pi(1) \frac{1}{N}, \quad \pi(N) = \pi(N - 1) \frac{1}{N}.$$

Laissant $\pi(0)$ indéterminé, on résout les équations pour $i = 0, 1, \dots, N$ successivement pour obtenir

$$\pi(i) = \pi(0) \binom{N}{i}.$$

La quantité $\pi(0)$ est alors déterminée en écrivant que π est un vecteur de probabilité :

$$1 = \sum_{i=0}^N \pi(i) = \pi(0) \sum_{i=0}^N \binom{N}{i} = \pi(0) 2^N.$$

Ceci donne pour π distribution binomiale de taille N et de paramètre $\frac{1}{2}$:

$$\pi(i) = \frac{1}{2^N} \binom{N}{i}.$$

C'est la distribution qu'on obtiendrait en plaçant indépendamment les particules dans les compartiments avec la probabilité $\frac{1}{2}$ pour chaque compartiment.

Les distributions stationnaires peuvent être nombreuses. Ainsi, si on prend la matrice de transition qui est la matrice unité, toute probabilité sur l'espace d'état est une probabilité stationnaire. Il se peut aussi qu'il n'y ait pas de probabilité stationnaire (voir l'Exercice 8.5.10).

EXEMPLE 8.1.9: PROCESSUS DE NAISSANCE ET DE MORT, I. De tels processus généralisent le modèle de diffusion d'Ehrenfest. L'espace d'état est $E = \{0, 1, \dots, N\}$, et la matrice de transition :

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & & & & & \\ q_1 & r_1 & p_1 & & & & \\ & q_2 & r_2 & p_2 & & & \\ & & \ddots & & & & \\ & & & q_i & r_i & p_i & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & q_{N-1} & r_{N-1} & p_{N-1} \\ & & & & & & 1 & 0 \end{pmatrix},$$

où $p_i > 0$ et $q_i > 0$ pour tout état i tel que $1 \leq i \leq N-1$. Les équations de balance globale pour les états $i \neq 0, N$ sont :

$$\pi(i) = p_{i-1}\pi(i-1) + r_i\pi(i) + q_{i+1}\pi(i+1),$$

et pour les états 0 et N :

$$\pi(0) = \pi(1)q_1, \quad \pi(N) = \pi(N-1)p_{N-1}$$

(barrières réfléchissantes). En remarquant que $r_i = 1 - p_i - q_i$ et en regroupant des termes, on a, pour $2 \leq i \leq N-1$,

$$\pi(i+1)q_{i+1} - \pi(i)p_i = \pi(i)q_i - \pi(i-1)p_{i-1}$$

et

$$\begin{aligned} \pi(1)q_1 - \pi(0) &= 0, \\ \pi(2)q_2 - \pi(1)p_1 &= \pi(1)q_1 - \pi(0). \end{aligned}$$

Donc, $\pi(1)q_1 = \pi(0)$, et pour $2 \leq i \leq N-1$,

$$\pi(i)q_i = \pi(i-1)p_{i-1}.$$

Ceci donne

$$\pi(1) = \pi(0) \frac{1}{q_1},$$

et pour $2 \leq i \leq N$,

$$\pi(i) = \pi(0) \frac{p_1 p_2 \cdots p_{i-1}}{q_1 q_2 \cdots q_i}. \quad (8.8)$$

L'inconnue $\pi(0)$ est déterminée par l'égalité $\sum_{i=0}^N \pi(i) = 1$ (π est une probabilité), ce qui donne,

$$\pi(0) \left\{ 1 + \frac{1}{q_1} + \frac{p_1}{q_1 q_2} + \cdots + \frac{p_1 p_2 \cdots p_{N-1}}{q_1 q_2 \cdots q_{N-1} q_N} \right\} = 1. \quad (8.9)$$

Retournement du temps

Les notions de retournement de temps et de réversibilité sont très productives en théorie des chaînes de Markov.

Soit $\{X_n\}_{n \geq 0}$ une CMH de matrice de transition \mathbf{P} admettant une distribution stationnaire π positive ($\pi(i) > 0$ pour tout état i). La matrice \mathbf{Q} , indexée par E , définie par

$$\pi(i)q_{ij} = \pi(j)p_{ji}, \quad (8.10)$$

est une matrice stochastique. En effet,

$$\sum_{j \in E} q_{ij} = \sum_{j \in E} \frac{\pi(j)}{\pi(i)} p_{ji} = \frac{1}{\pi(i)} \sum_{j \in E} \pi(j) p_{ji} = \frac{\pi(i)}{\pi(i)} = 1,$$

où la troisième égalité tient compte des équations de balance globale. Elle reçoit l'interprétation suivante : supposons que la distribution initiale soit π , auquel cas pour tout $n \geq 0$, tout $i \in E$,

$$P(X_n = i) = \pi(i).$$

Alors la formule de rétrodition de Bayes donne

$$P(X_n = j \mid X_{n+1} = i) = \frac{P(X_{n+1} = i \mid X_n = j)P(X_n = j)}{P(X_{n+1} = i)},$$

c'est-à-dire, au vu de (8.10),

$$P(X_n = j \mid X_{n+1} = i) = q_{ji}.$$

On voit que \mathbf{Q} est la matrice de transition de la chaîne quand on "retourne le temps".

L'observation qui suit est promue au rang de théorème à cause de son efficacité.

Théorème 8.1.7 Soit \mathbf{P} une matrice stochastique indexée par l'ensemble dénombrable E , et soit π une distribution de probabilité positive sur E . Si la matrice \mathbf{Q} indexée par E et définie par (8.10) est une matrice stochastique (la somme de chacune des lignes est égale à 1), alors π est une distribution stationnaire de \mathbf{P} .

Démonstration. Pour $i \in E$ fixé, on somme les égalités (8.10) par rapport à $j \in E$, ce qui donne

$$\sum_{j \in E} \pi(i) q_{ij} = \sum_{j \in E} \pi(j) p_{ji}.$$

Mais le membre de gauche de cette égalité égale $\pi(i) \sum_{j \in E} q_{ij} = \pi(i)$, et donc, pour tout $i \in E$,

$$\pi(i) = \sum_{j \in E} \pi(j) p_{ji}.$$

□

Définition 8.1.3 On appelle réversible toute chaîne de Markov de distribution initiale π (une distribution stationnaire) positive telle que pour tout $i, j \in E$,

$$\pi(i) p_{ij} = \pi(j) p_{ji}. \quad (8.11)$$

Dans ce cas, $q_{ij} = p_{ij}$, et donc la chaîne et la chaîne retournée ont la même distribution puisque celle-ci est entièrement déterminée par la distribution initiale et la matrice de transition. Les équations (8.11) sont appelées les *équations de balance détaillée*. Le résultat suivant est un corollaire immédiat du Théorème 8.1.7.

Théorème 8.1.8 Soit \mathbf{P} une matrice de transition sur E , et soit π une distribution de probabilité sur E . Si pour tout $i, j \in E$, les équations de balance détaillée (8.11) sont vérifiées, alors π est une distribution stationnaire de \mathbf{P} .

EXEMPLE 8.1.10: L'URNE D'EHRENFEST, TAKE 3. (suite des Exemples 8.1.3 et 8.1.8)
On rappelle qu'on avait obtenu l'expression

$$\pi(i) = \frac{1}{2N} \binom{N}{i}$$

pour la distribution stationnaire. La vérification des équations de balance détaillée

$$\pi(i) p_{i,i+1} = \pi(i+1) p_{i+1,i}$$

est immédiate. L'urne d'Ehrenfest est donc réversible.

EXEMPLE 8.1.11: MARCHE ALÉATOIRE SUR UN GRAPHE. Soit un graphe non orienté où on note E l'ensemble de ses sommets, ou nœuds. Soit d_i le nombre d'arêtes adjacentes

au sommet i . Transformons ce graphe en un graphe orienté en décomposant chacune de ses arêtes en deux arêtes orientées de directions opposées, et faisons en un graphe de transition en associant à l'arête orientée de i vers j la probabilité de transition $\frac{1}{d_i}$.

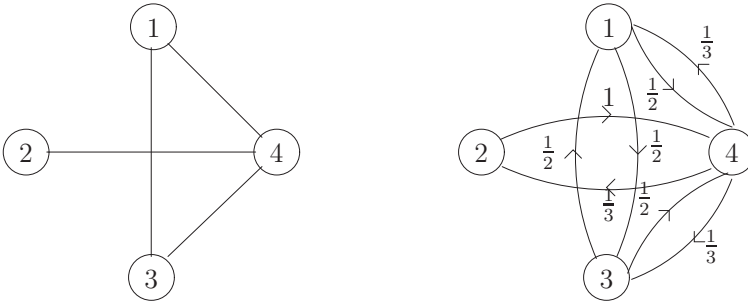
On supposera que $d_i > 0$ pour tout état i (pas de nœud isolé). Une distribution stationnaire (en fait, la distribution stationnaire comme nous le verrons bientôt) est donnée par

$$\pi(i) = \frac{d_i}{\sum_{j \in E} d_j}.$$

Pour cela on utilise le Théorème 8.1.8, en faisant le pari que la chaîne est réversible. Il nous faut juste vérifier que

$$\pi(i) \frac{1}{d_i} = \pi(j) \frac{1}{d_j},$$

ce qui est bien le cas.



Marche aléatoire sur un graphe

Communication

Les propriétés que l'on va définir dans cette fin de section (communication et période) sont purement *topologiques*, en ce sens qu'elles concernent le graphe de transition "nu", c'est-à-dire sans les étiquettes indiquant les probabilités de transition.

Définition 8.1.4 *L'état j est dit accessible depuis l'état i s'il existe un entier $M \geq 0$ tel que $p_{ij}(M) > 0$. En particulier, un état i est toujours accessible depuis lui-même, puisque $p_{ii}(0) = 1$. Les états i et j sont dits communiquer si i est accessible depuis j et si j est accessible depuis i . On note ceci $i \leftrightarrow j$.*

Pour $M \geq 1$, $p_{ij}(M) = \sum_{i_1, \dots, i_{M-1}} p_{ii_1} \cdots p_{i_{M-1}j}$, et donc $p_{ij}(M) > 0$ si et seulement si il existe au moins un chemin $i, i_1, \dots, i_{M-1}, j$ de i à j tel que

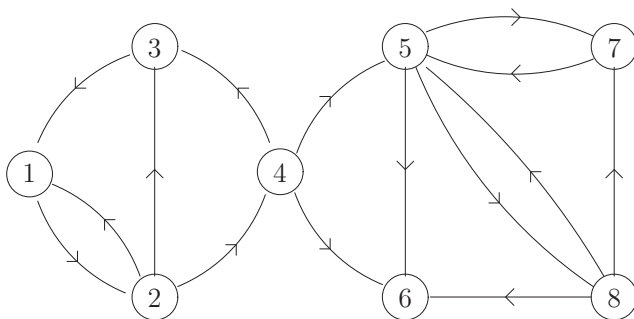
$$p_{ii_1} p_{i_1 i_2} \cdots p_{i_{M-1} j} > 0,$$

ou, de manière équivalente, si il existe un chemin orienté de i vers j dans le graphe de transition G . On vérifie immédiatement que

$$\begin{aligned} i &\leftrightarrow i && \text{(reflexivité),} \\ i &\leftrightarrow j \Rightarrow j &\leftrightarrow i & \text{(symétrie),} \\ i &\leftrightarrow j, j &\leftrightarrow k \Rightarrow i &\leftrightarrow k \text{ (transitivité).} \end{aligned}$$

Donc la relation de communication (\leftrightarrow) est une relation d'équivalence. Elle engendre une partition de l'espace d'état E en classes d'équivalences appelées *classes de communication*.

Définition 8.1.5 Un état i tel que $p_{ii} = 1$ est dit fermé. Plus généralement, un ensemble C d'états tel que $\sum_{j \in C} p_{ij} = 1$ pour tout $i \in C$ est dit fermé.



Un graphe de transition avec 3 classes de communications

EXEMPLE 8.1.12: UN GRAPHE DE TRANSITION. Le graphe de transition de la figure ci-dessous a trois classes de communication : $\{1, 2, 3, 4\}$, $\{5, 7, 8\}$, et $\{6\}$. L'état 6 est fermé. La classe de communication $\{5, 7, 8\}$ n'est pas fermée, mais l'ensemble $\{5, 6, 7, 8\}$ l'est.

On observe sur cet exemple qu'il peut y avoir des arêtes orientées reliant deux classes de communication différentes E_k et E_ℓ . Cependant, toutes les arêtes orientées entre deux classes de communication données ont la même orientation, toutes de E_k à E_ℓ ou toutes de E_ℓ à E_k .

Définition 8.1.6 S'il n'y a qu'une seule classe de communication, la chaîne, sa matrice de transition et son graphe de transition sont dits irréductibles.

Période

Considérons la marche aléatoire sur \mathbb{Z} (Exemple 8.1.1). Comme $0 < p < 1$, elle est irréductible. On observe que $E = C_0 + C_1$, où C_0 et C_1 , l'ensemble des entiers relatifs pairs et impairs respectivement, ont la propriété suivante. Si on part de $i \in C_0$ (resp., C_1), on ne peut se rendre en une seule étape que dans un état $j \in C_1$ (resp., C_0). La chaîne passe alternativement d'une classe à l'autre. En ce sens, la chaîne a un comportement périodique, correspondant à la période 2. Voici la définition exacte.

Définition 8.1.7 La période d_i de l'état $i \in E$ est par définition,

$$d_i = \text{PGCD}\{n \geq 1 ; p_{ii}(n) > 0\},$$

avec la convention $d_i = +\infty$ s'il n'y a pas d'indice $n \geq 1$ tel que $p_{ii}(n) > 0$ (on ne revient pas en i). Si $d_i = 1$, l'état i est dit apériodique.

Théorème 8.1.9 Si les états i et j communiquent, ils ont la même période.

Démonstration. Comme i et j communiquent, il existe des entiers M et N tels que $p_{ij}(M) > 0$ et $p_{ji}(N) > 0$. Pour tout $k \geq 1$,

$$p_{ii}(M + nk + N) \geq p_{ij}(M)(p_{jj}(k))^n p_{ji}(N)$$

(en effet, un chemin tel que $X_0 = i, X_M = j, X_{M+k} = j, \dots, X_{M+nk} = j, X_{M+nk+N} = i$ est un des moyens parmi d'autres d'aller de i à i en $M + nk + N$ étapes).

Donc, pour tout $k \geq 1$ tel que $p_{jj}(k) > 0$, on a $p_{ii}(M + nk + N) > 0$ pour tout $n \geq 1$. Par conséquent, d_i divise $M + nk + N$ pour tout $n \geq 1$, et en particulier, d_i divise k . On a donc montré que d_i divise tout k tel que $p_{jj}(k) > 0$, et en particulier, d_i divise d_j . Par symétrie, d_j divise d_i , et donc, finalement $d_i = d_j$. \square

Définition 8.1.8 Si la chaîne est irréductible, la période d commune à tous les états est appelée la période de \mathbf{P} , ou de la chaîne. Si $d = 1$, la matrice de transition et la chaîne sont dites apériodiques.

Théorème 8.1.10 Soit \mathbf{P} une matrice stochastique irréductible, de période d . Alors, pour tous états $i, j \in E$ il existe des entiers (dépendant de i, j) $m \geq 0$ et $n_0 \geq 0$ tels que

$$p_{ij}(m + nd) > 0, \text{ pour tout } n \geq n_0.$$

Démonstration. Il suffit de prouver le théorème pour $i = j$. En effet, il existe m tel que $p_{ij}(m) > 0$, parce que j est accessible depuis i , la chaîne étant irréductible, et donc, si pour un $n_0 \geq 0$ on a $p_{jj}(nd) > 0$ pour tout $n \geq n_0$, alors $p_{ij}(m + nd) > p_{ij}(m)p_{jj}(nd) > 0$ pour tout $n \geq n_0$.

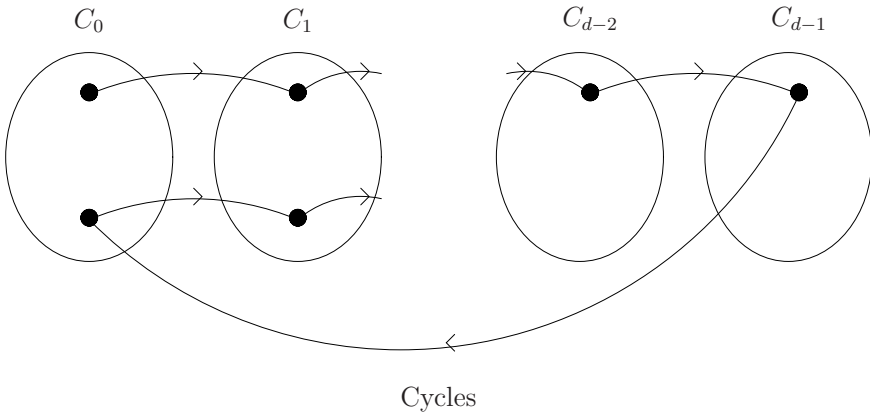
Le reste de la preuve découle d'un résultat classique de la théorie des nombres : tout ensemble A d'entiers positifs qui est fermé par addition et de PGCD d (comme c'est bien le cas pour $A = \{k \geq 1; p_{jj}(k) > 0\}$) contient tous les multiples de d , sauf peut-être un nombre fini d'entre eux. En d'autres termes, dans le cas qui nous intéresse, il existe n_0 tel que $n > n_0$ entraîne $p_{jj}(nd) > 0$. \square

Au vu du résultat précédent, il est clair que pour une CMH *irréductible* de période d , on peut trouver une unique partition de E en d classes C_0, C_1, \dots, C_{d-1} telles que pour tout $k, 0 \leq k \leq d$, et tout $i \in C_k$,

$$\sum_{j \in C_{k+1}} p_{ij} = 1,$$

où par convention $C_d = C_0$. Les classes C_0, C_1, \dots, C_{d-1} sont les *classes cycliques*.

Considérons une CMH irréductible de période d dont les classes cycliques sont C_0, C_1, \dots, C_d . En renumérotant les états de E si nécessaire, la matrice de transition a la structure de blocs ci-dessous (où on a choisi $d = 4$ pour être explicite),



$$\mathbf{P} = \begin{matrix} & C_0 & C_1 & C_2 & C_3 \\ \begin{matrix} C_0 \\ C_1 \\ C_2 \\ C_3 \end{matrix} & \begin{pmatrix} 0 & A_0 & 0 & \\ 0 & 0 & A_1 & 0 \\ 0 & 0 & 0 & A_2 \\ A_3 & 0 & 0 & 0 \end{pmatrix} \end{matrix},$$

et donc \mathbf{P}^2 , \mathbf{P}^3 , et \mathbf{P}^4 ont aussi une structure de blocs correspondant aux classes C_0, C_1, C_2, C_3 :

$$\mathbf{P}^2 = \begin{pmatrix} 0 & 0 & B_0 & 0 \\ 0 & 0 & 0 & B_1 \\ B_2 & 0 & 0 & 0 \\ 0 & B_3 & 0 & 0 \end{pmatrix} \quad \mathbf{P}^3 = \begin{pmatrix} 0 & 0 & 0 & D_0 \\ D_1 & 0 & 0 & 0 \\ 0 & D_2 & 0 & 0 \\ 0 & 0 & D_3 & 0 \end{pmatrix},$$

et finalement

$$\mathbf{P}^4 = \begin{pmatrix} E_0 & 0 & 0 & 0 \\ 0 & E_1 & 0 & 0 \\ 0 & 0 & E_2 & 0 \\ 0 & 0 & 0 & E_3 \end{pmatrix}.$$

On observe deux phénomènes : le décalage des blocs et le fait que \mathbf{P}^4 est bloc-diagonale. Ceci est évidemment général : \mathbf{P}^d est bloc-diagonale relativement aux classes cycliques C_0, C_1, \dots, C_{d-1} . La matrice de transition à d étapes, \mathbf{P}^d , est aussi une matrice stochastique, et les classes cycliques sont des classes de communication de \mathbf{P}^d comme le montre la forme bloc-diagonale de cette dernière matrice.

8.2 Récurrence

Cette section et la suivante sont consacrées à l'étude du comportement à long terme des chaînes de Markov homogènes. Les notions importantes sont celles d'*état récurrent* et d'*état transitoire*. Soit $\{X_n\}_{n \geq 0}$ une CMH de matrice de transition \mathbf{P} . On définit le *temps de retour* en i par

$$T_i = \inf\{n \geq 1; X_n = i\},$$

avec la convention usuelle : $\inf \emptyset = \infty$ (Ici : $T_i = \infty$ si $X_n \neq i$ pour tout $n \geq 1$).

Définition 8.2.1 *On dit que l'état i est récurrent si $P_i(T_i < \infty) = 1$. Un état i récurrent est dit récurrent positif si de plus $E_i[T_i] < \infty$, récurrent nul si $E_i[T_i] = \infty$. Un état qui n'est pas récurrent est dit transitoire.*

Pour un état i fixé, notons $\{\tau_k\}_{k \geq 1}$ la suite des temps de retour successifs dans l'état i . Formellement :

$$\begin{aligned} \tau_1 &= T_i \\ \dots \\ \tau_{k+1} &= \inf\{n > \tau_k; X_n = i\} \\ \dots \end{aligned}$$

Le résultat suivant est intuitif. Sa démonstration fait l'objet de l'Exercice 8.5.14.

Théorème 8.2.1 *Sachant que $\tau_k < \infty$, le processus $\{X_{\tau_k+n}\}_{n \geq 0}$ est une CMH de matrice de transition \mathbf{P} indépendante de $\{X_{n \wedge \tau_k}\}_{n \geq 0}$.*

En particulier, si l'état i est récurrent, les "cycles" $\{X_{n+\tau_k}\}_{0 \leq n < \tau_{k+1}$, $k \geq 1$, sont indépendants et identiquement distribués.

Notons $f_{ji} = P_j(T_i < \infty)$ la probabilité de retourner en i en partant de j , et notons

$$N_i = \sum_{n=1}^{\infty} 1_{\{X_n=i\}}$$

le nombre total de visites en i à partir de $n = 1$. On a

$$P_j(N_i = 0) = 1 - f_{ji}$$

et, lorsque $r \geq 1$,

$$P_j(N_i = r) = f_{ji} \times (f_{ji})^{r-1} \times (1 - f_{ji}).$$

En effet, en utilisant le Théorème 8.2.1, cette probabilité est égale à la probabilité de visiter i au moins une fois (f_{ji}) multiplié par la probabilité de visiter i encore $r - 1$ fois ($(f_{ji})^{r-1}$) en succession, et enfin après la r -ème visite, de ne plus jamais revenir en i (probabilité $1 - f_{ji}$). En particulier :

$$P_i(N_i = r) = (f_{ii})^r \times (1 - f_{ii}).$$

On en déduit que :

(a) Si $f_{ii} = 1$, $P_i(N_i = r) = 0$ pour tout $r \geq 0$, et donc $P_i(N_i = \infty) = 1$, et bien entendu $E_i[N_i] = \infty$.

(b) Si $f_{ii} < 1$, on a $\sum_{r=0}^{\infty} P_i(N_i = r) = 1$, et donc $P_i(N_i = \infty) = 0$; d'autre part un calcul élémentaire donne : $E_i[N_i] = \frac{f_{ii}}{1-f_{ii}} < \infty$.

En particulier :

Théorème 8.2.2 *Pour que l'état i soit récurrent, il faut et il suffit que*

$$\sum_{n=1}^{\infty} p_{ii}(n) = \infty.$$

Démonstration. Les remarques précédant l'énoncé du théorème montrent que $f_{ii} = 1 \Leftrightarrow E_i[N_i] = \infty$. D'autre part,

$$\begin{aligned} E_i[N_i] &= E_i \left[\sum_{n=1}^{\infty} 1_{\{X_n=i\}} \right] \\ &= \sum_{n=1}^{\infty} E_i[1_{\{X_n=i\}}] \\ &= \sum_{n=1}^{\infty} P_i(X_n = i) = \sum_{n=1}^{\infty} p_{ii}(n). \end{aligned}$$

□

Le critère de récurrence ci-dessus s'appelle le *critère de la matrice potentiel* car $\sum_{n=0}^{\infty} p_{ii}(n)$ est le terme (i, i) de la *matrice potentiel*

$$\mathbf{G} = \sum_{n=0}^{\infty} \mathbf{P}^n.$$

EXEMPLE 8.2.1: MARCHE ALÉATOIRE SUR \mathbb{Z} , TAKE 2. (suite de l'Exercice 8.1.1) Comme $0 < p < 1$, cette chaîne est irréductible et tous ses états sont de même nature (transitoires ou récurrents). Considérons par exemple l'état 0. On a $p_{00}(2n+1) = 0$ et

$$p_{00}(2n) = \frac{(2n)!}{n!n!} p^n (1-p)^n.$$

D'après la formule d'équivalence de Stirling ($n! \sim (n/e)^n \sqrt{2\pi n}$), la quantité ci-dessus est équivalente à

$$\frac{[4p(1-p)]^n}{\sqrt{\pi n}}. \quad (8.12)$$

La nature de la série $\sum_{n=0}^{\infty} p_{00}(n)$ est donc la même que celle de la série de terme général (8.12). Si $p \neq \frac{1}{2}$, auquel cas $4p(1-p) < 1$, cette dernière converge, et si $p = \frac{1}{2}$, auquel cas $4p(1-p) = 1$, elle diverge. Donc, la marche aléatoire sur \mathbb{Z} est transiente si $p \neq \frac{1}{2}$, récurrente si $p = \frac{1}{2}$.

Les exemples où l'on peut utiliser le critère de la matrice potentiel pour vérifier si un état est récurrent ou non sont rares. Ce critère va cependant nous servir à démontrer l'important résultat suivant, à savoir que la récurrence est une "propriété de classe" (de communication).

Théorème 8.2.3 *Soit une CMH de matrice de transition \mathbf{P} . Deux états qui communiquent sont, soit tous deux récurrents, soit tous deux transitoires. En particulier, si \mathbf{P} est irréductible, les états sont soit tous récurrents, soit tous transitoires. La chaîne (ou sa matrice de transition) est alors dite, respectivement, récurrente, transiente.*

Démonstration. Si i et j communiquent, il existe M et N tels que $p_{ij}(M) > 0$ et $p_{ji}(N) > 0$. Posons $\alpha = p_{ij}(M)p_{ji}(N) > 0$. On a l'inégalité

$$p_{ii}(M+n+N) \geq p_{ij}(M)p_{jj}(n)p_{ji}(N) = \alpha p_{jj}(n).$$

(En effet, le premier membre est la probabilité d'aller de i en j en exactement $M+n+N$ étapes, tandis que le second membre est la probabilité d'aller de i en j en exactement $M+n+N$ étapes, mais de façon particulière, avec la contrainte que $X_M = j$ et $X_{M+n} = j$.) De même :

$$p_{jj}(N+n+M) \geq \alpha p_{ii}(n).$$

Donc, si i et j communiquent, les séries $\sum_{n=1}^{\infty} p_{ii}(n)$ et $\sum_{n=1}^{\infty} p_{jj}(n)$ ont le même comportement. Les états i et j sont donc, d'après le Théorème 8.2.2, ou bien tous deux transitoires, ou bien tous deux récurrents. \square

Critère de la probabilité stationnaire

Le but que l'on se fixe maintenant est de démontrer un critère plus maniable que celui de la matrice potentiel, à savoir, le *critère de la probabilité stationnaire*, qui dit qu'une CMH irréductible est récurrente positive si et seulement si elle admet une distribution stationnaire.

La démonstration passe par des préliminaires concernant la *mesure invariante* d'une CMH irréductible récurrente.

Définition 8.2.2 *Un vecteur non nul $x = \{x_i\}_{i \in E}$ de coordonnées non négatives est appelé mesure invariante de la matrice stochastique $\mathbf{P} = \{p_{ij}\}_{i,j \in E}$ si pour tout $i \in E$,*

$$x_i = \sum_{j \in E} x_j p_{ji}. \quad (8.13)$$

(En notation abrégée, $0 \leq x < \infty$ et $x^T \mathbf{P} = x^T$.)

Théorème 8.2.4 *A. Soit \mathbf{P} la matrice de transition d'une CMH irréductible récurrente $\{X_n\}_{n \geq 0}$. Soit 0 un état arbitraire et soit T_0 le temps de retour en 0. On définit, pour tout $i \in E$, la quantité*

$$x_i = E_0 \left[\sum_{n=1}^{T_0} 1_{\{X_n=i\}} \right]. \quad (8.14)$$

Alors, pour tout $i \in E$,

$$0 < x_i < \infty, \quad (8.15)$$

et x est une mesure invariante de \mathbf{P} .

B. La mesure invariante d'une matrice stochastique irréductible récurrente est unique à une constante multiplicative près.

C. Une CMH irréductible récurrente est récurrente positive si et seulement si sa mesure invariante x vérifie

$$\sum_{i \in E} x_i < \infty. \quad (8.16)$$

Démonstration.

A. Faisons deux observations préliminaires. Premièrement, quand $1 \leq n \leq T_0$, $X_n = 0$ si et seulement si $n = T_0$, et donc

$$x_0 = 1.$$

Deuxièmement,

$$\sum_{i \in E} \sum_{n=1}^{T_0} 1_{\{X_n=i\}} = \sum_{n=1}^{T_0} \left\{ \sum_{i \in E} 1_{\{X_n=i\}} \right\} = \sum_{n=1}^{T_0} 1 = T_0,$$

et donc

$$\sum_{i \in E} x_i = E_0[T_0]. \quad (8.17)$$

Passons à la démonstration de (8.13). Pour cela, on introduit la quantité

$${}_0p_{0i}(n) := E_0[1_{\{X_n=i\}}1_{\{n \leq T_0\}}] = P_0(X_1 \neq 0, \dots, X_{n-1} \neq 0, X_n = i).$$

C'est la probabilité que, partant de l'état 0, on visite i au temps n sans être au préalable retourné en 0. On a

$$x_i = E_0 \left[\sum_{n \geq 1} 1_{\{X_n=i\}} 1_{\{n \leq T_0\}} \right],$$

et donc,

$$x_i = \sum_{n \geq 1} {}_0p_{0i}(n). \quad (8.18)$$

On observe que

$${}_0p_{0i}(1) = p_{0i}.$$

D'autre part, en utilisant la méthode d'analyse à un pas, pour tout $n \geq 2$,

$${}_0p_{0i}(n) = \sum_{j \neq 0} {}_0p_{0j}(n-1)p_{ji}. \quad (8.19)$$

En faisant la somme de toutes ces égalités, et en tenant compte de (8.18), on obtient

$$x_i = p_{0i} + \sum_{j \neq 0} x_j p_{ji},$$

c'est-à-dire (8.13), puisque $x_0 = 1$.

Ensuite, nous montrons que $x_i > 0$ pour tout $i \in E$. En effet, en itérant (8.13), on a $x^T = x^T \mathbf{P}^n$, c'est-à-dire, puisque $x_0 = 1$,

$$x_i = \sum_{j \in E} x_j p_{ji}(n) = p_{0i}(n) + \sum_{j \neq 0} x_j p_{ji}(n).$$

Si x_i était nul pour un $i \in E$, $i \neq 0$, cela impliquerait que $p_{0i}(n) = 0$ pour tout $n \geq 0$, et donc que 0 et i ne communiquent pas, en contradiction avec l'hypothèse d'irréductibilité.

Il reste à montrer que $x_i < \infty$ pour tout $i \in E$. Comme précédemment, on trouve que

$$1 = x_0 = \sum_{j \in E} x_j p_{j0}(n)$$

pour tout $n \geq 1$, et donc, si $x_i = \infty$ pour un i , nécessairement $p_{i0}(n) = 0$ pour tout $n \geq 1$, et ceci contredit à nouveau l'hypothèse d'irréductibilité.

B. Dans la preuve de la partie A, on a montré que pour toute mesure invariante y d'une matrice stochastique irréductible, $y_i > 0$ pour tout $i \in E$. On peut donc définir, pour tout $i, j \in E$, la matrice \mathbf{Q} par

$$q_{ji} = \frac{y_i}{y_j} p_{ij}. \quad (8.20)$$

C'est une matrice de transition, puisque $\sum_{i \in E} q_{ji} = \frac{1}{y_j} \sum_{i \in E} y_i p_{ij} = \frac{y_j}{y_j} = 1$. Le terme général de \mathbf{Q}^n est

$$q_{ji}(n) = \frac{y_i}{y_j} p_{ij}(n). \quad (8.21)$$

En effet, en supposant (8.21) vraie pour n ,

$$\begin{aligned} q_{ji}(n+1) &= \sum_{k \in E} q_{jk} q_{ki}(n) = \sum_{k \in E} \frac{y_k}{y_j} p_{kj} \frac{y_i}{y_k} p_{ik}(n) \\ &= \frac{y_i}{y_j} \sum_{k \in E} p_{ik}(n) p_{kj} = \frac{y_i}{y_j} p_{ij}(n+1), \end{aligned}$$

et (8.21) suit par induction.

La matrice \mathbf{Q} est irréductible, puisque \mathbf{P} est irréductible. En effet, au vu de (8.21), $q_{ji}(n) > 0$ si et seulement si $p_{ij}(n) > 0$. Aussi, $p_{ii}(n) = q_{ii}(n)$, et donc $\sum_{n \geq 0} q_{ii}(n) = \sum_{n \geq 0} p_{ii}(n)$, et ceci garantit que \mathbf{Q} est récurrente d'après le critère de la matrice potentiel. Notons $g_{ji}(n)$ la probabilité, relative à la chaîne de matrice de transition \mathbf{Q} , de retourner dans l'état i pour la première fois à l'étape n en partant de j . L'analyse à un pas donne :

$$g_{i0}(n+1) = \sum_{j \neq 0} q_{ij} g_{j0}(n). \quad (8.22)$$

En multipliant les deux membres par y_i et en utilisant (8.20), on obtient

$$y_i g_{i0}(n+1) = \sum_{j \neq 0} (y_j g_{j0}(n)) p_{ji}.$$

Rappelons que ${}_0 p_{0i}(n+1) = \sum_{j \neq 0} {}_0 p_{0j}(n) p_{ji}$, ou encore :

$$y_0 {}_0 p_{0i}(n+1) = \sum_{j \neq 0} (y_0 {}_0 p_{0j}(n)) p_{ji}.$$

On voit donc que les suites $\{y_0 {}_0 p_{0i}(n)\}$ et $\{y_i g_{i0}(n)\}$ vérifient la même équation de récurrence. Leurs premiers termes ($n = 1$), respectivement $y_0 {}_0 p_{0i}(1) = y_0 p_{0i}$ et $y_i g_{i0}(1) = y_i q_{i0}$, sont égaux, d'après (8.20). Donc, pour tout $n \geq 1$,

$${}_0 p_{0i}(n) = \frac{y_i}{y_0} g_{i0}(n).$$

En sommant par rapport à $n \geq 1$ et en utilisant l'égalité $\sum_{n \geq 1} g_{i0}(n) = 1$ (\mathbf{Q} est récurrente), on obtient le résultat annoncé : $x_i = \frac{y_i}{y_0}$.

C. La preuve découle immédiatement de l'égalité (8.17) et de la définition de la récurrence positive. \square

Voici enfin le critère de récurrence positive de la probabilité stationnaire.

Théorème 8.2.5 1. Une CMH irréductible est récurrente positive si et seulement si elle admet une probabilité stationnaire π . Dans ce cas, la probabilité stationnaire est unique, et $\pi(i) > 0$ pour tout $i \in E$.

2. Soit π l'unique distribution stationnaire d'une chaîne irréductible récurrente positive, et soit T_i le temps de retour en i . Alors

$$\pi(i)E_i[T_i] = 1. \quad (8.23)$$

3. Toute CMH irréductible d'espace d'état fini est récurrente positive.

Démonstration.

1. La partie directe est une conséquence immédiate du Théorème 8.2.4. Pour la réciproque, supposons l'existence d'une distribution stationnaire π . En itérant $\pi^T = \pi^T \mathbf{P}$, on obtient $\pi^T = \pi^T \mathbf{P}^n$, c'est-à-dire, pour tout $i \in E$,

$$\pi(i) = \sum_{j \in E} \pi(j)p_{ji}(n).$$

Si la chaîne était transiente, on aurait, pour tous états i, j ,

$$\lim_{n \uparrow \infty} p_{ji}(n) = 0.$$

(En effet, $\lim_{n \uparrow \infty} p_{ji}(n) = \lim_{n \uparrow \infty} E_j[1_{\{X_n=i\}}]$. D'autre part, $\lim_{n \uparrow \infty} 1_{\{X_n=i\}} = 0$ (j est transient) et $1_{\{X_n=i\}} \leq 1$, et donc, par convergence dominée, $\lim_{n \uparrow \infty} E_j[1_{\{X_n=i\}}] = 0$.) Puisque $p_{ji}(n)$ est uniformément (en j et n) bornée par 1, on a, par convergence dominée à nouveau,

$$\pi(i) = \lim_{n \uparrow \infty} \sum_{j \in E} \pi(j)p_{ji}(n) = \sum_{j \in E} \pi(j) \left(\lim_{n \uparrow \infty} p_{ji}(n) \right) = 0.$$

Ceci contredit $\sum_{i \in E} \pi(i) = 1$. La chaîne ne peut donc être que récurrente, et d'après le Théorème 8.2.4, partie C, elle est récurrente positive.

La distribution stationnaire π d'une chaîne irréductible récurrente positive est unique (d'après le Théorème 8.2.4, partie B, et le fait que le seul choix du facteur multiplicatif est 1). Aussi, on rappelle que $\pi(i) > 0$ pour tout $i \in E$ (Théorème 8.2.4, partie A).

2. Cette égalité est une conséquence directe de l'expression (8.14) de la mesure invariante. En effet, π est obtenue par normalisation de x : pour tout $i \in E$,

$$\pi(i) = \frac{x_i}{\sum_{j \in E} x_j},$$

et en particulier, pour $i = 0$, en se rappelant que $x_0 = 1$ et en utilisant (8.17),

$$\pi(0) = \frac{x_0}{\sum_{j \in E} x_j} = \frac{1}{E_0[T_0]}.$$

L'état 0 ne jouant pas un rôle spécial dans cette analyse, (8.23) est vraie pour tout $i \in E$.

3. On prouve d'abord la récurrence. Si la chaîne était transiente, alors, pour tout $i, j \in E$,

$$\lim_{n \uparrow \infty} p_{ij}(n) = 0,$$

et donc, puisque l'espace d'état est fini,

$$\lim_{n \uparrow \infty} \sum_{j \in E} p_{ij}(n) = 0.$$

Mais pour tout n ,

$$\sum_{j \in E} p_{ij}(n) = 1,$$

une contradiction. Donc, la chaîne est récurrente. D'après le Théorème 8.2.4 elle admet une mesure invariante x . Puisque E est fini, $\sum_{i \in E} x_i < \infty$, et donc la chaîne est récurrente positive, d'après le Théorème 8.2.4, partie C. \square

EXEMPLE 8.2.2: PROCESSUS DE NAISSANCE ET DE MORT, II. Le modèle est le même que celui de l'Exemple 8.1.9, sauf que l'espace d'état est $E = \mathbb{N}$. Les mêmes calculs que précédemment conduisent à l'expression (8.8). Pour que cette solution soit une probabilité, on doit avoir $\pi(0) > 0$. Aussi en écrivant que $\sum_{i=1}^{\infty} \pi(i) = 1$, on obtient

$$\pi(0) \left\{ 1 + \frac{1}{q_1} + \sum_{j=1}^{\infty} \frac{p_1 p_2 \cdots p_j}{q_1 q_2 \cdots q_{j+1}} \right\} = 1. \quad (8.24)$$

Donc une distribution stationnaire existe si et seulement si

$$\sum_{j=1}^{\infty} \frac{p_1 p_2 \cdots p_j}{q_1 q_2 \cdots q_{j+1}} < \infty. \quad (8.25)$$

Dans ce cas $\pi(i)$ est donné par (8.8), où $\pi(0)$ est déterminé par (8.24).

Un cas particulier important est

$$E = \mathbb{N}, \quad p_i = p, \quad q_i = 1 - p, \quad r_i = 0$$

où $0 < p < 1$. Il s'agit d'une marche aléatoire du type de l'Exemple 8.1.1 où l'état 0 est "réfléchissant". La condition (8.25) se lit

$$\sum_j \left(\frac{p}{1-p} \right)^j < \infty,$$

c'est-à-dire : la chaîne est récurrente positive si et seulement si $p < \frac{1}{2}$.

EXEMPLE 8.2.3: L'ATELIER DE RÉPARATION. Durant le jour n , Z_{n+1} machines tombent en panne, et elles sont admises dans l'atelier de réparation dans la journée $n+1$. Chaque jour, une machine parmi celles qui attendent est réparée. Donc, en notant X_n le nombre de machines dans l'atelier au jour n ,

$$X_{n+1} = (X_n - 1)^+ + Z_{n+1}, \quad (8.26)$$

où $a^+ = \max(a, 0)$. En particulier, si $\{Z_n\}_{n \geq 1}$ est une suite IID indépendante de l'état initial X_0 , alors $\{X_n\}_{n \geq 0}$ est une CMH. En termes de la distribution de probabilité

$$P(Z_1 = k) = a_k, \quad k \geq 0,$$

la matrice de transition de cette chaîne est

$$\mathbf{P} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

En effet, la formule (8.6) donne,

$$p_{ij} = P((i-1)^+ + Z_1 = j) = P(Z_1 = j - (i-1)^+) = a_{j-(i-1)^+}.$$

Nous allons montrer qu'une condition nécessaire et suffisante d'irréductibilité est que $P(Z_1 = 0) > 0$ ($a_0 > 0$) et $P(Z_1 \geq 2) > 0$ ($a_0 + a_1 < 1$).

L'équation de récurrence (8.26) nous permet de faire les observations suivantes. Si $a_0 = 0$, alors $X_{n+1} \geq X_n$ et il y a donc une probabilité nulle d'aller de i à $i-1$ quel que soit $i \geq 1$. Si $a_0 + a_1 = 1$, alors $X_{n+1} \leq X_n$ et il y a donc une probabilité nulle d'aller de i à $i+1$ quel que soit $i \geq 0$. Donc les deux conditions $a_0 > 0$ et $a_0 + a_1 < 1$ sont *nécessaires* pour l'irréductibilité.

Elles sont également *suffisantes*. En effet, s'il existe $k \geq 2$ tel que $P(Z_{n+1} = k) > 0$, alors on peut aller de n'importe quel $i > 0$ à $i+k-1 > i$ ou de $i = 0$ à $k > 0$ avec une probabilité positive. De plus, si $P(Z_{n+1} = 0) > 0$, on peut aller de $i > 0$ à $i-1$ avec une probabilité positive. En particulier, on peut aller de i à $j < i$ avec une probabilité positive. Donc, pour aller de i à $j \geq i$, on peut procéder par sauts vers le haut de hauteur strictement positive, pour atteindre un état $l \geq i$, et ensuite, dans le cas où $l > i$, descendre par sauts d'une unité vers le bas de l à i . Tout ceci avec une probabilité positive.

Supposons que la chaîne soit irréductible récurrente positive, avec la distribution stationnaire π . Soit z un nombre complexe de module ≤ 1 . De (8.26), on tire

$$\begin{aligned} z^{X_{n+1}+1} &= \left(z^{(X_n-1)^++1} \right) z^{Z_{n+1}} \\ &= \left(z^{X_n} 1_{\{X_n > 0\}} + z 1_{\{X_n = 0\}} \right) z^{Z_{n+1}} \\ &= \left(z^{X_n} - 1_{\{X_n = 0\}} + z 1_{\{X_n = 0\}} \right) z^{Z_{n+1}}, \end{aligned}$$

et donc

$$zz^{X_{n+1}} - z^{X_n} z^{Z_{n+1}} = (z - 1)1_{\{X_n=0\}} z^{Z_{n+1}}.$$

Comme X_n et Z_{n+1} sont indépendantes, $E[z^{X_n} z^{Z_{n+1}}] = E[z^{X_n}]g_Z(z)$, où $g_Z(z)$ est la fonction génératrice de Z_{n+1} , et $E[1_{\{X_n=0\}} z^{Z_{n+1}}] = \pi(0)g_Z(z)$, où $\pi(0) = P(X_n = 0)$. Donc,

$$zE[z^{X_{n+1}}] - g_Z(z)E[z^{X_n}] = (z - 1)\pi(0)g_Z(z).$$

Si on est en régime stationnaire, alors $E[z^{X_{n+1}}] = E[z^{X_n}] = g_X(z)$, et donc :

$$g_X(z)(z - g_Z(z)) = \pi(0)(z - 1)g_Z(z). \quad (8.27)$$

Ceci nous donne $g_X(z) = \sum_{i=0}^{\infty} \pi(i)z^i$, du moins si on dispose de $\pi(0)$. Pour obtenir $\pi(0)$, on dérive (8.27) :

$$g'_X(z)(z - g_Z(z)) + g_X(z)(1 - g'_Z(z)) = \pi(0)(g_Z(z) + (z - 1)g'_Z(z)), \quad (8.28)$$

et on fait $z = 1$, pour obtenir, en tenant compte des identités $g_X(1) = g_Z(1) = 1$ et $g'_Z(1) = E[Z]$,

$$\pi(0) = 1 - E[Z].$$

Comme $\pi(0)$ doit être non négative, ceci conduit à la condition nécessaire de récurrence positive : $E[Z] \leq 1$. En fait, on a nécessairement $E[Z] < 1$. En effet, si $E[Z] = 1$, ce qui implique $\pi(0) = 0$, il découle de (8.27) que

$$g_X(x)(x - g_Z(x)) = 0$$

pour tout $x \in [0, 1]$. Mais comme la chaîne est supposée irréductible, le cas $Z_{n+1} \equiv 1$ (c'est-à-dire, $g_Z(x) \equiv x$) est exclus et l'équation $x - g_Z(x) = 0$ a seulement $x = 1$ comme solution quand $g'_Z(1) = E[Z] \leq 1$. Donc $g_X(x) \equiv 0$ pour tout $x \in [0, 1]$, et en conséquence $g_X(z) \equiv 0$ sur $\{|z| < 1\}$ (une fonction analytique sur un ouvert ne peut avoir de points d'accumulation de zéros à l'intérieur de cet ouvert, sauf si elle est identiquement nulle). Ceci mène à une contradiction puisque la fonction génératrice d'une variable à valeurs entières ne peut être identiquement nulle.

On a donc démontré qu'une condition nécessaire de récurrence positive est $E[Z] < 1$. Nous allons voir que c'est aussi une condition suffisante. On commence par vérifier que la condition $E[Z] < 1$ entraîne la récurrence. Il suffit de montrer que la probabilité de l'événement $A =$ "non retour en 0 en partant de 0" est nulle. En effet dans A , pour tout n ,

$$X_n = 1 + \sum_{k=1}^n (Z_k - 1) \geq 1$$

et en particulier

$$\frac{\sum_{k=1}^n Z_k}{n} - 1 > 0,$$

ce qui donne, en faisant tendre n vers l'infini, d'après la loi forte des grands nombres, $E[Z] \geq 1$. Une contradiction.

Supposons la récurrence, c'est-à-dire, supposons que le temps de retour T_0 est presque sûrement fini. On a l'égalité :

$$\sum_{n=1}^{T_0} Z_n = (T_0 - 1). \quad (8.29)$$

Observons que

$$\begin{aligned} E[Z_n 1_{n \leq T_0}] &= E[Z_n] - E[Z_n 1_{n > T_0}] \\ &= E[Z_n] - E[Z_n] E[1_{n > T_0}] \\ &= E[Z_n] E[1_{n \leq T_0}], \end{aligned}$$

où on a utilisé le fait que l'événement $\{n > T_0\}$, qui ne dépend que de Z_1, \dots, Z_{n-1} , est indépendant de Z_n . On a donc, à partir de (8.29)

$$\begin{aligned} E_0[T_0] &= 1 + E_0 \left[\sum_{n=1}^{\infty} Z_n 1_{n \leq T_0} \right] \\ &= 1 + \sum_{n=1}^{\infty} E_0[Z_n 1_{n > T_0}] \\ &= 1 + \sum_{n=1}^{\infty} E_0[Z_n] E[1_{n \leq T_0}] \\ &= 1 + E[Z_1] \sum_{n=1}^{\infty} E_0[1_{n \leq T_0}] \\ &= 1 + E[Z_1] E_0 \left[\sum_{n=1}^{\infty} 1_{n \leq T_0} \right] \\ &= 1 + E[Z_1] E_0[T_0], \end{aligned}$$

et donc, si $E[Z] < 1$ (en particulier, la chaîne est récurrente),

$$E_0[T_0] = (1 - E[Z_1])^{-1} < \infty,$$

et donc la chaîne est récurrente positive.

Si $E[Z] = 1$ (en particulier, la chaîne est récurrente), on ne peut avoir $E_0[T_0] < \infty$, car alors on aurait $E_0[T_0] = 1 + E_0[T_0]$. La chaîne récurrente est donc dans ce cas récurrente nulle.

Reste à examiner le cas $E[Z] > 1$. On sait que cela exclue la récurrence positive. Il se trouve qu'en fait, la chaîne est alors transiente, mais nous ne le démontrerons pas ici.

En résumé, sous la condition d'irréductibilité ($P(Z = 0) > 0$ et $P(Z \geq 2) > 0$) :

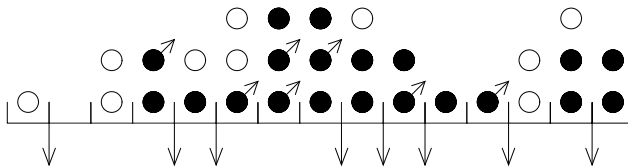
A. Si $E[Z] < 1$, la chaîne est récurrente positive et la fonction génératrice de sa distribution stationnaire est

$$\sum_{i=0}^{\infty} \pi(i) z^i = (1 - E[Z]) \frac{(z - 1) g_Z(z)}{z - g_Z(z)}.$$

- B. Si $E[Z] = 1$, la chaîne est récurrente nulle.
 C. Si $E[Z] > 1$, la chaîne est transiente.

EXEMPLE 8.2.4: INSTABILITÉ DU PROTOCOLE ALOHA. Le protocole ALOHA est une méthode d'accès à un canal satellite. C'est un *protocole distribué*, en ce sens que les utilisateurs du canal ne se concertent pas pour éviter les *collisions* (plus d'un message transmis en même temps, auquel cas aucun de ces messages n'est considéré comme transmis, et chacun des messages en collision redemande la transmission à un temps ultérieur, de manière suffisamment intelligente — évidemment pas de retransmission immédiate). Plus précisément, on considère le protocole ALOHA avec fenêtres de transmission périodiques. Ce protocole impose les règles suivantes (voir aussi la figure ci-dessous) :

- (i) Les transmissions et retransmissions des messages ont lieu dans des intervalles de temps équidistants, les *slots*, dont la durée est supérieure à celle du temps nécessaire pour transmettre un message. (Ici un message est une suite de longueur fixe de symboles binaires).
 (ii) Les messages qui au début d'un *slot* sont en attente de retransmission demandent leur retransmission indépendamment les uns des autres chacun avec la probabilité $\nu \in (0, 1)$.
 (iii) Les *messages frais*—ceux qui se présentent pour la première fois— tentent immédiatement de passer.



- message frais
 ● message en attente, non autorisé à retransmettre
 ● ↗ message en attente, autorisé à retransmettre
 ↓ transmission réussie

Le protocole ALOHA

Soit X_n le nombre de *messages en attente de retransmission* au début du *slot* n . La probabilité pour que i parmi $X_n = k$ messages en attente de retransmission demandent la retransmission dans le *slot* suivant est donc

$$b_i(k) = \binom{k}{i} \nu^i (1 - \nu)^{k-i}.$$

Soit A_n le nombre de requêtes nouvelles dans le n -ème *slot*. La suite $\{A_n\}_{n \geq 0}$ est supposée IID avec la distribution

$$P(A_n = j) = a_j.$$

La quantité

$$\lambda = E[A_n] = \sum_{i=1}^{\infty} i a_i$$

est *l'intensité de trafic*. On suppose que $0 < a_0 + a_1 < 1$, ce qui garantit que $\{X_n\}_{n \geq 0}$ est une CMH irréductible. Sa matrice de transition est :

$$\begin{aligned} p_{ij} &= b_1(i) a_0 \text{ si } j = i - 1, \\ &= [1 - b_1(i)] a_0 + b_0(i) a_1 \text{ si } j = i, \\ &= [1 - b_0(i)] a_1 \text{ si } j = i + 1, \\ &= a_{j-i} \text{ si } j \geq i + 2. \end{aligned}$$

La preuve consiste à comptabiliser les possibilités. Par exemple, la première ligne correspond à un parmi les i messages en attente (de transmission ou de retransmission) qui réussit à passer, et pour cela il faut qu'il y ait soit pas de message nouveau (probabilité a_0) et un seul parmi les i messages en attente qui est admis à retransmettre (probabilité $b_1(i)$). La seconde ligne correspond à un des deux événements suivants : (1), "pas de nouveau message et zéro ou plus de deux requêtes de retransmission parmi les messages en attente" et (2), "un nouveau message et zéro requête de retransmission parmi les messages en attente".

Le but de cet exemple est de démontrer que ce protocole n'est pas stable, en ce sens que la CMH $\{X_n\}_{n \geq 0}$ n'est *pas récurrente positive*. Pour cela, il suffit, d'après le Théoreme 8.3.1 de contredire l'existence d'une distribution stationnaire π .

Si une telle distribution stationnaire existait, elle satisferait aux équations de balance globale

$$\begin{aligned} \pi(i) &= \pi(i) \{ [1 - b_1(i)] a_0 + b_0(i) a_1 \} + \pi(i - 1) [1 - b_0(i - 1)] a_1 \\ &\quad + \pi(i + 1) b_1(i + 1) a_0 + \sum_{\ell=2}^{\infty} \pi(i - \ell) a_{\ell} \end{aligned}$$

(où $\pi(i) = 0$ si $i < 0$). Posons

$$P_N = \sum_{i=0}^N \pi(i)$$

et faisons la somme des équations de balance globale de $i = 0$ à N . On obtient :

$$P_N = \pi(N) b_0(N) a_1 + \pi(N + 1) b_1(N + 1) a_0 + \sum_{\ell=0}^N a_{\ell} P_{N-\ell},$$

qui donne à son tour :

$$P_N(1 - a_0) = \pi(N)b_0(N)a_1 + \pi(N+1)b_1(N+1)a_0 + \sum_{\ell=1}^N a_\ell P_{N-\ell}.$$

Mais comme P_N croît avec N et $\sum_{\ell=1}^N a_\ell \leq \sum_{\ell=1}^{\infty} a_\ell = 1 - a_0$, nous avons

$$\sum_{\ell=1}^N a_\ell P_{N-\ell} \leq P_{N-1}(1 - a_0),$$

et donc

$$P_N(1 - a_0) \leq \pi(N)b_0(N)a_1 + \pi(N+1)b_1(N+1)a_0 + P_{N-1}(1 - a_0),$$

d'où il suit que

$$\frac{\pi(N+1)}{\pi(N)} \geq \frac{1 - a_0 - b_0(N)a_1}{b_1(N+1)a_0}.$$

Faisant usage de la forme explicite des $b_i(k)$, on obtient

$$\frac{\pi(N+1)}{\pi(N)} \geq \frac{(1 - a_0) - (1 - \nu)^N a_1}{(N+1)\nu(1 - \nu)^N a_0}.$$

Pour toutes les valeurs de $\nu \in (0, 1)$, le membre de droite de cette inégalité tend vers l'infini, ce qui contredit $\sum_{N=1}^{\infty} \pi(N) = 1$ et les inégalités $\pi(N) > 0$ que π doit vérifier en tant que distribution stationnaire d'une CMH irréductible.

8.3 Comportement asymptotique

Convergence vers l'équilibre

Considérons une CMH irréductible et récurrente positive. En particulier, si la distribution initiale est la distribution stationnaire, elle conserve cette distribution pour tous les temps. La chaîne est alors dite en *régime stationnaire*.

Quel est son comportement à long terme quand la distribution initiale est arbitraire ? Le résultat fondamental est le suivant :

Théorème 8.3.1 *Soit une CMH de matrice de transition \mathbf{P} irréductible, récurrente positive et apériodique. Alors, pour tout $j \in E$, et toute distribution initiale μ ,*

$$\lim_{n \uparrow \infty} P(X_n = j) = \pi(j),$$

où π est la distribution stationnaire de la chaîne.

Définition 8.3.1 *Une CMH irréductible récurrente positive et APÉRIODIQUE est dite ergodique.*

Supposons que l'on trouve un processus $\{X'_n\}_{n \geq 0}$ tel que $X_n \stackrel{\mathcal{D}}{\sim} X'_n$ pour tout $n \geq 0$, et un processus $\{X''_n\}_{n \geq 0}$ tel que $X''_n \stackrel{\mathcal{D}}{\sim} \pi$ pour tout $n \geq 0$. Alors, le résultat sera démontré si l'on peut prouver que

$$\lim_{n \uparrow \infty} |P(X'_n = i) - P(X''_n = i)| = 0. \quad (8.30)$$

Nous allons construire $\{X'_n\}_{n \geq 0}$ et $\{X''_n\}_{n \geq 0}$ pour que cette situation ait bien lieu. Nous allons le faire de telle sorte qu'il existe un temps aléatoire presque sûrement fini τ tel que $X'_n = X''_n$ pour tout $n \geq \tau$. On montrera ci-dessous que, dans ce cas,

$$|P(X'_n = i) - P(X''_n = i)| \leq P(\tau > n). \quad (8.31)$$

Alors, la finitude de τ entraîne que $\lim_{n \uparrow \infty} P(\tau > n) = 0$, et le résultat sera donc prouvé. Voici la démonstration de (8.31) :

$$\begin{aligned} P(X'_n = j) - P(X''_n = j) &= P(X'_n = j, \tau \leq n) + P(X'_n = j, \tau > n) \\ &\quad - P(X''_n = j, \tau \leq n) - P(X''_n = j, \tau > n) \\ &= P(X'_n = j, \tau > n) - P(X''_n = j, \tau > n) \\ &\leq P(X'_n = j, \tau > n) \\ &\leq P(\tau > n). \end{aligned}$$

De même, $P(X''_n = j) - P(X'_n = j) \leq P(\tau > n)$.

Il nous reste à construire $\{X'_n\}_{n \geq 0}$ et $\{X''_n\}_{n \geq 0}$ avec les propriétés annoncées.

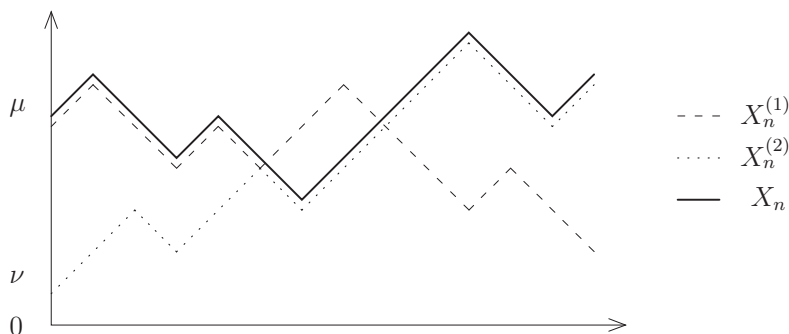
Théorème 8.3.2 Soit $\{X_n^{(1)}\}_{n \geq 0}$ et $\{X_n^{(2)}\}_{n \geq 0}$ deux CMH ergodiques indépendantes de même matrice de transition \mathbf{P} et de distributions initiales μ et ν , respectivement. Soit $\tau = \inf\{n \geq 0; X_n^{(1)} = X_n^{(2)}\}$, avec $\tau = \infty$ si les deux chaînes ne se recoupent jamais. Alors τ est en fait presque sûrement fini et, de plus, le processus $\{X'_n\}_{n \geq 0}$ défini par

$$X'_n = \begin{cases} X_n^{(1)} & \text{if } n \leq \tau, \\ X_n^{(2)} & \text{if } n \geq \tau \end{cases} \quad (8.32)$$

(voir la figure ci-dessous) est une CMH de matrice de transition \mathbf{P} .

Démonstration. Considérons la CMH $\{Z_n\}_{n \geq 0}$ définie par $Z_n = (X_n^{(1)}, X_n^{(2)})$ (chaîne produit), prenant ses valeurs dans $E \times E$. Sa probabilité de transition de (i, k) à (j, ℓ) en n étapes est $p_{ij}(n)p_{k\ell}(n)$, elle est irréductible, et elle admet $\{\pi(i)\pi(j)\}_{(i,j) \in E^2}$ comme distribution stationnaire (voir l'Exercice 8.5.16, où le rôle de l'hypothèse d'apériodicité est mis en évidence). D'après le critère de la distribution stationnaire, la chaîne produit est récurrente positive. En particulier, elle atteint la diagonale de E^2 en temps fini, et donc $P(\tau < \infty) = 1$.

Il reste à prouver que le processus $\{X'_n\}_{n \geq 0}$ défini par (8.32) est une CMH de matrice de transition \mathbf{P} . Ceci est fait dans l'Exercice 8.5.18. \square



EXEMPLE 8.3.1: L'URNE D'EHRENFEST, TAKE 4. (suite des Exemples 8.1.3, 8.1.8 et 8.1.10) La célébrité du modèle d'Ehrenfest est due à l'éclaircissement qu'il apporte au phénomène d'irréversibilité thermodynamique, un sujet de controverses du temps de Boltzmann. Selon la théorie macroscopique de la thermodynamique de ce physicien, les systèmes progressent d'une manière ordonnée vers l'équilibre thermodynamique.

Considérons, par exemple, un système de N particules gazeuses dans une boîte divisée en deux compartiments, A et B , séparés par une membrane fictive. Si à l'origine des temps on place toutes les particules dans le compartiment A , ces particules vont se redistribuer, et le système atteint l'équilibre, un état pour lequel les contenus deux compartiments sont thermodynamiquement équivalents. Boltzmann disait qu'il existait une flèche du temps en direction de l'entropie croissante, et en effet dans l'expérience de diffusion modélisée par le modèle des Ehrenfest, l'équilibre thermodynamique correspond à l'entropie maximale du système.

Boltzmann eut un redoutable contradicteur, un certain Zermelo, qui, au vu de la réversibilité dans le temps des lois de la physique, doutait de la flèche du temps évoquée par Boltzmann, ou du moins, demandait une explication. Sa position était renforcée par d'irréfutables mathématiques, en particulier le théorème de récurrence de Poincaré qui prédisait que si l'expérience débutait avec toutes les particules dans le compartiment A , on les retrouverait toutes, tôt ou tard dans le compartiment d'origine. Comme chacun sait, ce comportement n'est jamais observé dans la vie quotidienne, et on n'a jamais vu le sucre dissous dans la tasse de café reprendre sa forme initiale.

La théorie de Boltzmann était mise à mal par ce type d'arguments. Les choses devaient être clarifiées, et elles le furent par Tatiana et Paul Ehrenfest, dont le modèle markovien permit de sauver tout l'édifice.

Ce modèle ne rentre pas dans les détails de la physique du phénomène de diffusion, mais il en conserve les traits essentiels du point de vue de la physique statistique. C'est un système réversible (la chaîne de Markov est réversible) et il est récurrent, repassant une infinité de fois dans n'importe quel état, comme par exemple celui où le compartiment A est vide. L'irréversibilité du système consiste en ceci : en partant d'un état quelconque,

la distribution au temps n converge vers la distribution stationnaire¹ qui met pratiquement toute la masse sur les états proches du partage équilibré des particules entre les deux compartiments. Il n'en reste pas moins que ceci n'empêche pas la récurrence et en particulier le retour en l'état 0 (correspondant au compartiment A vide).

Mais en réalité ce retour n'est *pas observable* dans le sens suivant. On peut montrer que le temps moyen pour aller à l'état 0 en partant de $L = \frac{N}{2}$ (on suppose N pair) est

$$\frac{1}{2L} 2^{2L} (1 + O(L))$$

tandis que le temps moyen pour aller de L à 0 est inférieur à

$$L + L \log L + O(1).$$

Avec $L = 10^6$ et une unité de temps mathématique égale à 10^{-5} seconde, le retour à l'état L en partant de 0 est de l'ordre d'une seconde, tandis qu'il faudrait de l'ordre de

$$\frac{1}{2 \cdot 10^{11}} \times 2^{2^{10^6}} \text{ secondes}$$

pour passer de L à 0. Il s'agit d'un temps astronomique, et c'est en ce sens que le retour en 0 n'est pas observable.

Ces nombres nous apprennent qu'il est inutile de passer trop de temps à touiller son café, ou de l'avaler rapidement de peur que le morceau de sucre se reforme. Plus sérieusement : rien n'empêche la chaîne de se trouver dans un état rare (au sens probabiliste, c'est-à-dire, de faible probabilité stationnaire), seulement elle ne s'y trouve que rarement (au sens temporel), extrêmement rarement, pour nous autres mortels, jamais !

Boltzmann était conscient du fait que les temps de récurrence dans le théorème de Poincaré devaient être très longs, mais ses arguments n'avaient pas réussi à convaincre, ce que put faire le modèle Ehrenfest.

Théorème ergodique

Nous allons donner des conditions générales garantissant que les moyennes empiriques du type

$$\frac{1}{N} \sum_{k=1}^N g(X_k, \dots, X_{k+L})$$

convergent vers les moyennes probabilistes.

On obtiendra le résultat recherché comme conséquence de la proposition suivante.

¹Dans ce modèle, ce n'est pas vrai à cause de la périodicité de la chaîne, mais cette objection disparaît au prix d'une légère modification.

Proposition 8.3.1 *Soit $\{X_n\}_{n \geq 0}$ une CMH irréductible récurrente, et soit x la mesure invariante canonique associée à l'état $0 \in E$,*

$$x_i = E_0 \left[\sum_{n \geq 1} 1_{\{X_n=i\}} 1_{\{n \leq T_0\}} \right], \quad (8.33)$$

où T_0 est le temps de retour en 0. Définissons pour tout $n \geq 1$

$$\nu(n) = \sum_{k=1}^n 1_{\{X_k=0\}}. \quad (8.34)$$

Soit maintenant $f : E \rightarrow \mathbb{R}$ une fonction telle que

$$\sum_{i \in E} |f(i)| x_i < \infty. \quad (8.35)$$

Alors, pour toute distribution initiale μ , presque sûrement,

$$\lim_{N \uparrow \infty} \frac{1}{\nu(N)} \sum_{k=1}^N f(X_k) = \sum_{i \in E} f(i) x_i. \quad (8.36)$$

Démonstration. Soit $T_0 = \tau_1, \tau_2, \tau_3, \dots$ les temps de retour successifs en 0. Posons

$$U_p = \sum_{n=\tau_p+1}^{\tau_{p+1}} f(X_n).$$

La suite $\{U_p\}_{p \geq 1}$ est, d'après le Théorème 8.2.1, IID. De plus, si on suppose $f \geq 0$,

$$\begin{aligned} E[U_1] &= E_0 \left[\sum_{n=1}^{T_0} f(X_n) \right] \\ &= E_0 \left[\sum_{n=1}^{T_0} \sum_{i \in E} f(i) 1_{\{X_n=i\}} \right] = \sum_{i \in E} f(i) E_0 \left[\sum_{n=1}^{T_0} 1_{\{X_n=i\}} \right] \\ &= \sum_{i \in E} f(i) x_i. \end{aligned}$$

Cette quantité est finie par hypothèse et, d'après la loi forte des grands nombres,

$$\lim_{n \uparrow \infty} \frac{1}{n} \sum_{p=1}^n U_p = \sum_{i \in E} f(i) x_i,$$

c'est-à-dire :

$$\lim_{n \uparrow \infty} \frac{1}{n} \sum_{k=T_0+1}^{\tau_{n+1}} f(X_k) = \sum_{i \in E} f(i) x_i. \quad (8.37)$$

En observant que

$$\tau_{\nu(n)} \leq n < \tau_{\nu(n)+1},$$

on a :

$$\frac{\sum_{k=1}^{\tau_{\nu(n)}} f(X_k)}{\nu(n)} \leq \frac{\sum_{k=1}^n f(X_k)}{\nu(n)} \leq \frac{\sum_{k=1}^{\tau_{\nu(n)+1}} f(X_k)}{\nu(n)}.$$

Comme la chaîne est récurrente, $\lim_{n \uparrow \infty} \nu(n) = \infty$, et donc, d'après (8.37), les termes extrêmes de la chaîne d'inégalités ci-dessus tendent vers $\sum_{i \in E} f(i)x_i$ quand n tend vers l'infini, et ceci donne (8.36). Le cas où f est de signe arbitraire s'obtient en écrivant (8.36) pour $f^+ = \max(0, f)$ et $f^- = \max(0, -f)$, et en prenant les différences des égalités obtenues de cette manière (la différence n'est pas une forme indéterminée $\infty - \infty$, grâce à l'hypothèse (8.35)). \square

Nous sommes maintenant en mesure de donner le *théorème ergodique* pour les chaînes de Markov.

Théorème 8.3.3 *Soit $\{X_n\}_{n \geq 0}$ une CMH irréductible récurrente positive de distribution stationnaire π , et soit $f : E \rightarrow \mathbb{R}$ une fonction telle que*

$$E_\pi [|f(X_0)|] := \sum_{i \in E} |f(i)|\pi(i) < \infty. \quad (8.38)$$

Alors, pour toute distribution initiale μ , presque sûrement,

$$\lim_{n \uparrow \infty} \frac{1}{N} \sum_{k=1}^N f(X_k) = \sum_{i \in E} f(i)\pi(i). \quad (8.39)$$

Démonstration. On applique la Proposition 8.3.1 à $f \equiv 1$. La condition (8.35) est satisfaite, puisque dans le cas positif récurrent, $\sum_{i \in E} x_i = E_0[T_0] < \infty$. Donc, presque sûrement,

$$\lim_{N \uparrow \infty} \frac{N}{\nu(N)} = \sum_{j \in E} x_j.$$

Si la fonction f satisfait (8.38), elle satisfait aussi (8.35), puisque x et π sont proportionnelles, et donc, presque sûrement,

$$\lim_{N \uparrow \infty} \frac{1}{\nu(N)} \sum_{k=1}^N f(X_k) = \sum_{i \in E} f(i)x_i.$$

En combinant les égalités ci-dessus, on obtient

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N f(X_k) = \lim_{N \rightarrow \infty} \frac{\nu(N)}{N} \frac{1}{\nu(N)} \sum_{k=1}^N f(X_k) = \frac{\sum_{i \in E} f(i)x_i}{\sum_{j \in E} x_j},$$

d'où (8.39), puisque π est obtenue par normalisation de x . \square

Le corollaire suivant étend le domaine d'application du Théorème 8.3.3.

Corollaire 8.3.1 Soit $\{X_n\}_{n \geq 0}$ une CMH irréductible récurrente positive de distribution stationnaire π , et soit $g : E^{L+1} \rightarrow \mathbb{R}$ une fonction telle que

$$E_\pi[|g(X_0, \dots, X_L)|] < \infty.$$

Alors, pour toute distribution initiale μ , presque sûrement,

$$\lim_{N \uparrow \infty} \frac{1}{N} \sum_{k=1}^N g(X_k, X_{k+1}, \dots, X_{k+L}) = E_\pi[|g(X_0, \dots, X_L)|].$$

Démonstration. On applique le Théorème 8.3.3 à la chaîne $\{(X_n, X_{n+1}, \dots, X_{n+L})\}_{n \geq 0}$ d'espace d'état

$$F = \{(i_0, i_1, \dots, i_L) \in E^{L+1}; p_{i_0 i_1} \cdots p_{i_{L-1} i_L} > 0\}$$

qui est (voir Exercice 8.5.19) irréductible récurrente positive de distribution stationnaire

$$\pi(i_0) p_{i_0 i_1} \cdots p_{i_{L-1} i_L}.$$

□

EXEMPLE 8.3.2: ESTIMATEUR EMPIRIQUE DES PROBABILITÉS DE TRANSITION. Soit $\{X_n\}_{n \geq 1}$ une CMH irréductible récurrente positive de matrice de transition \mathbf{P} et de distribution stationnaire π . Alors, pour tous $i_0, i_1 \in E$,

$$\lim_{N \uparrow \infty} \frac{1}{N} \sum_{k=1}^N 1_{\{X_n = i_0\}} = \pi(i_0)$$

et

$$\lim_{N \uparrow \infty} \frac{1}{N} \sum_{k=1}^N 1_{\{X_n = i_0, X_{n+1} = i_1\}} = \pi(i_0) p_{i_0, i_1}.$$

En particulier,

$$\lim_{N \uparrow \infty} \frac{\sum_{k=1}^N 1_{\{X_n = i_0, X_{n+1} = i_1\}}}{\sum_{k=1}^N 1_{\{X_n = i_0\}}} = p_{i_0, i_1}.$$

Démonstration. Soit $f : E \rightarrow \mathbb{R}$ la fonction définie par $f(i) = 1_{\{i_0\}}(i)$. On a :

$$\begin{aligned} f(X_n) &= 1_{\{i_0\}}(X_n) \\ &= 1_{\{X_n = i_0\}}, \end{aligned}$$

et donc, d'après le théorème ergodique (Théorème 8.3.3),

$$\begin{aligned} \lim_{N \uparrow \infty} \frac{1}{N} \sum_{k=1}^N 1_{\{X_n = i_0\}} &= E_\pi [1_{\{X_0 = i_0\}}] \\ &= P_\pi(X_0 = i_0) = \pi(i_0). \end{aligned}$$

Avec $f : E \times E \rightarrow \mathbb{R}$ définie par $f(i, j) = 1_{\{i_0, i_1\}}(i, j)$, on a :

$$\begin{aligned} f(X_n, X_{n+1}) &= 1_{\{i_0, i_1\}}(X_n, X_{n+1}) \\ &= 1_{\{X_n=i_0, X_{n+1}=i_1\}}. \end{aligned}$$

et donc, toujours d'après le théorème ergodique (Corollaire 8.3.1) :

$$\begin{aligned} \lim_{N \uparrow \infty} \frac{1}{N} \sum_{k=1}^N 1_{\{X_n=i_0, X_{n+1}=i_1\}} &= E_\pi [1_{\{X_0=i_0, X_1=i_1\}}] \\ &= P_\pi(X_0 = i_0, X_1 = i_1) = \pi(i_0)p_{i_0, i_1}. \end{aligned}$$

□

On voit donc que si l'on ne connaît pas la matrice de transition d'une CMH irréductible récurrente positive, on peut, en principe, obtenir cette matrice de transition si on dispose d'une trajectoire complète de la CMH en question.

EXEMPLE 8.3.3: UNE POLITIQUE DE MAINTENANCE. Soit $\{U_n\}_{n \geq 1}$ une suite de variables IID à valeurs dans \mathbb{N}_+ . La variable U_n est interprétée comme la durée de vie d'une machine, la n -ème, qui est remplacée par une $(n+1)$ -ème dès qu'elle tombe en panne. Donc, au temps 0, la machine 1 est mise en service jusqu'à ce qu'elle "casse" au temps U_1 , où elle est immédiatement remplacée par la machine 2, qui casse au temps $U_1 + U_2$, et ainsi de suite. Au temps n , le temps restant avant la prochaine panne est X_n . Plus précisément, le processus $\{X_n\}_{n \geq 0}$ prend ses valeurs dans $E = \mathbb{N}$ (On suppose que ces variables prennent des valeurs arbitrairement grandes : $P(U_1 > k) > 0$ pour tout $k \geq 1$; l'autre cas se traite de manière analogue), est égal à 0 au temps $R_k = \sum_{i=1}^k U_i$, à $U_{k+1} - 1$ au temps $R_k + 1$, et alors décroît d'une unité par unité de temps jusqu'à ce qu'il atteigne la valeur 0 au temps R_{k+1} . On supposera que pour tout $k \in \mathbb{N}_+$, $P(U_1 > k) > 0$, de sorte que l'espace d'état E est \mathbb{N} . Alors $\{X_n\}_{n \geq 0}$ est une CMH de probabilités de transition :

$$\begin{aligned} p_{i, i+1} &= P(U > i+1 | U > i) = \frac{P(U > i+1)}{P(U > i)}, \\ p_{i, 0} &= P(U = i+1 | U > i) = \frac{P(U = i+1)}{P(U > i)}, \end{aligned}$$

où U est une variable aléatoire de même distribution que U_1 . Cette CMH est irréductible, comme on le vérifie facilement. Elle est récurrente positive si et seulement si $E[U] < \infty$ ($U_1 = T_0$). Dans ce cas, sa distribution stationnaire est

$$\pi(i) = \frac{P(U > i)}{E[U]}. \quad (8.40)$$

En effet, on vérifie facilement les équations de balance globale

$$\pi(i) = p_{i-1,i}\pi(i-1)$$

et

$$\pi(0) = \sum_{i=0}^{\infty} \pi(i)p_{i,0}.$$

Une visite de la CMH à l'état 0 correspond à une panne de la machine machine, et donc d'après le théorème ergodique,,

$$\pi(0) = \lim_{N \uparrow \infty} \frac{1}{N} \sum_{k=1}^N 1_{\{X_k=0\}}$$

est la fréquence empirique des pannes. On a

$$\pi(0) = E_0[T_0]^{-1},$$

où T_0 est le temps de retour en 0. Ici,

$$E_0[T_0] = E[U],$$

et donc

$$\lim_{N \uparrow \infty} \frac{1}{N} \sum_{k=1}^N 1_{\{X_k=0\}} = \frac{1}{E[U]}. \quad (8.41)$$

Le coût d'une panne peut être si important que l'on préfère remplacer une machine avant qu'elle tombe en panne (une panne entraîne des réparations coûteuses, peut-être même une catastrophe humaine, tandis qu'un remplacement se traduit par de simples coûts de maintenance). Dans la politique de *de retraite à âge fixe*, on choisit un entier $T \geq 1$ et on impose que toute machine atteignant l'âge T soit immédiatement remplacée. On veut calculer la fréquence empirique des pannes (pas des remplacements).

La CMH correspondant à cette politique est du même type que celle décrite plus haut, on remplace simplement U_n par $V_n = U_n \wedge T$. Un remplacement (pas une panne) a lieu au temps n si et seulement si $X_n = 0$ et $X_{n-1} = T-1$. Mais $X_{n-1} = T-1$ implique $X_n = 0$, et donc un remplacement a lieu au temps n si et seulement si

$$X_{n-1} = T-1.$$

La fréquence empirique de remplacements non dus à des pannes est donc, d'après le théorème ergodique,

$$\lim_{N \uparrow \infty} \frac{1}{N} \sum_{k=1}^N 1_{\{X_k=T-1\}} = \pi(T-1).$$

La formule (1) appliquée à cette nouvelle situation donne

$$\pi(T-1) = \frac{P(V \geq T)}{E[V]},$$

et donc, comme $V = U \wedge T$,

$$\pi(T - 1) = \frac{P(U \geq T)}{E[U \wedge T]}.$$

La fréquence empirique des visites à 0 est, d'après (8.41),

$$\frac{1}{E[U \wedge T]}.$$

La fréquence empirique des pannes est donc

$$\frac{1}{E[U \wedge T]} - \frac{P(U \geq T)}{E[U \wedge T]} = \frac{P(U < T)}{E[U \wedge T]}.$$

8.4 Méthode de Monte Carlo

Principe de la méthode de Monte Carlo

On a vu (Section 3.2) que pour échantillonner une distribution de probabilité π sur un espace fini $E = \{1, 2, \dots, r\}$, on dispose d'une méthode conceptuellement très simple : On tire un nombre aléatoire U uniformément distribué sur $[0, 1]$ et on définit la variable aléatoire Z en posant $Z = i$ si $\sum_{\ell=1}^{i-1} \pi(\ell) \leq U < \sum_{\ell=1}^i \pi(\ell)$. La variable aléatoire Z est alors bien un échantillon de π (c'est-à-dire : elle admet π comme distribution).

Cette méthode, dite de l'inverse, a un certain nombre d'inconvénients quand r est très grand :

- (a) Des problèmes peuvent se poser à cause de la petitesse des intervalles qui forment la partition de $[0, 1]$ adaptée à la distribution π , et du coût de la précision alors nécessaire dans les calculs.
- (b) L'espace des états ne se présente pas d'une manière naturelle sous la forme $\{1, 2, \dots, r\}$. En traitement des images et en physique statistique, un état est, par exemple, un tableau de symboles binaires $\{a_{ij}, 1 \leq i, j \leq M; \quad a_{ij} \in \{0, 1\}$, avec M très grand. Pour implémenter la méthode de l'inverse, il faut d'abord "coder" E , c'est-à-dire associer à chacun de ses éléments un nombre de 1 à r , obtenir Z qui est un nombre de 1 à r , puis "décoder" le résultat (dire quelle image correspond à ce nombre). Ces opérations, qui nécessitent des recherches dans de très grandes listes, sont coûteuses en temps de calcul.
- (c) Enfin, il existe de nombreuses situations, surtout en physique, où π n'est connu qu'à un facteur de normalisation près. La méthode de l'inverse est alors tout simplement inapplicable.

La recherche d'échantillonneurs qui surmontent ces difficultés (surtout la dernière) est un sujet important. La méthode dite MCMC ("Monte Carlo Markov chain") est basé

sur le principe suivant : on construit une CMH $\{X_n\}_{n \geq 0}$ irréductible récurrente positive apériodique dont l'espace des états est $E = \{1, 2, \dots, r\}$ et qui admet π comme distribution stationnaire. On laisse ce processus évoluer jusqu'à un temps n assez grand pour que la distribution de X_n soit assez proche de la distribution stationnaire π , et on prend comme échantillon de π la variable X_n . La distribution de X_n n'est qu'approximativement égale à π . Ceci n'est pas trop grave si on peut donner une vitesse de convergence de la distribution au temps n vers la distribution stationnaire ce qui permet de contrôler la qualité de l'approximation en choisissant n en fonction des exigences de précision. Le problème des vitesses de convergence est un domaine de recherche très actif que nous n'aborderons pas ici. Pour l'instant nous allons simplement choisir une matrice de transition irréductible récurrente positive apériodique qui admet π comme distribution stationnaire.

Il y a un nombre infini de telles matrices et, parmi elles, il y en a une infinité qui correspondent à une paire (\mathbf{P}, π) réversible, c'est-à-dire telle que

$$\pi(i)p_{ij} = \pi(j)p_{ji}. \quad (8.42)$$

Nous allons chercher des solutions de la forme

$$p_{ij} = q_{ij}\alpha_{ij} \quad (8.43)$$

pour $j \neq i$, où $Q = \{q_{ij}\}_{i,j \in E}$ est une matrice de transition sur E irréductible. La chaîne évolue de la façon suivante : Lorsque l'état présent est i , on choisit un état j avec la probabilité q_{ij} . Cet état j est accepté comme nouvel état avec la probabilité α_{ij} . S'il n'est pas accepté, on ne change pas d'état, on reste en i . La probabilité de transition de i à j quand $i \neq j$ est bien donné par (8.42). Il reste maintenant à choisir q_{ij} et α_{ij} . Nous allons décrire les deux algorithmes les plus célèbres.

EXEMPLE 8.4.1: L'ALGORITHME DE METROPOLIS. On suppose que la distribution π est de la forme

$$\pi(i) = \frac{e^{-U(i)}}{K}, \quad (8.44)$$

où $U : E \rightarrow \mathbb{R}$ est une fonction, dite "fonction énergie" dans un contexte de physique, et K est la "constante de partition", la constante de normalisation assurant que π est bien un vecteur de probabilité. On notera que la forme (8.44) n'a pas vraiment à être postulée puisque on peut toujours choisir $U(i) = -\log \pi(i)$ et $K = 1$. En pratique, la fonction d'énergie U est donnée et la constante de normalisation K est impossible à calculer numériquement. Mais nous allons voir que l'algorithme de Metropolis n'en a pas besoin. Cet algorithme préconise une matrice Q symétrique, et la probabilité d'acceptation

$$\alpha_{ij} = \min \left(1, e^{-(U(j)-U(i))} \right).$$

Il est facile de vérifier que les équations de balance détaillée (8.42) sont satisfaites, et que la CMH en question est bien irréductible et, lorsque la fonction énergie n'est pas une constante, apériodique.

EXEMPLE 8.4.2: L'ALGORITHME DE BARKER. Cet algorithme utilise lui aussi une matrice Q symétrique. Sa probabilité d'acceptation est

$$\alpha_{ij} = \frac{e^{-U(i)}}{e^{-U(i)} + e^{-U(j)}}.$$

Ce choix correspond au principe de base de la physique statistique : quand la Nature a le choix entre deux états 1 et 2 d'énergies respectives E_1 et E_2 , elle choisit $i = 1, 2$ avec la probabilité $\frac{e^{-E_i}}{e^{-E_1} + e^{-E_2}}$. Là encore, on vérifie que les équations de balance détaillée (8.42) sont satisfaites, et que la CMH en question est bien irréductible et, lorsque la fonction énergie n'est pas une constante, apériodique.

EXEMPLE 8.4.3: L'ALGORITHME DE GIBBS. L'espace des états est $E = \Lambda^N$, où N est un entier positif et Λ est un ensemble fini. La distribution à échantillonner est donc de la forme

$$\pi(z) = \pi(z(1), \dots, z(N))$$

Le mécanisme de transition est le suivant. Si on se trouve dans l'état $(z(1), \dots, z(N))$, on choisit un "site" ℓ , $1 \leq \ell \leq N$, au hasard, et on change la coordonnée $z(\ell)$ (et elle seule) en $y(\ell)$, cette nouvelle coordonnée étant choisie en fonction de $z(1), \dots, z(\ell - 1), z(\ell + 1), \dots, z(N)$ avec la probabilité

$$\pi(y(\ell) \mid z(1), \dots, z(\ell - 1), z(\ell + 1), \dots, z(N)). \quad (8.45)$$

Là encore, on vérifie que les équations de balance détaillée (8.42) sont satisfaites, que la CMH en question est bien irréductible et, lorsque π n'est pas une constante, apériodique.

Cette méthode est spécialement intéressante lorsque Λ (l'"espace des phases") est petit, et lorsque la probabilité conditionnelle $\pi(\cdot \mid z(1), \dots, z(\ell - 1), z(\ell + 1), \dots, z(N))$ ne dépend que d'un petit nombre des arguments $z(1), \dots, z(\ell - 1), z(\ell + 1), \dots, z(N)$. Voici un exemple qui présente un grand intérêt pour les physiciens :

EXEMPLE 8.4.4: L'ÉCHANTILLONNEUR DE GIBBS POUR LE MODÈLE D'ISING. Le modèle d'Ising est une idéalisation d'un matériau ferromagnétique. On a $N = M^2$ dipôles magnétiques placés sur une grille finie. Plus précisément les sites sur lesquels se trouvent les dipôles forment un ensemble $S = \{(i, j); 1 \leq i, j \leq M\}$. Un site $s \in S$ a donc la forme $s = (i, j)$. La distance entre deux sites $s_1 = (i_1, j_1)$ et $s_2 = (i_2, j_2)$ est $d(s_1, s_2) = |i_1 - i_2| + |j_1 - j_2|$. On dit que s_1 et s_2 sont voisins si $d(s_1, s_2) = 1$. Deux sites voisins s_1 et s_2 forment une paire notée $\langle s_1, s_2 \rangle$. L'espace des phases est $\Lambda = \{-1, +1\}$, la valeur -1 correspond à une orientation de dipôle, disons vers le bas, tandis que $+1$ correspond à la direction opposée. L'espace d'états $E = \Lambda^S$ est l'ensemble des "configurations" $(z(s), s \in S)$ où $z(s) \in \Lambda = \{-1, +1\}$. Une telle configuration représente des dipôles placés sur les sites de S , l'orientation du dipôle placé en s étant $z(s)$. Si on

énumère les sites de S de 1 à $N = M^2$, on retrouve bien la situation décrite plus haut. Précisons maintenant la distribution π . On prendra

$$\pi(z(s), s \in S) = \frac{\exp\{-\mathcal{E}(z)\}}{K}$$

où $\mathcal{E}(z)$ est l'énergie de la configuration $z = (z(s), s \in S) : \mathcal{E}(z) = H \sum_{s \in S} z(s) + J \sum_{\langle s, t \rangle} z(s)z(t)$ (la deuxième somme porte sur toutes les paires de sites voisins) et K est une constante de normalisation, en général incalculable numériquement. (Pour les physiciens H est le champ magnétique externe, et J est l'énergie interne d'une paire de dipôles voisins orientés dans le même sens.) La probabilité conditionnelle jouant le rôle de (8.45),

$$\pi(y(s) | z(t), t \in S - \{s\}) = \frac{\pi(y(s), z(t), t \in S - \{s\})}{\sum_{z(s) \in \Lambda} \pi(z(s), z(t), t \in S - \{s\})},$$

prend la forme (faire le calcul)

$$\pi(y(s) | z(t), t \in S - \{s\}) = \frac{\exp\{\mathcal{E}(s, z)\}}{\exp\{\mathcal{E}_{+1}(s, z)\} + \exp\{\mathcal{E}_{-1}(s, z)\}}, \quad (8.46)$$

où $\mathcal{E}(s, z) = y(s)(H + J \sum z(v(s)))$ et où la somme porte sur tous les sites $v(s)$ voisins de s . En physique, on appelle $\mathcal{E}(s, z)$ l'énergie locale au site s de la configuration $(y(s), z(t), t \in S - \{s\})$, et $\mathcal{E}_{+1}(s, z)$ et $\mathcal{E}_{-1}(s, z)$ sont les valeurs de cette énergie locale correspondant aux directions $+1$ et -1 respectivement de l'orientation du dipôle placé en s . L'échantillonneur de Gibbs fonctionne donc dans le cas présent de la façon suivante : si au temps n on a la configuration $z = (z(s), s \in S)$, on choisit un site complètement au hasard (distribution uniforme). Si c'est le site s qui a été tiré au sort, on tire au sort la nouvelle phase $y(s)$ de ce site s selon la probabilité (8.46). On notera que ce choix est fait selon les principes de la physique statistique décrit quelques lignes plus haut.

8.5 Exercices

Exercice 8.5.1. UN CONTRE-EXEMPLE.

La propriété de Markov ne dit pas que le présent et le futur sont indépendants étant donné une information *quelconque* sur le présent. Trouvez un exemple simple de CMH $\{X_n\}_{n \geq 0}$ avec l'espace d'état $E = \{1, 2, 3, 4, 5, 6\}$ tel que

$$P(X_2 = 6 | X_1 \in \{3, 4\}, X_0 = 2) \neq P(X_2 = 6 | X_1 \in \{3, 4\}).$$

Exercice 8.5.2.

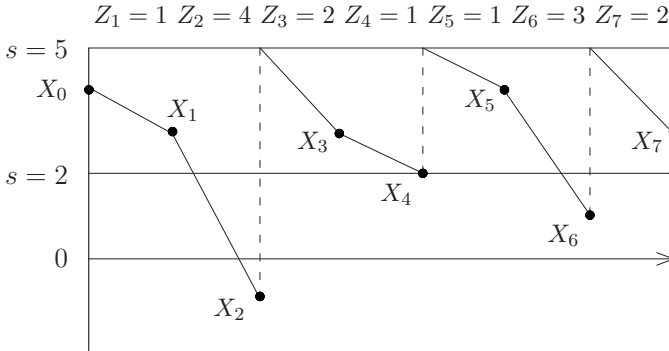
Démontrez l'égalité (8.2).

Exercice 8.5.3.

Démontrez le Théorème 8.1.5.

Exercice 8.5.4. GESTION DES STOCKS.

Une marchandise donnée A est stockée en vue de satisfaire à la demande. La demande totale entre le temps n et le temps $n + 1$ est de Z_{n+1} unités, et on suppose que la suite $\{Z_n\}_{n \geq 1}$ est IID, et indépendante de la valeur initiale X_0 du stock. Le remplissage du stock a lieu aux temps $n + 0$ (c'est-à-dire, immédiatement après le temps n) pour tout $n \geq 1$.



Une stratégie de gestion populaire est la stratégie (s, S) , où s et S sont des entiers tels que $0 < s < S$. Avec cette politique de gestion, si le niveau du stock au temps n est plus gret que s , alors le stock est ramené au niveau S autemps $n + 0$. Autrement, rien n'est fait. Le stock initial X_0 est supposé inférieur ou égal à S , et donc $\{X_n\}_{n \geq 1}$ prend ses valeurs dans $E = \{S, S - 1, S - 2, \dots\}$. (Voir la figure.) Les valeurs négatives du stock sont admises, avec interprétation qu'une commande non satisfaite est immédiatement honorée après restockage. Montrez que $\{X_n\}_{n \geq 1}$ est une CMH et donnez sa matrice de transition.

Exercice 8.5.5. * RECORDS.

Soit $\{Z_n\}_{n \geq 1}$ une suite IID de variables géométriques (pour $k \geq 0$, $P(Z_n = k) = (1-p)^k p$, où $p \in (0, 1)$). Soit $X_n = \max(Z_1, \dots, Z_n)$ la *valeur record* au temps n , où on suppose que X_0 est une variable à valeurs entières et indépendante de la suite $\{Z_n\}_{n \geq 1}$. Montrez que $\{X_n\}_{n \geq 0}$ est une CMH et donnez sa matrice de transition.

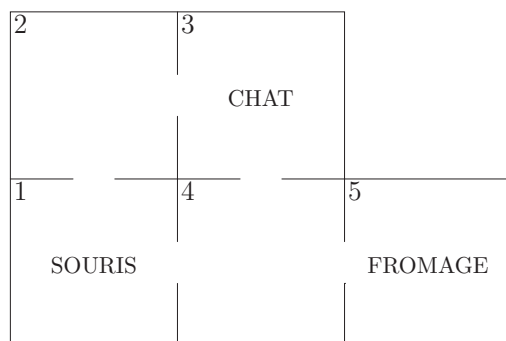
Exercice 8.5.6. * LA VIE DES GANGSTERS.

Trois personnages armés, A , B , et C , se trouvent soudainement en présence au carrefour d'une rue de Washington, D.C., et sur ce, se mettent tout naturellement à se tirer dessus. Chaque survivant tire sur un autre survivant de son choix toutes les 10 secondes. Les probabilités d'atteindre la cible pour A , B , et C sont respectivement α , β , et γ . A est le plus haï des trois, et donc, tant qu'il vit, B et C s'ignorent et lui tirent dessus. Pour des raisons historiques que nous ne développerons pas, A ne peut pas sentir B , et donc il ne

tire que sur B tant que ce dernier est vivant. Le bienheureux C n'est visé que lorsqu'il se trouve en présence de A seul ou B seul. Quelles sont les chances de survie de A, B , et C , respectivement ?

Exercice 8.5.7. * LE CHAT, LA SOURIS ET LE GRUYÈRE.

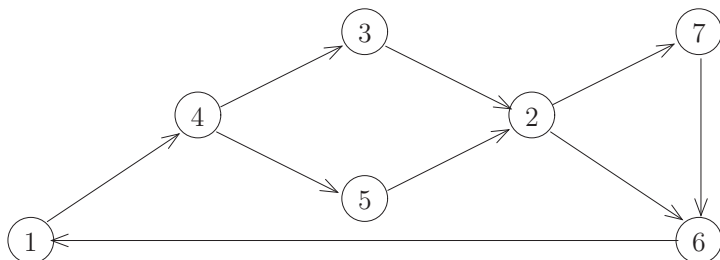
Une souris affairée se promène dans un labyrinthe. Si au temps n elle se trouve dans une pièce avec k portes, elle en choisit une avec la probabilité $\frac{1}{k}$ et se retrouve à l'instant $n+1$ dans la pièce à laquelle cette porte conduit. Un chat paresseux attend dans la pièce numéro 3, et il y a un morceau de fromage dans la pièce numéro 5. La souris commence son périple dans la pièce 1. Avec quelle probabilité goûtera-t-elle du fromage avant que le chat ne la dévore ?



La chambre au gruyère

Exercice 8.5.8.

Montrez que le graphe de transition de la figure ci-dessous est irréductible. Donnez sa période et ses classes cycliques.



Exercice 8.5.9.

Montrez qu'une matrice de transition \mathbf{P} avec au moins un état $i \in E$ tel que $p_{ii} > 0$ est apériodique.

Exercice 8.5.10. *

Montrez que la marche aléatoire symétrique sur \mathbb{Z} n'a pas de distribution stationnaire.

Exercice 8.5.11.

Est-ce que la CMH de l'Exemple 8.1.9 est réversible ?

Exercice 8.5.12.

Soit $\{X_n\}_{n \geq 0}$ la CMH de l'Exemple 8.1.7.

(1) Montrez qu'elle est réversible ;

(2) Sachant $X_0 = 1$, calculez la distribution de probabilité de $T_1 = \inf \{n \geq 1; X_n = 1\}$, où on utilise la convention $\inf \{\emptyset\} = \infty$.

Exercice 8.5.13. *

Calculez la distribution stationnaire de la CMH d'espace d'état $E = \{1, 2, 3\}$ et de matrice de transition

$$\mathbf{P} = \begin{pmatrix} 1 & 1 - \alpha & \alpha & 0 \\ 0 & 1 - \beta & \beta & 0 \\ \gamma & 0 & 1 - \gamma & 0 \end{pmatrix},$$

où $\alpha, \beta, \gamma \in (0, 1)$. Est-elle réversible ?

Exercice 8.5.14. *

Démontrez le Théorème 8.2.1

Exercice 8.5.15. LES CAILLOUX.

Des cailloux S_1, \dots, S_M sont alignés. Au temps n un caillou est choisi au hasard, et ce caillou échange sa place avec le caillou placé juste devant lui. Si le caillou sélectionné est en tête, on ne change rien. Par exemple, avec $M = 5$: Si la situation juste avant le temps n est $S_2 S_3 S_1 S_5 S_4$ (S_2 est en tête), et si S_5 est tiré au sort, la nouvelle situation est $S_2 S_3 S_5 S_1 S_4$, tandis que si S_2 est sélectionné, la configuration reste la même. À chaque tour de l'horloge, S_i est sélectionné avec la probabilité $\alpha_i > 0$. Notons X_n la situation au temps n , par exemple $X_n = S_{i_1} \dots S_{i_M}$, avec l'interprétation que S_{i_j} est dans la j -ème position. Montrez que $\{X_n\}_{n \geq 0}$ est une CMH irréductible récurrente positive et que sa distribution stationnaire est

$$\pi(S_{i_1} \dots S_{i_M}) = C \alpha_{i_1}^M \alpha_{i_2}^{M-1} \dots \alpha_{i_M},$$

où C est une constante de normalisation.

Exercice 8.5.16. CHAÎNE PRODUIT.

Soit $\{X_n^{(1)}\}_{n \geq 0}$ et $\{X_n^{(2)}\}_{n \geq 0}$ deux CMH avec la même matrice de transition \mathbf{P} . Montrez que le processus $\{Z_n\}_{n \geq 0}$ à valeurs dans $E \times E$ défini par $Z_n = (X_n^{(1)}, X_n^{(2)})$ est une CMH. Quelle est sa matrice de transition en n étapes ? Montrez qu'elle est irréductible si \mathbf{P} est

irréductible et apériodique. Donnez un contre-exemple lorsqu'on abandonne l'hypothèse d'apériodicité.

Exercice 8.5.17.

Soit $X_0^1, X_0^2, Z_n^1, Z_n^2$ ($n \geq 1$) des variables aléatoires indépendantes, et telles que, de plus, Z_n^1, Z_n^2 ($n \geq 1$) sont identiquement distribuées. Soit τ une variable aléatoire à valeurs entières non négatives telle que pour tout $m \in \mathbb{N}$, l'événement $\{\tau = m\}$ est exprimable en fonction de $X_0^1, X_0^2, Z_n^1, Z_n^2$ ($n \leq m$). On définit $\{Z_n\}_{n \geq 1}$ par

$$Z_n = \begin{cases} = Z_n^1 & \text{si } n \leq \tau \\ = Z_n^2 & \text{si } n > \tau \end{cases}$$

Montrez que $\{Z_n\}_{n \geq 1}$ a la même distribution que $\{Z_n^1\}_{n \geq 1}$ et est indépendante de X_0^1, X_0^2 .

Exercice 8.5.18. FUSION.

Soit $\{X_n^1\}_{n \geq 0}$ et $\{X_n^2\}_{n \geq 0}$ deux CMH avec la même matrice de transition \mathbf{P} . Soit τ le temps défini par

$$\tau = \inf\{n \geq 0; X_n^1 = X_n^2\}$$

(avec la convention usuelle: $\inf \emptyset = \infty$). Supposons que $P(\tau < \infty) = 1$. On définit $\{X_n\}_{n \geq 1}$ par

$$X_n = \begin{cases} X_n^1 & \text{if } n \leq \tau \\ X_n^2 & \text{if } n > \tau \end{cases}$$

Montrez que $\{X_n\}_{n \geq 1}$ a la même distribution que $\{X_n^1\}_{n \geq 1}$.

Exercice 8.5.19. LA CHAÎNE SERPENT.

Soit $\{X_n\}_{n \geq 0}$ une CMH d'espace d'état E et de matrice de transition \mathbf{P} . Pour $L \geq 1$, on définit $Y_n = (X_n, X_{n+1}, \dots, X_{n+L})$.

(a) Le processus $\{Y_n\}_{n \geq 0}$ prend ses valeurs dans $F = E^{L+1}$. Montrez que c'est une CMH et donnez sa matrice de transition.

(b) Montrez que si $\{X_n\}_{n \geq 0}$ est irréductible, il en est de même pour $\{Y_n\}_{n \geq 0}$ si on restreint l'espace d'état de cette dernière à $F = \{(i_0, \dots, i_L) \in E^{L+1}; p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{L-1} i_L} > 0\}$.

(c) Montrez que si $\{X_n\}_{n \geq 0}$ a une distribution stationnaire π , alors $\{Y_n\}_{n \geq 0}$ a aussi une distribution stationnaire. Laquelle?

Exercice 8.5.20. * RETOUR À L'ÉTAT INITIAL.

Soit τ le temps de retour à l'état initial d'une CMH irréductible récurrente positive $\{X_n\}_{n \geq 0}$, c'est-à-dire,

$$\tau = \inf\{n \geq 1; X_n = X_0\},$$

Calculez l'espérance de τ lorsque la distribution initiale est la distribution stationnaire π . Conclure que cette espérance est finie si et seulement si E est fini. Quand E est infini, est-ce que ceci est en contradiction avec l'hypothèse de récurrence positive ?

Exercice 8.5.21. * LE CAVALIER RENTRE À LA MAISON.

Un cavalier circule de manière aléatoire sur un échiquier, choisissant chaque mouvement parmi ceux qui lui sont permis avec la même probabilité, et débutant son périple d'un coin de l'échiquier. Combien de temps en moyenne lui faudra-t-il pour se retrouver sur sa case de départ ?

Exercice 8.5.22. CODAGE ALTERNATIF.

Dans certains systèmes de communication numérique, une suite de 0 et de 1 (symboles d'entrée) est codée en une suite de 0, +1 et -1 (symboles de sortie) de la manière suivante. Un symbole d'entrée 0 est codé en 0, tandis qu'un symbole d'entrée 1 est codé en -1 ou +1. Le choix entre -1 et +1 est fait de telle sorte que les -1 et les +1 alternent. Le premier 1 est codé en +1. Par exemple la suite de symboles d'entrée 011101 devient 0, +1, -1, +1, 0, -1.

a. Trouvez un automate avec 4 états +1, -1, 0_+ et 0_- , pour lequel la suite des états visités, à part l'état initial fixé à 0_+ , est, lorsque 0_+ et 0_- sont remplacés par 0, exactement la suite de sortie.

b. On suppose que la suite d'entrée $\{Z_n\}_{n \geq 1}$ est IID, avec 0 et 1 équiprobables. La suite des états de l'automate est alors une CMH dont on demande de calculer la matrice de transition \mathbf{P} et ses itérées \mathbf{P}^n , et la distribution stationnaire π .

c. Notons $\{Y_n\}_{n \geq 0}$ la suite de symboles de sortie (prenant les valeurs $\{0, -1, +1\}$). Montrez que $Y_n = f(X_n)$ pour une fonction f à identifier, et calculez $\lim_{n \rightarrow \infty} \{E[Y_n Y_{n+k}] - E[Y_n]E[Y_{n+k}]\}$ pour tout $k \geq 0$.

Exercice 8.5.23. * ABBABAA.

Une suite de A et de B est formée comme suit. La première lettre est choisie au hasard, $P(A) = P(B) = \frac{1}{2}$, ainsi que la deuxième, indépendamment de la première. Quand les $n \geq 2$ premières lettres ont été sélectionnées, la $(n+1)$ -ème est choisie, indépendamment des lettres dans les positions $k \leq n-2$, et conditionnellement à la paire formée par les lettres en position $n-1$ et n , comme suit :

$$P(A | AA) = \frac{1}{2}, P(A | AB) = \frac{1}{2}, P(A | BA) = \frac{1}{4}, P(A | BB) = \frac{1}{4}.$$

Quelles sont les proportions de A et de B au long terme ?

Exercice 8.5.24. RECHERCHE D'UN MOTIF.

Considérons le tableau de a et de b de la figure A ci-dessous où une lettre dans une position donnée est choisie au hasard et équiprobablement dans l'ensemble $\{a, b\}$,

indépendamment des autres lettres. On veut calculer la fréquence empirique asymptotique du motif de la figure *B*, sans compter les chevauchements. Par exemple, avec la suite de la figure *A*, on compte 2 occurrences du motif ; La troisième est ignorée car elle chevauche la deuxième (Figure *C*).

b a b b a b a b b b
b a a a b b a a a a A

b . b
. a a B

b a b *b a* *b a b b*
b a a *a b* *b a a a a* C
 OUI OUI NON

Trouvez un automate qui lit successivement les colonnes à deux lettres de gauche à droite, et qui possède un état privilégié *** avec la propriété suivante : l'automate entre ou reste dans l'état *** si et seulement si il vient de découvrir un motif qui ne chevauche pas un autre précédemment découvert. Quelle est la fréquence empirique asymptotique du motif ?

Solutions des exercices

Solutions des exercices du chapitre 1

SOLUTION (Exercice 1.4.1).

Utilisez la formule de Morgan, $\cap_{i=1}^n A_i = \overline{(\cup_{i=1}^n \overline{A_i})}$ ($1 \leq n \leq \infty$).

SOLUTION (Exercice 1.4.2).

Posons $A_i = \{r_i\} = [r_i, r_i]$ où $(r_i, i \geq 1)$ est une énumération de l'ensemble des rationnels \mathbf{Q} . On a $\mathbf{Q} = \cup_{i=1}^{\infty} A_i$. Mais \mathbf{Q} n'est pas une somme *finie* d'intervalles. En effet supposons qu'il en soit ainsi. Alors de deux choses l'une : ou bien l'un des intervalles en question ne se réduit pas à un point et a donc la puissance du continu ce qui contredit le fait que \mathbf{Q} est dénombrable ; ou bien tous les intervalles dont la somme est \mathbf{Q} sont des points ce qui contredit le fait que \mathbf{Q} contient un nombre infini de points.

SOLUTION (Exercice 1.4.3).

Évident.

SOLUTION (Exercice 1.4.4).

$\Omega = [0, 1]^2$, $\mathcal{A} = \{\emptyset, \Omega, A_1, A_2\}$ et $\mathcal{B} = \{\emptyset, \Omega, B_1, B_2\}$, où $A_1 = [0, \frac{1}{2}] \times [0, 1]$, $A_2 = \overline{A_1}$, $B_1 = [0, 1] \times [0, \frac{1}{2}]$, $B_2 = \overline{B_1}$. On a $A_1 \cap B_1 \notin \mathcal{A} \cup \mathcal{B}$.

SOLUTION (Exercice 1.4.5).

\emptyset et Ω sont dans $\cup_{n=1}^{\infty} \mathcal{A}_n$ (évident). Si $A \in \cup_{n=1}^{\infty} \mathcal{A}_n$, alors $A \in \mathcal{A}_m$ pour un m au moins (définition de $\cup_{n=1}^{\infty} \mathcal{A}_n$), donc $\overline{A} \in \mathcal{A}_m \subset \cup_{n=1}^{\infty} \mathcal{A}_n$. Si $A \in \cup_{n=1}^{\infty} \mathcal{A}_n, B \in \cup_{n=1}^{\infty} \mathcal{A}_n$, alors il existe m_1 et m_2 tel que $A \in \mathcal{A}_{m_1}, B \in \mathcal{A}_{m_2}$. Posons $m = \sup(m_1, m_2)$. Alors $A \in \mathcal{A}_m$ et $B \in \mathcal{A}_m$, donc $A \cap B \in \mathcal{A}_m \subset \cup_{n=1}^{\infty} \mathcal{A}_n$.

SOLUTION (Exercice 1.4.6).

Prendre $A'_1 = A_1$ et, pour $i \geq 1$, $A'_{i+1} = A_{i+1} \cap \left(\overline{\bigcup_{j=1}^i A_j}\right)$.

SOLUTION (Exercice 1.4.8).

Non.

SOLUTION (Exercice 1.4.9).

Posons $C_n = \bigcap_{m \geq n} A_m$. C'est l'ensemble des ω qui appartiennent à A_m pour tout $m \geq n$.
 $\omega \in B = \bigcup_{n \geq 1} C_n$ si et seulement si il existe un $N = N(\omega)$ tel que $\omega \in C_N$, et donc si $\omega \in A_m$ pour tous les $n \geq N$.

SOLUTION (Exercice 1.4.10).

On utilise le fait que la suite $(X_n(\omega), n \geq 1)$ à valeurs dans $\{0, 1\}$ tend vers 0 si et seulement si à partir d'un certain rang tous les $X_n(\omega)$ sont égaux à 0, et on se sert du résultat de l'Exercice 1.4.9.

SOLUTION (Exercice 1.4.11).

Dans le cas $n = 2$, la formule de Poincaré se vérifie directement. En effet $A \cup B = A + (B - A \cap B)$ et donc $P(A \cup B) = P(A) + P(B - A \cap B) = P(A) + P(B) - P(A \cap B)$. Le cas général se démontre par induction. Supposons la formule vraie pour n . Posons

$$\bigcup_{j=1}^{n+1} A_j = B \cup A_{n+1} \text{ où } B = \bigcup_{j=1}^n A_j.$$

D'après la formule de Poincaré pour $n = 2$,

$$P\left(\bigcup_{j=1}^{n+1} A_j\right) = P(B) + P(A_{n+1}) - P(B \cap A_{n+1}).$$

Pour $P(B) = P(\bigcup_{j=1}^{n+1} A_j)$ on emploie la formule supposée vraie pour n . Les termes qui manquent pour le cas $n + 1$ sont obtenus en remarquant que

$$P(B \cap A_{n+1}) = P\left(\bigcup_{j=1}^n \tilde{A}_j\right) \text{ où } \tilde{A}_j = A_j \cap A_{n+1},$$

et en utilisant la formule de Poincaré à l'ordre n pour les \tilde{A}_j .

SOLUTION (Exercice 1.4.12).

On passe aux complémentaires par la formule de de Morgan :

$$\cap_{n=1}^{\infty} C_n = \overline{\left(\cup_{n=1}^{\infty} \overline{C_n}\right)}.$$

Comme la suite $(C_n, n \geq 1)$ décroît, la suite $(\overline{C_n}, n \geq 1)$ croît. Il suffit alors d'appliquer le Théorème 1.2.1.

SOLUTION (Exercice 1.4.13).

Non car $P(A \cap B) = P(\emptyset) = 0$ ne peut être égal à $P(A)P(B) > 0$.

SOLUTION (Exercice 1.4.14).

On a

$$\begin{aligned} P(A \cap B \cap C) &= P(A|B \cap C)P(B \cap C), \\ P(B \cap C) &= P(B|C)P(C), \end{aligned}$$

et

$$P(A \cap B \cap C) = P(A \cap B|C)P(C),$$

d'où le résultat.

SOLUTION (Exercice 1.4.15).

En raisonnant par induction, il suffit de montrer que si l'on change un A_i en $\overline{A_i}$, la famille $\{A_1, \dots, A_{i-1}, \overline{A_i}, A_{i+1}, \dots, A_n\}$ est une famille d'événements mutuellement indépendants. Pour simplifier les notations, prenons $i = 1$. Il suffit donc de prouver, pour tout événement B produit d'événements choisis dans la famille $\{A_2, \dots, A_n\}$, par exemple $B = A_2 \cap A_4$, que $P(\overline{A_1} \cap B) = P(\overline{A_1})P(B)$ puisque l'on sait que $P(B) = P(A_2 \cap A_4) = P(A_2)P(A_4)$. Mais, comme $\overline{A_1} \cap B + A_1 \cap B = B$, on a

$$\begin{aligned} P(\overline{A_1} \cap B) &= P(B) - P(A_1 \cap B) \\ &= P(B) - P(A_1)P(B) \\ &= P(B)(1 - P(A_1)) = P(B)P(\overline{A_1}). \end{aligned}$$

SOLUTION (Exercice 1.4.16).

Il s'agit d'un imposteur. En effet, d'après la formule de Poincaré et en utilisant l'hypothèse $\Omega = \cup_{i=1}^n A_i$ et l'indépendance des A_i ,

$$1 = P(\Omega) = P\left(\bigcup_{i=1}^n A_i\right) = nx - C_n^2 x^2 + C_n^3 x^3 - \dots + (-1)^{n+1} x^n$$

où x est la probabilité commune des A_i . On aurait donc $(1 - x)^n = 0$ ce qui n'est possible que si $x = 1$.

SOLUTION (Exercice 1.4.17).

On vérifie d'abord que $P_B : \mathcal{F} \rightarrow 1$ et $P_B(\Omega) = 1$. Ensuite, on vérifie la propriété de sigma-additivité :

$$\begin{aligned}
 P_B \left(\sum_{i=1}^{\infty} A_i \right) &= \frac{P((\sum_{i=1}^{\infty} A_i) \cap B)}{P(B)} \\
 &= \frac{P(\sum_{i=1}^{\infty} A_i \cap B)}{P(B)} \\
 &= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} \\
 &= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} \\
 &= \sum_{i=1}^{\infty} P(A_i | B) = \sum_{i=1}^{\infty} P_B(A_i) .
 \end{aligned}$$

SOLUTION (Exercice 1.4.18).

On applique la formule de Poincaré $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ pour obtenir $P(A \cap B) = P(A) + P(B) - P(A \cup B)$ qu'on compare ensuite à $P(A)P(B)$:

	$P(A)$	$P(B)$	$P(A \cup B)$	$P(A \cap B)$	$P(A)P(B)$	résultats
Cas I	0.1	0.9	0.91	0.09	0.09	indépendants
Cas II	0.4	0.6	0.76	0.24	0.24	indépendants
Cas III	0.5	0.3	0.73	0.07	0.15	non-indépendants

SOLUTION (Exercice 1.4.19).

$P(A) = P(\omega_1) + P(\omega_2) = \frac{1}{2}$, de même $P(B) = P(C) = \frac{1}{2}$; $P(A \cap B) = P(\omega_2) = \frac{1}{4}$, de même $P(A \cap C) = P(B \cap C) = \frac{1}{4}$. On a donc les indépendances 2 à 2 annoncées, car, par exemple $P(A \cap B) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(A)P(B)$. Mais

$$P(A \cap B \cap C) = P(\emptyset) = 0 \neq P(A)P(B)P(C) = \frac{1}{8}.$$

SOLUTION (Exercice 1.4.20).

Appelons X_i l'événement " A_i est ouvert", Y_i l'événement " B_i est ouvert" et Z_i l'événement " C_i est ouvert". Les données sont :

$$\begin{cases} P(X_1) = 0.5 & P(X_2) = 0.1 \\ P(Y_1) = 0.8 & P(Y_2) = 0.1 & P(Y_3) = 0.4 \\ P(Z_1) = 0.3 . \end{cases}$$

A cela il faut ajouter la donnée qualitative : les relais sont indépendants. On a donc des égalités du type

$$P(\overline{X}_1 \cap \overline{X}_2) = P(\overline{X}_1)P(\overline{X}_2) ,$$

et de même si on appelle A l'événement "la branche A passe", on aura par exemple

$$P(\overline{A} \cap \overline{B} \cap \overline{C}) = P(\overline{A})P(\overline{B})P(\overline{C}) .$$

Comme $A = \overline{X}_1 \cap \overline{X}_2$, on a $P(A) = P(\overline{X}_1)P(\overline{X}_2)$. De même $P(B) = P(\overline{Y}_1)P(\overline{Y}_2)P(\overline{Y}_3)$ et $P(C) = P(\overline{Z}_1)$. Si K est l'événement "le circuit passe", alors $\overline{K} = \overline{A} \cap \overline{B} \cap \overline{C}$ si bien que

$$P(K) = 1 - P(\overline{K}) = 1 - P(\overline{A})P(\overline{B})P(\overline{C})$$

et donc

$$P(K) = 1 - (1 - P(\overline{X}_1)P(\overline{X}_2))(1 - P(\overline{Y}_1)P(\overline{Y}_2)P(\overline{Y}_3))(1 - P(\overline{Z}_1)) .$$

SOLUTION (Exercice 1.4.21).

Soit X_1, X_2, X_3 trois variables aléatoires indépendantes prenant les valeurs 0 et 1, et indistinctement distribuées selon

$$P(X_i = 1) = 1 - P(X_i = 0) = p .$$

L'interprétation est la suivante. Si $X_1 = 1$ le bagage a été perdu à Los Angeles, si $X_1 = 0$ et $X_2 = 1$ il a été perdu à New York, si $X_1 = 0$ et $X_2 = 0$ et $X_3 = 1$ il a été perdu à Londres. Soit M l'événement : le bagage n'arrive pas à Paris. On a

$$\overline{M} = (X_1 = 0, X_2 = 0, X_3 = 0) ,$$

et donc en utilisant l'indépendance des X_i (hypothèse naturelle)

$$P(M) = 1 - P(\overline{M}) = 1 - (1 - p)^3 .$$

On a à calculer les probabilités x , y et z que le bagage soit resté à Los Angeles, New York ou Londres respectivement sachant que le bagage n'est pas arrivé à Paris. En se souvenant que

$$M = \{X_1 = 1\} + \{X_1 = 0, X_2 = 1\} + \{X_1 = 0, X_2 = 0, X_3 = 1\} ,$$

on a

$$\begin{aligned} x &= P(X_1 = 1|M) = P(M, X_1 = 1)/P(M) \\ &= P(X_1 = 1)/P(M) = \frac{p}{1 - (1 - p)^3} . \end{aligned}$$

De même,

$$y = P(X_1 = 0, X_2 = 1)/P(M) = \frac{p(1-p)}{1 - (1-p)^3}$$

$$z = P(X_1 = 0, X_2 = 0, X_3 = 1)/P(M) = \frac{p^2(1-p)}{1 - (1-p)^3}.$$

SOLUTION (Exercice 1.4.22).

Soit X_i l'état de la i -ème montre examinée, $X_i = 0$ si elle est défectueuse, $X_i = 1$ si elle est acceptable. Soit Y la provenance du lot, $Y = J$ ou C selon que la montre provient de Junkcity ou de Cheaptown. On a les données explicites

$$P(X_i = 0|Y = J) = 1/200, P(X_i = 0|Y = C) = 1/1000.$$

On suppose que *dans une usine donnée*, les états des montres successives sont indépendants, et donc par exemple,

$$P(X_1 = 1, X_2 = 1|Y = J) = P(X_1 = 1|Y = J)P(X_2 = 1|Y = J)$$

et

$$P(X_1 = 1, X_2 = 1|Y = C) = P(X_1 = 1|Y = C)P(X_2 = 1|Y = C).$$

Deux données manquent, à savoir les probabilités a priori que le lot provienne de Junkcity ou de Cheaptown. Faute de renseignements, on prendra l'hypothèse symétrique

$$P(Y = J) = P(Y = C) = \frac{1}{2}.$$

Nous pouvons maintenant calculer $x = P(X_2 = 1|X_1 = 1)$ en appliquant d'abord la définition de la probabilité conditionnelle :

$$x = P(X_1 = 1, X_2 = 1)/P(X_1 = 1),$$

puis la règle des causes totales :

$$P(X_1 = 1, X_2 = 1) = P(X_1 = 1, X_2 = 1|Y = J)P(Y = J) \\ + P(X_1 = 1, X_2 = 1|Y = C)P(Y = C)$$

et

$$P(X_1 = 1) = P(X_1 = 1|Y = J)P(Y = J) + P(X_1 = 1|Y = C)P(Y = C).$$

On a donc

$$x = \frac{\frac{1}{2}P(X_1 = 1|Y = J)^2 + \frac{1}{2}P(X_1 = 1|Y = C)^2}{\frac{1}{2}P(X_1 = 1|Y = J) + \frac{1}{2}P(X_1 = 1|Y = C)}.$$

Numériquement :

$$x = \frac{(190/200)^2 + (999/1000)^2}{190/200 + 999/1000}.$$

Commentaire. Pour une usine *donnée*, les états de 2 montres successives sont indépendants par exemple,

$$P(X_1 = 1, X_2 = 1 | J) = P(X_1 = 1 | J)P(X_2 = 1 | J) ,$$

avec l'égalité analogue pour Cheaptown. Mais ces indépendances *conditionnelles* n'entraînent pas l'indépendance tout court, car on vérifie que

$$P(X_1 = 1, X_2 = 1) \neq P(X_1 = 1)P(X_2 = 1) .$$

SOLUTION (Exercice 1.4.23).

Soit X , Y et Z les résultats des tirs de chacun des 3 chasseurs : 0 si la balle n'atteint pas l'éléphant, 1 si le tir a touché la bête. Les données explicites sont

$$P(X = 1) = \frac{1}{4} , P(Y = 1) = \frac{1}{2} , P(Z = 1) = \frac{3}{4} .$$

On suppose que les chasseurs tirent indépendamment l'un de l'autre. L'événement "2 balles et 2 balles seulement ont atteint l'éléphant" s'écrit

$$A = \{X = 1, Y = 1, Z = 0\} + \{X = 1, Y = 0, Z = 1\} + \{X = 0, Y = 1, Z = 1\} .$$

Calculons d'abord $x = P(X = 0 | A)$. On a d'après la définition de la probabilité conditionnelle, $x = P(X = 0, A) / P(A)$. Mais

$$P(X = 0, A) = P(X = 0, Y = 1, Z = 1) = P(X = 0)P(Y = 1)P(Z = 1)$$

et

$$P(A) = P(X = 1)P(Y = 1)P(Z = 0) + P(X = 1)P(Y = 0)P(Z = 1) \\ + P(X = 0)P(Y = 1)P(Z = 1) ,$$

où on a utilisé l'hypothèse d'indépendance des tirs et la règle d'additivité. Il vient alors

$$x = \frac{(1 - \frac{1}{4})(\frac{1}{2})(\frac{3}{4})}{(1 - \frac{1}{4})(\frac{1}{2})(\frac{3}{4}) + (\frac{1}{4})(1 - \frac{1}{2})(\frac{3}{4}) + (\frac{1}{4})(\frac{1}{2})(1 - \frac{3}{4})} = \frac{9}{9 + 3 + 1} = \frac{9}{13} .$$

De façon analogue on trouve les probabilités y et z pour que ce soit les 2-ème et 3-ème chasseurs qui ont raté la cible,

$$y = \frac{3}{9 + 3 + 1} = \frac{3}{13} \text{ et } z = \frac{1}{9 + 3 + 1} = \frac{1}{13} .$$

SOLUTION (Exercice 1.4.24).

Soit X_1 et X_2 les positions des 2 points sur $[0, 1]$. La donnée est

$$P(X_1 \in [a_1, b_1], X_2 \in [a_2, b_2]) = (b_1 - a_1)(b_2 - a_2)$$

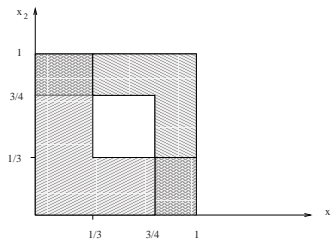
(chaque point est réparti uniformément sur le segment et ils sont tirés indépendamment l'un de l'autre). On doit calculer

$$x = P\left(\sup(X_1, X_2) \geq \frac{3}{4} \mid \inf(X_1, X_2) \leq \frac{1}{2}\right).$$

D'après la définition de Bayes,

$$x = P\left(\sup(X_1, X_2) \geq \frac{3}{4}, \inf(X_1, X_2) \leq \frac{1}{3}\right) / P\left(\inf(X_1, X_2) \leq \frac{1}{3}\right).$$

Appelons A et B les événements $\{\sup(X_1, X_2) \geq \frac{3}{4}\}$ et $\{\inf(X_1, X_2) \leq \frac{1}{3}\}$. (B a la forme du L dans la figure ci-dessous, et $A \cap B$ est la partie la plus sombre.)



Le calcul se réduit à un calcul d'aires : $P(A \cap B) = 2(\frac{1}{4} \times \frac{1}{3})$ et $P(B) = 2 \times \frac{1}{3} - \frac{1}{9}$ d'où $x = \frac{3}{10}$.

SOLUTION (Exercice 1.4.25).

Soit M l'événement, "le patient est atteint", et T_+ l'événement "le test est positif". On cherche à calculer $P(M|T_+)$. Les données sont

$$P(M) = 0.001, P(T_+|M) = 0.99, P(T_+|\overline{M}) = 0.02.$$

D'après la règle de rétrodictio de Bayes,

$$P(M|T_+) = \frac{P(T_+|M)P(M)}{P(T_+)}.$$

La règle des causes totales nous donne d'autre part

$$P(T_+) = P(T_+|M)P(M) + P(T_+|\overline{M})P(\overline{M}),$$

d'où

$$P(M|T_+) = \frac{P(T_+|M)P(M)}{P(T_+|M)P(M) + P(T_+|\overline{M})P(\overline{M})}.$$

Le résultat numérique est

$$P(M|T_+) = \frac{(0.99)(0.001)}{(0.99)(0.001) + (0.02)(0.999)} \sim \frac{1}{21} ,$$

ce qui est très faible ! On voit qu'il faut un $P(T_+|\overline{M})$ plus faible. Par exemple si $P(T_+|\overline{M}) = 0.002$ on trouve $P(M|T_+) = \frac{1}{3}$, et si $P(T_+|\overline{M}) = 0.0002$ on trouve $P(M|T_+) = \frac{99}{101}$.

Dans la pratique des dispensaires, plutôt que de faire un test avec faible $P(T_+|\overline{M})$ qui risque de coûter cher, on refait passer un test plus sûr aux patients ayant eu un premier test positif. L'essentiel est que le premier test ne laisse pas passer un malade atteint. On doit donc avoir un $P(T_+|M)$ très proche de 1. La situation décrite est celle de certains tests de dépistage effectués sur les recrues de l'armée. On groupe les échantillons, ce qui augmente évidemment la probabilité de fausse alarme.

Solutions des exercices du chapitre 2

SOLUTION (Exercice 2.4.1).

De l'inégalité $|X| \leq 1 + X^2$ et des propriétés de monotonie et de linéarité de l'espérance, on tire

$$\begin{aligned} E[|X|] &\leq E[1 + X^2] \\ &= E[1] + E[X^2] \\ &= 1 + E[X^2] < \infty \end{aligned}$$

SOLUTION (Exercice 2.4.3).

$(X - m_X)^2 = X^2 - 2m_X X + m_X^2$. Donc (linéarité de l'espérance) :

$$\begin{aligned} E[(X - m_X)^2] &= E[X^2] - 2m_X E[X] + m_X^2 \\ &= E[X^2] - m_X^2. \end{aligned}$$

SOLUTION (Exercice 2.4.4).

$$\begin{aligned} m_X &= \sum_{n=0}^{\infty} n e^{-\lambda} \frac{\lambda^n}{n!} \\ &= \lambda e^{-\lambda} \sum_{n=0}^{\infty} n \frac{\lambda^{n-1}}{n!} \\ &= \lambda e^{-\lambda} \frac{de^{\lambda}}{d\lambda} = \lambda. \end{aligned}$$

$$\begin{aligned} E[X^2] &= \sum_{n=0}^{\infty} n^2 e^{-\lambda} \frac{\lambda^n}{n!} \\ &= \lambda^2 e^{-\lambda} \sum_{n=2}^{\infty} \frac{n(n-1)\lambda^{n-2}}{n!} + \lambda e^{-\lambda} \sum_{n=1}^{\infty} \frac{n\lambda^{n-1}}{n!} \\ &= \lambda^2 e^{-\lambda} \sum_{n=2}^{\infty} \frac{\lambda^{n-2}}{(n-2)!} + \lambda e^{-\lambda} \sum_{n=1}^{\infty} \frac{\lambda^{n-1}}{(n-1)!} \\ &= \lambda^2 + \lambda. \end{aligned}$$

$$\begin{aligned}\sigma_X^2 &= E[X^2] - m_X^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda.\end{aligned}$$

SOLUTION (Exercice 2.4.5).

$$\begin{aligned}m_X &= \sum_{k=0}^n kP(X=k) \\ &= \sum_{k=1}^n kC_k^n p^k q^{n-k}.\end{aligned}$$

Puisque $kC_k^n = nC_{k-1}^{n-1}$ pour $k \leq n$, on a

$$\begin{aligned}m_X &= \sum_{k=1}^n nC_{k-1}^{n-1} \left(\frac{p}{q}\right)^{k-1} pq^{n-1} \\ &= npq^{n-1} \sum_{\ell=0}^{n-1} C_\ell^{n-1} \left(\frac{p}{q}\right)^\ell \\ &= npq^{n-1} \left(1 + \frac{p}{q}\right)^{n-1} = npq^{n-1} \left(\frac{p+q}{q}\right)^{n-1} = np\end{aligned}$$

$$\begin{aligned}E[X^2] &= \sum_{k=0}^n k^2 P(X=k) \\ &= \sum_{k=1}^n (k(k-1) + k) C_k^n p^k q^{n-k}.\end{aligned}$$

Puisque $k(k-1)C_k^n = n(n-1)C_{k-2}^{n-2}$ pour $2 \leq k \leq n$, on a

$$\begin{aligned}\sum_{k=1}^n k(k-1)C_k^n p^k q^{n-k} &= \sum_{k=2}^n n(n-1)C_{k-2}^{n-2} \left(\frac{p}{q}\right)^{k-2} p^2 q^{n-2} \\ &= n(n-1)p^2 q^{n-2} \sum_{\ell=0}^{n-2} C_\ell^{n-2} \left(\frac{p}{q}\right)^\ell \\ &= n(n-1)p^2 q^{n-2} \left(1 + \frac{p}{q}\right)^{n-2} = n(n-1)p^2.\end{aligned}$$

On a donc

$$\begin{aligned}E[X^2] &= n(n-1)p^2 + \sum_{k=1}^n kC_k^n p^k q^{n-k} = n(n-1)p^2 + m_X \\ &= n(n-1)p^2 + np.\end{aligned}$$

Finalement

$$\sigma_X^2 = E[X^2] - m_X^2 = n(n-1)p^2 + np - n^2p^2 = npq .$$

SOLUTION (Exercice 2.4.7).

$$\begin{aligned} P(T \geq n) &= P(T = n) + P(T = n+1) + \dots \\ &= pq^{n-1}(1 + q + \dots) \\ &= \frac{pq^{n-1}}{1-q} = q^{n-1} , \end{aligned}$$

d'où

$$\begin{aligned} P(T \geq n + n_0 | T > n_0) &= \frac{P(T \geq n + n_0, T \geq n_0 - 1)}{P(T \geq n_0 - 1)} \\ &= \frac{P(T \geq n + n_0)}{P(T \geq n_0 - 1)} \\ &= \frac{q^{n+n_0-1}}{q^{n_0-2}} = q^{n-1} . \end{aligned}$$

SOLUTION (Exercice 2.4.8).

$P(X = n) = P(X \geq n) - P(X \geq n+1)$, d'où

$$\begin{aligned} E[X] &= \sum_{n=0}^{\infty} nP(X = n) \\ &= \sum_{n=0}^{\infty} (nP(X \geq n) - nP(X \geq n+1)) \\ &= \sum_{n=0}^{\infty} (nP(X \geq n) - (n+1)P(X \geq n+1) + P(X \geq n+1)) \\ &= \sum_{n=0}^{\infty} P(X \geq n+1) = \sum_{n=1}^{\infty} P(X \geq n) . \end{aligned}$$

SOLUTION (Exercice 2.4.10).

Soit X cette variable. On a

$$\begin{aligned} g_X(s) &= \sum_{n \geq 0} s^n e^{-\lambda} \frac{\lambda^n}{n!} \\ &= e^{-\lambda} \sum_{n \geq 0} \frac{(\lambda s)^n}{n!} \\ &= e^{-\lambda} e^{\lambda s} = e^{\lambda(s-1)}. \end{aligned}$$

On applique les formules (2.17) et (2.18) : $g'_X(s) = \lambda e^{\lambda(s-1)}$ et donc $E[X] = g'_X(1) = \lambda$; d'autre part $g''_X(s) = \lambda^2 e^{\lambda(s-1)}$ et donc $g''_X(1) = \lambda^2$. La formule $\sigma_X^2 = g''_X(1) + g'_X(1) - g'_X(1)^2$ donne $\sigma_X^2 = \lambda$.

SOLUTION (Exercice 2.4.11).

Soit $S_n = X_1 + \dots + X_n$ où les X_i sont IID, de distribution de probabilité commune donnée par $P(X_i = 1) = 1 - P(X_i = 0) = p$. C'est une variable aléatoire binomiale de paramètres n et p . On a $g_{S_n}(s) = \prod_{i=1}^n g_{X_i}(s) = (ps + q)^n$, et donc

$$E[S_n] = g'_{S_n}(1) = np,$$

et

$$E[S_n^2] = g''_{S_n}(1) + g'_{S_n}(1) = n(n-1)p + np,$$

d'où

$$\begin{aligned} \text{Var}(S_n) &= E[S_n^2] - E[S_n]^2 \\ &= n(n-1)p + np - n^2p^2 \\ &= np(1-p) = npq. \end{aligned}$$

SOLUTION (Exercice 2.4.12).

$g_T(s) = \sum_{n=1}^{\infty} q^{n-1} p s^n = ps \sum_{n=1}^{\infty} q^{n-1} s^{n-1} = ps \sum_{n=0}^{\infty} (qs)^n = (qs)^n$. D'où

$$g_T(s) = \frac{ps}{1-qs}.$$

Les dérivées première et seconde de $g_T(s)$ sont

$$g'_T(s) = \frac{p}{(1-qs)^2} \text{ et } g''_T(s) = \frac{2qp}{(1-qs)^3},$$

d'où,

$$\begin{aligned} E[T] &= g'_T(1) \\ &= \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}, \end{aligned}$$

et

$$\begin{aligned} E[T^2] &= g_T''(1) + g_T'(1) \\ &= \frac{2q}{p^2} + \frac{1}{p} , \end{aligned}$$

et finalement

$$\sigma_T^2 = E[T^2] - E[T]^2 = \frac{q}{p^2} .$$

SOLUTION (Exercice 2.4.13).

On utilise la formule (3.1) du Théorème 2.3.2. Ici, $g_X(s) = ps + q$ et $g_T(s) = e^{\lambda(s-1)}$ d'où $g_S(s) = e^{\lambda p(s-1)}$. S est donc une variable aléatoire de Poisson de moyenne λp .

SOLUTION (Exercice 2.4.14).

$E[S] = g_S'(1)$. Mais en utilisant le résultat du Théorème 2.3.3 : $g_S'(s) = g_T'(g_X(s))g_X'(s)$ et donc, puisque $g_X(1) = 1$,

$$g_S'(1) = g_T'(1)g_X'(1) ,$$

d'où le résultat puisque $E[T] = g_T'(1)$ et $E[X] = g_X'(1)$.

SOLUTION (Exercice 2.4.15).

On a pour $n \geq 1$,

$$m_n = \varphi_n'(1) = g_Z'(\varphi_{n-1}(1))\varphi_{n-1}'(1) = g_Z'(1)\varphi_{n-1}'(1) = m_Z m_{n-1} ,$$

et donc, puisque $m_0 = 1$,

$$m_n = m_Z^n ,$$

pour $n \geq 0$. D'autre part

$$v_n = \varphi_n''(1) + m_n - m_n^2 \text{ et } \sigma_Z^2 = g_Z''(1) + m_Z - m_Z^2 ,$$

et donc, puisque

$$\begin{aligned} \varphi_n''(1) &= g''(\varphi_{n-1}(1))\varphi_{n-1}'(1)^2 + g'(\varphi_{n-1}(1))\varphi_{n-1}''(1) \\ &= g''(1)\varphi_{n-1}'(1)^2 + g'(1)\varphi_{n-1}''(1) , \end{aligned}$$

on a

$$v_n = \sigma_Z^2 m_Z^{2n-2} + m_Z v_{n-1} \quad (n \geq 1) .$$

En tenant compte de la condition initiale $v_0 = 0$, on a donc pour $n \geq 0$,

$$v_n = \begin{cases} \sigma_Z^2 m_Z^{n-1} \frac{1-m_Z^2}{1-m_Z} & \text{si } m_Z \neq 0 \\ n \sigma_Z^2 & \text{si } m_Z = 1 . \end{cases}$$

Si $X_0 = k$, on a k processus de branchement indépendants, d'où :

$$m_n = km_Z^n,$$

et

$$v_n = \begin{cases} k^2 \sigma_Z^2 m_Z^{n-1} \frac{1-m_Z^n}{1-m_Z} & \text{si } m \neq 0 \\ k^2 n \sigma_Z^2 & \text{si } m = 1. \end{cases}$$

SOLUTION (Exercice 2.4.16).

$$\begin{aligned} P(p(X) > 0) &= E [1_{\{p(X) > 0\}}] \\ &= \sum_{x \in \mathcal{X}} p(x) 1_{\{p(x) > 0\}} \\ &= \sum_{x \in \mathcal{X}} p(x) = 1. \end{aligned}$$

SOLUTION (Exercice 2.4.17).

Considérons les variables indépendantes Z_1, Z_2, \dots où $Z_i = 1$ si et seulement si le i -ème insecte capturé est une femelle, $Z_i = 0$ autrement. Lorsque $k < M$, on a évidemment $P(X = k) = 0$. Lorsque $k \geq M$,

$$\begin{aligned} P(X = k) &= P(Z_1 + \dots + Z_{k-1} = M - 1, Z_k = 1) \\ &= P(Z_1 + \dots + Z_{k-1} = M - 1) \times P(Z_k = 1) \\ &= \frac{(k-1)!}{(k-M)!(M-1)!} \theta^{M-1} (1-\theta)^{k-M} \times \theta \\ &= \frac{(k-1)!}{(k-M)!(M-1)!} \theta^M (1-\theta)^{k-M}. \end{aligned}$$

où on a utilisé le fait que $Z_1 + \dots + Z_{k-1}$ est une variable binomiale.

Solutions des exercices du chapitre 3

SOLUTION (Exercice 3.5.1).

D'après le théorème de Fubini,

$$\left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right)^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dx dy .$$

En passant des coordonnées cartésiennes (x, y) aux coordonnées polaires (ρ, θ) :

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dx dy = \int_0^{2\pi} \int_0^{\infty} e^{-\frac{\rho^2}{2}} \rho d\rho d\theta .$$

En utilisant à nouveau le théorème de Fubini, on trouve pour le dernier membre de l'égalité précédente

$$\int_0^{2\pi} \int_0^{\infty} e^{-\frac{\rho^2}{2}} \rho d\rho d\theta = 2\pi \int_0^{\infty} e^{-\frac{\rho^2}{2}} \rho d\rho = 2\pi \int_0^{\infty} e^{-v} dv = 2\pi .$$

(i) En faisant le changement $y = \frac{x-m}{\sigma}$, on trouve

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}y^2} dy = 1 .$$

(ii) Le changement $z = x - m$ donne

$$\begin{aligned} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (z+m) e^{-\frac{1}{2}\frac{z^2}{\sigma^2}} dz \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} z e^{-\frac{1}{2}\frac{z^2}{\sigma^2}} dz + m \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\frac{z^2}{\sigma^2}} dz . \end{aligned}$$

L'avant dernière intégrale est nulle pour des raisons de symétrie ; quant à la dernière, elle est égale à m , d'après (i).

(iii) Le changement $y = \frac{x-m}{\sigma}$ donne

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-m)^2 e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx = \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y^2 e^{-\frac{1}{2}y^2} dy .$$

Au vu de l'égalité

$$\frac{d}{dy}(y e^{-\frac{1}{2}y^2}) = -y^2 e^{-\frac{1}{2}y^2} + e^{-\frac{1}{2}y^2} ,$$

$$\text{on a } \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y^2 e^{-\frac{1}{2}y^2} dy$$

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}y^2} dy - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{d}{dy}(y e^{-\frac{1}{2}y^2}) dy \\ &= 1 - (y e^{-\frac{1}{2}y^2})_{-\infty}^{+\infty} = 1 , \end{aligned}$$

d'où le résultat annoncé.

(iv) Il suffit de prouver que la fonction caractéristique d'une gaussienne standard est $e^{-\frac{1}{2}u^2}$, c'est-à-dire

$$e^{-\frac{1}{2}u^2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{iux - \frac{1}{2}x^2} dx ,$$

ou encore :

$$1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\frac{1}{2}(u+ix)^2} dx ,$$

ce qu'on on laisse au lecteur le soin de démontrer (faire un calcul d'intégrale dans le plan complexe, en utilisant la méthode des résidus).

SOLUTION (Exercice 3.5.2).

(i) découle de la définition de $\Gamma(\alpha)$, puisque ($y \rightarrow bx$)

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx = \beta^\alpha \int_0^\infty y^{\alpha-1} e^{-\beta y} dy .$$

(ii)

$$\begin{aligned} m_X &= \int_0^\infty x \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{(\alpha-1)-1} e^{-\beta x} dx \\ &= \frac{1}{\beta} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \frac{\alpha}{\beta} . \end{aligned}$$

(iii)

$$E[X^2] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{(\alpha+2)-1} e^{-\beta x} dx = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} \frac{1}{\beta^2} .$$

Or

$$\Gamma(\alpha+2) = (\alpha+1)\alpha\Gamma(\alpha)$$

et donc

$$E[X^2] = \frac{(\alpha+1)\alpha}{\beta} = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} ,$$

d'où

$$\sigma_X^2 = E[X^2] - E[X]^2 = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}$$

(iv)

$$\frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{(iu-\beta)x} x^{\alpha-1} dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(iu-\beta)^\alpha} = \frac{1}{(1-\frac{iu}{\beta})^\alpha} .$$

(Cette dernière égalité se justifie grâce à une intégration dans le plan complexe.)

SOLUTION (Exercice 3.5.3).

Les fonctions caractéristiques de X_1 et X_2 sont respectivement

$$\varphi_{X_1}(u) = \left(1 - \frac{iu}{\beta}\right)^{-\alpha_1} , \quad \varphi_{X_2}(u) = \left(1 - \frac{iu}{\beta}\right)^{-\alpha_2} .$$

Comme X_1 et X_2 sont indépendantes, la fonction caractéristique $\varphi_{X_1+X_2}(u)$ de $X_1 + X_2$ est donnée par le produit $\varphi_{X_1}(u)\varphi_{X_2}(u)$:

$$\varphi_{X_1+X_2}(u) = \left(1 - \frac{iu}{\beta}\right)^{-(\alpha_1+\alpha_2)},$$

ce qui montre que $X_1 + X_2$ suit une loi gamma de paramètre $\alpha_1 + \alpha_2$ et β .

SOLUTION (Exercice 3.5.4).

On utilise la partition $\Omega = \sum_{i=1}^{\infty} \{Y = y_i\}$ pour écrire :

$$\{Z \leq x\} = \sum_{i=1}^{\infty} (\{Z \leq z\} \cap \{Y = y_i\})$$

Mais si $Y = y_i$, $Z = X + y_i$ et donc

$$\{Z \leq z\} = \sum_{i=1}^{\infty} (\{X + y_i \leq z\} \cap \{Y = y_i\}) .$$

D'où

$$P(Z \leq z) = \sum_{i=1}^{\infty} P(X \leq Z - y_i, Y = y_i) .$$

Mais X et Y sont indépendantes, et donc

$$P(Z \leq z) = \sum_{i=1}^{\infty} P(X \leq Z - y_i) P(Y = y_i) = \sum_{i=1}^{\infty} p_i \int_{-\infty}^{z-x_i} f_X(x) dx .$$

Finalement

$$P(Z \leq z) = \int_{-\infty}^z \sum_{i=1}^{\infty} p_i f_X(x - y_i) dx ,$$

ce qui prouve que Z admet la densité de probabilité

$$f_Z(z) = \sum_{i=1}^{\infty} p_i f_X(z - y_i) .$$

SOLUTION (Exercice 3.5.5).

On sait (Exemple 3.1.6) que si X_i est une gaussienne standard, alors X_i^2 suit une loi du χ^2 à 1 degré de liberté, dont la fonction caractéristique est

$$\varphi_{X_i^2}(u) = (1 - 2iu)^{-\frac{1}{2}} .$$

Comme les X_i^2 sont indépendantes entre elles,

$$\varphi_{X_1^2} + \dots + \varphi_{X_n^2}(u) = \prod_{i=1}^n \varphi_{X_i^2}(u) = (1 - 2iu)^{-\frac{n}{2}},$$

ce qui est la fonction caractéristique d'une loi du χ^2 à n degrés de liberté.

SOLUTION (Exercice 3.5.6).

$$P(X \geq x + y | X \geq y) = \frac{P(X \geq x + y, X \geq y)}{P(X \geq y)} = \frac{P(X \geq x + y)}{P(X \geq y)}.$$

Or $P(X \geq x) = e^{-\lambda x}$. D'où $P(X \geq x + y | X \geq y) = e^{-\lambda x} = P(X \geq x)$.

SOLUTION (Exercice 3.5.9).

a.

$$P(\max(X_1, \dots, X_n) \leq y) = P(X_1 \leq y, \dots, X_n \leq y) = P(X_1 \leq y) \dots P(X_n \leq y) = F(y)^n$$

$$\begin{aligned} P(\min(X_1, \dots, X_n) \leq z) &= 1 - P(\min(X_1, \dots, X_n) > z) \\ &= 1 - P(X_1 > z, \dots, X_n > z) \\ &= 1 - P(X_1 > z) \dots P(X_n > z) = 1 - (1 - F(z))^n. \end{aligned}$$

b.

$$P(Y \leq y) = F(y)^n = (1 - e^{-\lambda y})^n 1_{\{y > 0\}}$$

et donc

$$\begin{aligned} f_Y(y) &= n\lambda e^{-\lambda y} (1 - e^{-\lambda y})^{n-1} 1_{\{y > 0\}} \\ P(Z \leq z) &= 1 - (1 - F(z))^n = 1 - e^{-n\lambda z} 1_{\{z > 0\}} \end{aligned}$$

et donc

$$f_Z(z) = n\lambda e^{-n\lambda z} 1_{\{z > 0\}},$$

c'est-à-dire

$$Z = \min(X_1, \dots, X_n) \sim \mathcal{E}(n\lambda).$$

SOLUTION (Exercice 3.5.15).

(i) découle de la croissance de la fonction P : en effet si $x_1 \leq x_2$, alors $\{X \leq x_1\} \subseteq \{X \leq x_2\}$.

(ii) il suffit de prouver que $\lim_{n \rightarrow \infty} F_x(-n) = 0$. Pour cela on appliquera le théorème de continuité séquentielle en remarquant que $P(\phi) = 0, F_X(-n) = P(X \leq -n)$ et $\{X \leq -N\} \downarrow \phi$.

(iii) se démontre de façon analogue à (ii). En effet, $1 = P(\Omega)$, $F_X(+n) = P(X \leq n)$ et $\{X \leq n\} \uparrow \Omega$ d'où d'après le théorème de continuité séquentielle, $\lim \uparrow F_X(n) = 1$.

(iv) il suffit de prouver que pour tout $x \in \mathbb{R}$, $\lim F_X(X + \frac{1}{n}) = F_X(X)$. Cela découle encore du théorème de continuité séquentielle associé aux observations suivantes :

$$F_X(x) = P(X \leq x)$$

$$F_X\left(x + \frac{1}{n}\right) = P\left(X \leq x + \frac{1}{n}\right)$$

et

$$\cap_{n=1}^{\infty} \left\{ X \leq x + \frac{1}{n} \right\} = \{X \leq x\}.$$

SOLUTION (Exercice 3.5.16).

D'après le Théorème 3.1.9, la densité de probabilité du vecteur $Z = (Z_1, \dots, Z_n)$ est

$$f_Z(z_1, \dots, z_n) = \frac{1}{n!} 1_{[0,1]^n \cap C}(z_1, \dots, z_n)$$

où $C = \{x_1 < x_2 < \dots < x_n\}$. La densité de Z_i est obtenue en intégrant f_Z par rapport aux arguments $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$:

$$\begin{aligned} f_{Z_i}(z) &= n! \underbrace{\int_0^1 \dots \int_0^1}_{n-1} 1_{(z_1 < \dots < z_{i-1} < z < z_{i+1} < \dots < z_n)} dz_1 \dots dz_{i-1} dz_{i+1} \dots dz_n \\ &= n! \underbrace{\int_0^1 \dots \int_0^1}_{i-1} 1_{(z_1 < \dots < z_{i-1} < z)} dz_1 \dots dz_{i-1} \times \underbrace{\int_0^1 \dots \int_0^1}_{n-i} 1_{(z < z_{i+1} < \dots < z_n)} dz_{i+1} \dots dz_n, \end{aligned}$$

c'est-à-dire, en utilisant le résultat de l'Exemple 3.1.10,

$$f_{Z_i}(z) = n! \frac{z^{i-1}}{(i-1)!} \frac{(1-z)^{n-i}}{(n-i)!} = n \binom{n-1}{i-1} z^{i-1} (1-z)^{n-i}$$

SOLUTION (Exercice 3.5.17).

$$E[X] = \frac{1}{2\pi} \int_0^{2\pi} \sin z dz = 0, \quad E[Y] = \frac{1}{2\pi} \int_0^{2\pi} \cos z dz = 0.$$

D'où $E[(X - m_X)(Y - m_Y)] = E[XY] = E[\cos z \sin z] = \frac{1}{2} E[\sin 2z] = \frac{1}{4\pi} \int_0^{2\pi} \sin(2z) dz = 0$ et X et Y admettent toutes deux des densités comme on le voit facilement. Soit $f_X(x)$ et $f_Y(y)$ ces densités. Si X et Y étaient indépendantes, (X, Y) admettrait une densité

$$f_{XY}(x, y) = f_X(x) f_Y(y).$$

Or cela n'est pas possible car, d'une part,

$$P(X^2 + Y^2 = 1) = 1 ,$$

et d'autre part, quelle que soit la densité $f_{XY}(x,y)$,

$$\int \int_{\{x^2+y^2=1\}} f_{XY}(x,y) dx dy = 0 .$$

SOLUTION (Exercice 3.5.19).

Le ρ qui intervient dans l'expression de la densité est le coefficient de corrélation. On a donc l'équation de la droite de régression (D) de X_2 par rapport à X_1 :

$$\frac{x_2}{\sigma_2} = \rho \frac{x_1}{\sigma_1} .$$

L'intersection de cette droite avec l'ellipse d'équidensité est donnée par

$$\frac{x_1}{\sigma_1} = \frac{\lambda}{1-\rho} \quad , \quad \frac{x_2}{\sigma_2} = \frac{\lambda\rho}{1-\rho^2} .$$

En ce point-là, la tangente à l'ellipse d'équidensité est parallèle à $0x_1$ (ce qui prouve que (D) est l'axe conjugué de $0x$). En effet, pour ce point,

$$\frac{dQ}{dx_1}(x_1, x_2).1 + \frac{dQ}{dx_2}(x_1, x_2).0 = \frac{dQ}{dx_1}(x_1, x_2) = \frac{2x_1}{\sigma_1^2} - 2\rho \frac{x_2}{\sigma_1\sigma_2} = 0 .$$

SOLUTION (Exercice 3.5.20).

$$\begin{aligned} E[(\hat{Y} - Y)(\hat{Y} - Y)^T] &= E[\Sigma_{YX}\Sigma_X^{-1}(X - m_X) - (Y - m_Y)](\Sigma_{YX}\Sigma_X^{-1}(X - m_X) \\ &\quad - (Y - m_Y))^T] \\ &= \Sigma_{YX}.\Sigma_X^{-1}E[(X - m_X)(X - m_X)^T](\Sigma_X^{-1})^T\Sigma_{YX}^T \\ &\quad + E[(Y - m_Y)(Y - m_Y)^T] \\ &\quad - \Sigma_{YX}\Sigma_X^{-1}E[(X - m_X) - (Y - m_Y)^T] \\ &\quad - E[(Y - m_Y)(X - m_X)^T](\Sigma_X^{-1})' . \Sigma_{YX}^T \\ &= \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} + \Sigma_Y - \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} - \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} , \end{aligned}$$

où on a utilisé les égalités matricielles $\Sigma_X = \Sigma_X^T$, $(\Sigma_X^{-1})^T = \Sigma_X^{-1}$, $\Sigma_{YX}^T = \Sigma_{XY}$. Finalement :

$$P_{Y|X} = \Sigma_Y - \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} .$$

Commentaire : Si Σ_Y est inversible, la dernière expression prend la forme

$$P_{X|Y} = \Sigma_Y (I - (\Sigma_Y^{-1} \Sigma_{YX})(\Sigma_X^{-1} \Sigma_{XY})) .$$

Dans le cas où Y et X sont scalaires, on a donc :

$$E[|\hat{Y} - Y^2|] = \sigma_Y^2 (1 - \rho_{XY}^2) .$$

C'est donc la matrice $(\Sigma_Y^{-1} \Sigma_{YX})(\Sigma_X^{-1} \Sigma_{XY})$ qui joue le rôle du carré du coefficient de corrélation.

SOLUTION (Exercice 3.5.21).

(1) : On a :

$$\begin{aligned} \Sigma_{SX} &= \Sigma_{SS} + \Sigma_{SB} = \Sigma_{SS} + 0 = \Sigma_{SS} = \Sigma_S \\ \Sigma_{XX} &= \Sigma_{SS} + \Sigma_{SB} + \Sigma_{BS} + \Sigma_{BB} = \Sigma_{SS} + \Sigma_{BB} = \Sigma_S + \Sigma_B . \end{aligned}$$

D'où

$$\hat{S} = m_S + \Sigma_S (\Sigma_S + \Sigma_B)^{-1} (X - m_X) .$$

(2) : D'après le résultat de l'Exercice 3.5.20,

$$\begin{aligned} E[(\hat{S} - S)(\hat{S} - S)^T] &= \Sigma_S - \Sigma_S (\Sigma_S + \Sigma_B)^{-1} \Sigma_S = \Sigma_S (I - \Sigma_S (\Sigma_S + \Sigma_B)^{-1}) \\ &= \Sigma_S (\Sigma_S + \Sigma_B - \Sigma_S) (\Sigma_S + \Sigma_B)^{-1} = \Sigma_S \Sigma_B (\Sigma_S + \Sigma_B)^{-1} \\ &= \Sigma_B \Sigma_S (\Sigma_S + \Sigma_B)^{-1} \end{aligned}$$

Commentaire : dans le cas unidimensionnel,

$$E[|\hat{S} - S|^2] = \sigma_B^2 \frac{\sigma_S^2}{\sigma_S^2 + \sigma_B^2} .$$

Si on définit le rapport "signal sur bruit" S/B par

$$S/B = \frac{\sigma_S^2}{\sigma_B^2} ,$$

on a

$$E(|\hat{S} - S|^2) = \sigma_B^2 \frac{(S/B)}{1 + (S/B)} .$$

L'effet du bruit (σ_B^2) a donc été réduit dans le facteur $(S/B)/(1 + (S/B))$.

SOLUTION (Exercice 3.5.22).

Il faut prouver

$$E[(X + Y)^2] \leq E[X^2] + E[Y^2] + 2E[X^2]^{1/2} E[Y^2]^{1/2} ,$$

c'est-à-dire :

$$2E[XY] \leq 2E[X^2]^{1/2} E[Y^2]^{1/2} ,$$

ce qui n'est autre que l'inégalité de Schwarz.

SOLUTION (Exercice 3.5.23).

$$\begin{aligned} \phi_X(u) &= \int_{-\infty}^{+\infty} e^{iux} \frac{a}{2} e^{-a|x|} dx = \frac{a}{2} \left[\int_0^{+\infty} e^{(iu-a)x} dx + \int_0^{+\infty} e^{(-iu-a)x} dx \right] \\ &= \frac{a}{z} \left[-\frac{1}{iu-a} + \frac{1}{iu+a} \right] = \frac{a^2}{a^2 + u^2} . \end{aligned}$$

La densité d'une variable de Cauchy Y est

$$f_Y(y) = \frac{1}{\pi} \frac{1}{1+y^2} .$$

D'après la formule d'inversion de Fourier,

$$\frac{a}{2} e^{-a|x|} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-iux} \frac{a^2}{a^2 + u^2} du .$$

Cette dernière égalité s'écrit donc, pour $a = 1$,

$$e^{-|x|} = \int_{-\infty}^{+\infty} e^{+iux} \frac{1}{\pi} \frac{1}{1+u^2} du .$$

D'où

$$\phi_Y(u) = e^{-|u|} .$$

SOLUTION (Exercice 3.5.25).

Le vecteur (X, Y) étant obtenu comme combinaison linéaire du vecteur gaussien (X_1, X_2) est donc un vecteur gaussien. On a

$$\begin{aligned} V(X) &= V\left(\frac{X_1 - X_2}{2}\right) = \frac{1}{2} V(X_1 - X_2) = \frac{1}{2} (V(X_1) + V(X_2)) \\ &= \frac{1}{2} (V(X_1) + V(X_2)) = \frac{1}{2} (1 + 1) = 1 \end{aligned}$$

(où l'on a utilisé le fait que X_1 et X_2 sont indépendantes). De la même manière, $V(Y) = 1$. Aussi :

$$\text{cov}(X, Y) = \frac{1}{2} E[X_1^2 - X_2^2] = \frac{1}{2} (E[X_1^2] - E[X_2^2]) = 0 .$$

Les variables X et Y sont donc non-corrélées. Comme (X, Y) est un vecteur gaussien, cela entraîne que X et Y sont indépendantes. En résumé, X et Y sont deux gaussiennes réduites indépendantes. Calculons la fonction caractéristique de Z :

$$\begin{aligned}\phi_Z(u) &= E[e^{iuZ}] = E[e^{iuXY}] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{iuxy} f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} e^{iuxy} f_X(x) dx \right) f_Y(y) dy = \int_{-\infty}^{+\infty} \phi_X(uy) f_Y(y) dy .\end{aligned}$$

Mais comme X est une gaussienne réduite,

$$\phi_X(uy) = e^{-\frac{1}{2}u^2y^2} ,$$

d'où :

$$\phi_Z(u) = \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2y^2} f_Y(y) dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}y^2(u^2+1)} dy .$$

En faisant le changement de variable $v = y\sqrt{u^2+1}$ et en utilisant la formule $\int_{-\infty}^{+\infty} e^{-\frac{1}{2}v^2} dv = \sqrt{2\pi}$ on trouve :

$$\phi_Z(u) = \frac{1}{\sqrt{1+u^2}} .$$

SOLUTION (Exercice 3.5.26).

(1) En utilisant le fait que X et $-X$ ont la même distribution :

$$\begin{aligned}P(Y_a \leq x) &= P(Y_a \leq x, |X| < a) + P(Y_a \leq x, |X| \geq a) \\ &= P(X \leq x, |X| < a) + P(-X \leq x, |X| \geq a) \\ &= P(X \leq x, |X| < a) + P(X \leq x, |X| \geq a) = P(X \leq x) .\end{aligned}$$

(2) S'il en était ainsi, $X + Y_a$ serait une gaussienne. Or $X + Y_a = 2X1_{\{|X| < a\}}$, et donc $P(X + Y_a = 0) = P(|X| \geq a) \in (0, 1)$, ce qui n'est pas possible, même pour une gaussienne au sens de la définition 3.4.1.

(3)

$$\begin{aligned}\text{cov}(X, Y_a) &= E[X^2 1_{\{|X| < a\}} - X^2 1_{\{|X| \geq a\}}] \\ &= \frac{1}{\sqrt{2\pi}} \int_0^a t^2 e^{-\frac{t^2}{2}} dt - \frac{1}{\sqrt{2\pi}} \int_a^\infty t^2 e^{-\frac{t^2}{2}} dt = \frac{1}{4} - \frac{1}{4} = 0\end{aligned}$$

Solutions des exercices du chapitre 4

SOLUTION (Exercice 4.4.1).

Résultat :

$$f_{X|Y}(x|y) = \frac{1}{2\sqrt{1-y^2}} 1_{\{|x| \leq \sqrt{1-y^2}, |y| \leq 1\}}.$$

SOLUTION (Exercice 4.4.2).

On commence par calculer la densité de probabilité $f_{Y,Z}(y, z)$ du couple (Y, Z) . Tout d'abord la fonction de répartition de (Y, Z) , $F_{Y,Z}(y, z) = P(Y \leq y, Z \leq z) = P(Y \leq y) - P(Y \leq y, Z > z)$. En utilisant l'indépendance des X_i , on obtient

$$P(Y \leq y) = P(X_1 \leq y, \dots, X_n \leq y) = F(y)^n,$$

et

$$P(Y \leq y, Z > z) = P(z < X_1 \leq y, \dots, z < X_n \leq y) = (F(y) - F(z))^n 1_{\{y \geq z\}}$$

D'où :

$$F_{Y,Z}(y, z) = F(y)^n - (F(y) - F(z))^n 1_{\{y \geq z\}}$$

et donc

$$f_{Y,Z}(y, z) = \frac{\partial^2 F_{Y,Z}(y, z)}{\partial z \partial y} = n(n-1)(F(y) - F(z))^{n-2} f(y) f(z) 1_{\{y \geq z\}}$$

La densité conditionnelle $f_{Z|Y}(z|y)$ est donnée par la formule de définition

$$f_{Z|Y}(z|y) = \frac{f_{Y,Z}(y, z)}{f_Y(y)}.$$

Mais d'après un calcul fait quelques lignes plus haut ,

$$F_Y(y) = P(Y \leq y) = F(y)^n,$$

et donc

$$f_Y(y) = \frac{\partial F_Y(y)}{\partial y} = nF(y)^{n-1} f(y),$$

d'où :

$$f_{Z|Y}(z|y) = \frac{(n-1)(F(y) - F(z))^{n-2} f(z)}{F(y)^{n-1}} 1_{\{y \geq z\}}$$

L'espérance conditionnelle $E[Z|Y]$ est de la forme $g(Y)$ où

$$g(y) = \int_{-\infty}^{+\infty} z f_{Z|Y}(z|y) dz = (n-1) \int_{-\infty}^y z \frac{(F(y) - F(z))^{n-2}}{F(y)^{n-1}} f(z) dz.$$

Cas particulier :

$$f_{Z|Y}(z|y) = (n-1) \frac{(y-z)^{n-2}}{y^n} 1_{\{0 \leq z \leq y \leq 1\}}.$$

On trouve finalement

$$E[Z|Y] = \frac{Y}{n}.$$

SOLUTION (Exercice 4.4.3).

Soit X_n la variable aléatoire qui prend la valeur 1 si le soleil se lève le n -ème jour, 0 sinon. Soit Y la variable aléatoire tirée à l'origine des temps et qui représente le biais de la pièce du pile ou face cosmique lancée quotidiennement pour décider si le soleil se lève ou non :

$$P(X_n = 1|Y = p) = p.$$

Lorsque $Y = p$ est donné, $(X_n, n \geq 1)$ est une suite de variables aléatoires indépendantes :

$$P(X_1 = x_1, \dots, X_n = x_n|Y = p) = p^{\sum_{i=1}^n x_i} (1-p)^{\sum_{i=1}^n (1-x_i)}.$$

Les $(X_n, n \geq 1)$ sont *conditionnellement* indépendants étant donné Y , mais ils ne sont pas indépendants, comme on va le voir en calculant $P(X_{N+1} = 1|X_1 = 1, \dots, X_N = 1)$. Si on avait l'indépendance, on trouverait que

$$P(X_{N+1} = 1|X_1 = 1, \dots, X_N = 1) = P(X_{N+1} = 1) = \frac{1}{2}$$

puisque $P(X_n = 1) = \int_0^1 P(X_n = 1|Y = p) f_Y(p) dp = \int_0^1 p dp$ (où l'on a supposé que Y est uniformément distribuée sur $[0,1]$). En réalité

$$\begin{aligned} P(X_{N+1} = 1|X_1 = 1, \dots, X_N = 1) \\ &= \int_0^1 P(X_{N+1} = 1|X_1 = 1, \dots, X_N = 1, Y = p) f_{Y|X_1, \dots, X_N}(p|1, \dots, 1) dp \\ &= \int_0^1 p \frac{p^N}{\int_0^1 p^N dp} dp = \frac{1}{N+2} / \frac{1}{N+1} = \frac{N+1}{N+2}. \end{aligned}$$

SOLUTION (Exercice 4.4.4).

On écrit

$$f_{XY}(x, y) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} x e^{-\frac{1}{2}xy} 1_{\{x \geq 0, y \geq 0\}},$$

et donc (Théorème 4.1.1) :

$$f_X(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} 1_{\{x \geq 0\}}$$

(X est distribuée comme la valeur absolue d'une gaussienne standard) et

$$f_{X|Y}(x|y) = \frac{1}{2} x e^{-\frac{1}{2}xy} 1_{\{y \geq 0\}}$$

(Y est, sachant $X = x$, distribuée selon une loi exponentielle de paramètre $\frac{1}{2}x$).

SOLUTION (Exercice 4.4.5).

Pour $0 \leq k \leq n$,

$$\begin{aligned} P(X_1 = k | X_1 + X_2 = n) &= P(X_1 = k, X_1 + X_2 = n) / P(X_1 + X_2 = n) \\ &= P(X_1 = k, X_1 + X_2 = n - k) / P(X_1 + X_2 = n) \\ &= e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} / e^{-(\lambda_1 + \lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n-k}. \end{aligned}$$

SOLUTION (Exercice 4.4.6).

D'après le Théorème 4.1.8,

$$E[|E[X|Y_1] - X|^2] = \min_v E[|v(Y_1) - X|^2]$$

et

$$E[|E[X|Y_2] - X|^2] = \min_w E[|w(Y_2) - X|^2],$$

où les minimisations portent sur des fonctions v et w telles que $v(Y_1)$ et $w(Y_2)$ soient de carré intégrable. Mais $w(Y_2) = w(g(Y_1))$, ce qui entraîne

$$\min_w E[|w(Y_2) - X|^2] \geq \min_v E[|v(Y_1) - X|^2].$$

L'égalité inverse a lieu pour des raisons analogues et donc

$$E[|E[X|Y_1] - X|^2] = \min_w E[|w(Y_2) - X|^2].$$

Comme $E[X|Y_1]$ est une fonction de Y_1 , c'est aussi une fonction de Y_2 . D'après la partie "unicité" du Théorème 4.1.8, $E[X|Y_1]$ est donc égale à $E[X|Y_2]$.

SOLUTION (Exercice 4.4.7).

On applique le Théorème 4.2.1 avec $Y = (U, V)$. Or,

$$\Sigma_Y = \begin{pmatrix} \Sigma_U & 0 \\ 0 & \Sigma_V \end{pmatrix}$$

et donc

$$\Sigma_Y^{-1} = \begin{pmatrix} \Sigma_U^{-1} & 0 \\ 0 & \Sigma_V^{-1} \end{pmatrix}$$

D'autre part,

$$\Sigma_{XY} = (\Sigma_{XU} \quad \Sigma_{XV}) .$$

on a donc

$$\begin{aligned} E[X|U, V] &= m_X + (\Sigma_{XU} \quad \Sigma_{XV}) \begin{pmatrix} \Sigma_U^{-1} & 0 \\ 0 & \Sigma_V^{-1} \end{pmatrix} \begin{pmatrix} U - m_U \\ V - m_V \end{pmatrix} \\ &= m_X + \Sigma_{XU} \Sigma_U^{-1} (U - m_U) + \Sigma_{XV} \Sigma_V^{-1} (V - m_V) + m_X - m_X \\ &= E[X|U] + E[X|V] - m_X . \end{aligned}$$

SOLUTION (Exercice 4.4.8).

En notant que $a \oplus a = 0$,

$$P(X = x|H_i) = P(B \oplus v_i = x) = P(B = x \oplus v_i) = p^{h(x \oplus v_i)} q^{n-h(x \oplus v_i)}$$

où pour tout $y \in \{0,1\}^n$,

$$h(y) = \sum_{k=1}^n y_k .$$

D'où

$$\log P(X = x|H_i) = (\log p) h(x \oplus v_i) + (\log q) (n - h(x \oplus v_i)) = n(\log q) - h(x \oplus v_i) (\log \frac{q}{p}).$$

Comme $\log \frac{q}{p} > 0$, la stratégie optimale est donc :

$$h(X \oplus v_i) = \min_k \{h(X \oplus v_k)\} \rightarrow H_i$$

D'où le résultat annoncé en remarquant que $h(x \oplus v_i)$ est la distance de Hamming entre x et v_i .

Solutions des exercices du chapitre 5

SOLUTION (Exercice 5.5.1).

$$\begin{aligned} H(Y) &= - \sum_y P(Y = y) \log P(Y = y) \\ &= - \sum_x P(Y = \varphi(x)) \log P(Y = \varphi(x)) \\ &= - \sum_x P(X = x) \log P(X = x). \end{aligned}$$

Pour la deuxième question, on note simplement que la fonction φ définie par $\varphi(z) = (z, \psi(z))$ est une fonction bijective.

SOLUTION (Exercice 5.5.2).

Le code retourné est uniquement déchiffrable car en parcourant une concaténation de mots-code de ce nouveau code de droite à gauche, on peut faire de manière unique les césures permettant de séparer les mots-code. Considérons le code retourné d'un code ayant la propriété du préfixe, le code 3 de l'Exemple 5.3.1. Ce code retourné n'a pas la propriété du préfixe. C'est l'exemple recherché.

SOLUTION (Exercice 5.5.4).

$$\frac{\partial}{\partial p_i} \left(\sum_i p_i \log p_i + \lambda \sum_i p_i E_i \right) = p_i + 1 + \lambda E_i = 0,$$

d'où le résultat

$$p_i = ce^{-\lambda E_i} = \frac{e^{-\lambda E_i}}{\sum_k e^{-\lambda E_k}}.$$

La constante λ est choisie de telle sorte que

$$\frac{\sum_k e^{-\lambda E_k} E_k}{\sum_k e^{-\lambda E_k}} = E.$$

Il faut montrer qu'un tel choix est possible. On pose $E_{\min} = \inf_i E_i$ et $E_{\max} = \sup_i E_i$. En écrivant

$$F(\lambda) = \frac{\sum_k e^{-\lambda E_k} E_k}{\sum_k e^{-\lambda E_k}} = \frac{\sum_k e^{-\lambda(E_k - E_{\min})} E_k}{\sum_k e^{-\lambda(E_k - E_{\min})}},$$

on voit que si on fait tendre λ vers ∞ , seuls subsistent les termes tels que $E_k = E_{\min}$ et donc $F(\infty) = E_{\min}$. De même, en écrivant

$$F(\lambda) = \frac{\sum_k e^{+\lambda(E_k - E_{\max})} E_k}{\sum_k e^{+\lambda(E_k - E_{\max})}},$$

on voit que si on fait tendre λ vers $-\infty$, seuls subsistent les termes tels que $E_k = E_{\max}$ et donc $F(-\infty) = E_{\max}$. Comme $F(\lambda)$ est continue et strictement décroissante, et que $E_{\min} \leq E \leq E_{\max}$, il existe une et une seule solution λ .

SOLUTION (Exercice 5.5.5).

$$\begin{aligned}
 H(X|Y) &= - \sum_{x,y} p_{X,Y}(x,y) (\log p_{X|Y}(x,y)) \\
 &= - \sum_{x,y} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_Y(y)} \\
 &= - \sum_{x,y} p(x,y) \log p(x,y) - \left(- \sum_y \left(\sum_x p(x,y) \right) \log p(y) \right) \\
 &= - \sum_{x,y} p(x,y) \log p(x,y) - \left(- \sum_y p(y) \log p(y) \right) \\
 &= H(X,Y) - H(Y).
 \end{aligned}$$

SOLUTION (Exercice 5.5.6).

En faisant dans la première identité $X = (X_1, \dots, X_n)$ et $Y = X_{n+1}$, on obtient la deuxième. La troisième identité s'obtient par itération de la deuxième. Pour obtenir la troisième, on applique plusieurs fois le résultat de l'exercice précédent : $H(X, Y|Z) = H(X, Y, Z) - H(Z)$, $H(Y|Z) = H(Y, Z) - H(Z)$, $H(X|Y, Z) = H(X, Y, Z) - H(Y, Z)$, et le résultat en découle.

SOLUTION (Exercice 5.5.7).

Les identités découlent immédiatement des définitions des quantités concernées. L'inégalité annoncée n'est autre que l'inégalité de Gibbs avec $p = p_{X,Y}$ et $q = p_X * p_Y$ ($(p_X * p_Y)(x, y) = p_X(x)p_Y(y)$). D'après le Théorème 5.1.1, l'égalité a lieu si et seulement si p et q sont identiques, c'est-à-dire si $p_{X,Y} = p_X * p_Y$, ce qui exprime l'indépendance de X et Y .

SOLUTION (Exercice 5.5.8).

On a (exercice précédent) :

$$H(X_{n+1}|X_1, \dots, X_n) \leq H(X_{n+1}|X_2, \dots, X_n)$$

Et par stationnarité :

$$H(X_{n+1}|X_2, \dots, X_n) = H(X_n|X_1, \dots, X_{n-1})$$

donc :

$$H(X_{n+1}|X_1, \dots, X_n) \leq H(X_n|X_1, \dots, X_{n-1})$$

Donc la suite $(H(X_n|X_1, \dots, X_{n-1}))_{n \geq 2}$ est convergente car minorée par 0 et décroissante. Elle converge vers un certain $H \geq 0$. On a ensuite, d'après l'Exercice 5.5.6,

$$\frac{1}{n}H(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n H(X_i|X_1, \dots, X_{i-1}).$$

Comme :

$$\lim_{n \rightarrow \infty} H(X_n|X_1, \dots, X_{n-1}) = H,$$

on a (théorème de Césaro) :

$$\lim_{n \rightarrow \infty} \frac{1}{n}H(X_1, \dots, X_n) = \lim_{n \rightarrow \infty} H(X_n|X_1, \dots, X_{n-1}) = H.$$

SOLUTION (Exercice 5.5.9).

(1) On a l'entropie différentielle :

$$\begin{aligned} h(X) &= - \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \left(\underbrace{\ln \frac{1}{\sqrt{2\pi\sigma^2}}}_{-\frac{1}{2} \log(2\pi\sigma^2)} - \frac{x^2}{2\sigma^2} \right) dx \\ &= \frac{1}{2} \ln 2\pi\sigma^2 + \frac{1}{2} \int \frac{x^2}{\sigma^2} f(x) dx. \end{aligned}$$

(2) La densité de probabilité de ce vecteur est :

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

L'entropie différentielle de X est donc égale à

$$\frac{1}{2} \ln \left((2\pi)^d |\Gamma| \right) + \frac{1}{2} E \left[(X - \mu)^T \Gamma^{-1} (X - \mu) \right]$$

On a :

$$\begin{aligned} E \left[(X - \mu)^T \Gamma^{-1} (X - \mu) \right] &= \sum_i \sum_j E \left[(X_i - \mu_i)(X_j - \mu_j) \right] (\Gamma^{-1})_{ij} \\ &= \sum_i \sum_j E \left[(X_j - \mu_j)(X_i - \mu_i) \right] (\Gamma^{-1})_{ij} \\ &= \sum_j \sum_i \Gamma_{ji} (\Gamma^{-1})_{ij} \\ &= \sum_j (\Gamma \Gamma^{-1})_{jj} = \sum_j 1 = d, \end{aligned}$$

ce qui donne :

$$h(X) = \frac{1}{2} \ln \left((2\pi e)^d |\Gamma| \right) .$$

SOLUTION (Exercice 5.5.10).

On a :

$$\Pr(X_\Delta = x_i) = p_i = f(x_i)\Delta .$$

On a donc

$$H(X_\Delta) = - \sum_{\mathbb{Z}} f(x_i)\Delta \log (f(x_i)\Delta) .$$

Sous l'hypothèse d'existence de toutes les intégrales impliquées (en particulier celle de $f(x) \log f(x)$) :

$$\begin{aligned} H(X_\Delta) &= - \sum_{i \in \mathbb{Z}} \left(\int_{i\Delta}^{(i+1)\Delta} f(x) dx \right) \log f(x_i) - \sum_{i \in \mathbb{Z}} \left(\int_{i\Delta}^{(i+1)\Delta} f(x) dx \right) \log \Delta \\ &= - \sum_{i \in \mathbb{Z}} \left(\int_{i\Delta}^{(i+1)\Delta} f(x) dx \right) \log f(x_i) - \left(\int_{\mathbb{R}} f(x) dx \right) \log \Delta \\ &= - \sum_{i \in \mathbb{Z}} \left(\int_{i\Delta}^{(i+1)\Delta} f(x) dx \right) \log f(x_i) - \log \Delta \end{aligned}$$

D'autre part,

$$\lim_{\Delta \downarrow 0} \sum_{i \in \mathbb{Z}} \left(\int_{i\Delta}^{(i+1)\Delta} f(x) dx \right) \log f(x_i) = \int_{\mathbb{R}} f(x) \log f(x) dx = -h(X) ,$$

d'où le résultat annoncé :

$$\lim_{\Delta \downarrow 0} (H(X_\Delta) + \log \Delta) = h(X) .$$

SOLUTION (Exercice 5.5.11).

(1)

$$\begin{aligned} D(f|g) &= \int f(x) \log \frac{f(x)}{g(x)} dx \\ &= \int \frac{f(x)}{g(x)} \log \frac{f(x)}{g(x)} g(x) dx \\ &= E \left[\frac{f(Y)}{g(Y)} \log \frac{f(Y)}{g(Y)} \right] \\ &= E \left[\frac{f(Y)}{g(Y)} \log \frac{f(Y)}{g(Y)} - \left(\frac{f(Y)}{g(Y)} - 1 \right) \right] \end{aligned}$$

Car :

$$E \left[\frac{f(Y)}{g(Y)} \right] = \int \frac{f(x)}{g(x)} g(x) dx = \int f(x) dx = 1$$

De plus, $z \log z - (z - 1) \geq 0$ avec égalité si et seulement si $z = 1$, ce qui montre que $D(f|g) \geq 0$ avec égalité si et seulement si :

$$P \left(\frac{f(Y)}{g(Y)} = 1 \right) = P(f(Y) = g(Y)) = 1,$$

C'est à dire si $f = g$.

(2) On vérifie d'abord que

$$f_{X,Y} \ll f_X \otimes f_Y.$$

En effet :

$$\begin{aligned} P(f_X(X)f_Y(Y) = 0) &= P(f_X(X) \text{ ou } f_Y(Y) = 0) \\ &\leq P(f_X(X) = 0) + P(f_Y(Y) = 0) = 0. \end{aligned}$$

Le reste est un cas particulier de (1).

(3) On a,

$$\begin{aligned} h(X|Y) &= -E \left[\log \frac{f_{X,Y}}{f_Y(Y)} \right] \\ &= - \iint f(x, y) \log \frac{f(x, y)}{f(y)} dx dy \\ &= - \iint f(x, y) \log f(x|y) dx dy \end{aligned}$$

Donc $h(X|Y) = h(X, Y) - h(Y)$. De plus $h(X|Y) \leq h(X)$, d'où :

$$\begin{aligned} I(X; Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \leq 0 \end{aligned}$$

SOLUTION (Exercice 5.5.13).

Soit g distribuée suivant $\mathcal{N}(0, \Gamma)$. On alors :

$$\begin{aligned} 0 \leq D(f|g) &= \int f \ln \frac{f}{g} = \int f \ln f - \int f \ln g \\ &= \int f \ln f - \int g \ln g \\ &= -(h(f) - h(g)). \end{aligned}$$

$(\ln g(X))$ est une fonction linéaire de XX^T , et son espérance est la même lorsque X a la densité f ou g .)

Solutions des exercices du chapitre 6

SOLUTION (Exercice 6.4.1).

Supposons qu'elle prenne au moins deux valeurs distinctes, a_1 and a_2 . Les ensembles $A_1 = \{f = a_1\}$ et $A_2 = \{f = a_2\}$ sont non vides et ne sont pas identiques à X . Ils n'appartiennent donc pas à la tribu grossière.

SOLUTION (Exercice 6.4.3).

Réponse : $\mathcal{X} = f^{-1}(\mathcal{E})$.

SOLUTION (Exercice 6.4.4).

Non. Prendre $f = 1_A - 1_{\bar{A}}$ où A est un ensemble non mesurable. Cette fonction n'est pas mesurable (par exemple, $\{f = 1\} = A \notin \mathcal{X}$), mais $|f| \equiv 1$ est mesurable.

SOLUTION (Exercice 6.4.5).

Soit ν la mesure de comptage sur \mathbb{Z} et pour tout $n \geq 1$, définissons $B_n = \{i \in \mathbb{Z} : |i| \geq n\}$. Alors $\nu(B_n) = +\infty$ pour tout $n \geq 1$, et

$$\nu\left(\bigcap_{n=1}^{\infty} B_n\right) = \nu(\emptyset) = 0.$$

SOLUTION (Exercice 6.4.6).

(e) $f \geq \frac{1}{n}1_{\{f \geq \frac{1}{n}\}}$ et donc (monotonie) : $0 = \mu(f) \geq \frac{1}{n}\mu(\{f \geq \frac{1}{n}\})$. En particulier, $\mu(\{f \geq \frac{1}{n}\}) = 0$, et (continuité séquentielle) $\mu(\{f > 0\}) = \lim_{n \uparrow \infty} \mu(\{f \geq \frac{1}{n}\}) = 0$.

(f) Avec $A = \{f > 0\}$, on a $\mu(f1_{\{f > 0\}}) = 0$, et donc, d'après (e), $f1_{\{f > 0\}} = 0$, μ -p.p.. En particulier, $\mu(\{f > 0\}) = 0$. De manière semblable, $\mu(\{f < 0\}) = 0$. On a donc $f = 0$ μ -p.p..

(g) Supposons (sans perte de généralité) que $f \geq 0$. On a alors $f \geq n1_{\{f \geq n\}}$ et donc (monotonie) : $\mu(\{f \geq n\}) \leq \frac{1}{n}\mu(f)$. Donc (continuité séquentielle) $\mu(\{f = \infty\}) = \lim_{n \uparrow \infty} \mu(\{f \geq n\}) = 0$.

SOLUTION (Exercice 6.4.7).

La fonction $\inf(f_n, f)$ est bornée par la fonction intégrable f . De plus, elle converge μ -presque partout vers f . Donc, par le théorème de convergence dominée, $\lim_{n \uparrow \infty} \int_X \inf(f_n, f) d\mu = \int_X f d\mu$. Le reste de la preuve découle de

$$\int_X |f_n - f| d\mu = \int_X f_n d\mu + \int_X f d\mu - \int_X \inf(f_n, f) d\mu.$$

SOLUTION (Exercice 6.4.8).

En invoquant le théorème de Beppo Levi pour la deuxième égalité :

$$\begin{aligned} \int_{\mathbb{R}_+} \frac{t e^{-at}}{1 - e^{-bt}} dt &= \int_{\mathbb{R}_+} \left(\sum_{n=0}^{+\infty} t e^{-(a+nb)t} \right) dt \\ &= \sum_{n=0}^{+\infty} \int_{\mathbb{R}_+} t e^{-(a+nb)t} dt \\ &= \sum_{n=0}^{+\infty} \frac{1}{(a+nb)^2}. \end{aligned}$$

SOLUTION (Exercice 6.4.9).

La fonction d'ensembles h^{-1} préserve les opérations sur les ensembles usuelles (complémentation, union, intersection). En particulier

$$h^{-1}\left(\sum_{n \geq 1} A_n\right) = \sum_{n \geq 1} h^{-1}(A_n)$$

et donc

$$\mu(h^{-1}(\sum_{n \geq 1} A_n)) = \sum_{n \geq 1} \mu(h^{-1}(A_n)).$$

Ceci prouve la sigma-additivité. D'autre part, $h^{-1}(\emptyset) = \emptyset$, et donc $\mu(h^{-1}(\emptyset)) = \mu(\emptyset) = 0$.

SOLUTION (Exercice 6.4.10).

Soit $\{A_n\}_{n \geq 1}$ une suite d'éléments de \mathcal{X} deux à deux disjoints. La propriété de sigma-additivité de ν s'écrit :

$$\int_X \left(\sum_{n \geq 1} 1_{A_n}(x) \right) h(x) \mu(dx) = \sum_{n \geq 1} \left(\int_X 1_{A_n}(x) h(x) \mu(dx) \right)$$

Pour la démontrer, on écrit

$$\int_X \left(\sum_{n=1}^K 1_{A_n}(x) \right) h(x) \mu(dx) = \sum_{n=1}^K \left(\int_X 1_{A_n}(x) h(x) \mu(dx) \right)$$

et on fait tendre K vers l'infini. On utilise pour le premier membre le théorème de la convergence monotone pour obtenir à la limite $\int_X \left(\sum_{n \geq 1} 1_{A_n}(x) \right) h(x) \mu(dx)$. La limite du second membre est quant à elle $\sum_{n=1}^K \left(\int_X 1_{A_n}(x) h(x) \mu(dx) \right)$, d'où le résultat annoncé : ν est bien sigma-additive. D'autre part, il est clair que $\nu(\emptyset) = 0$.

SOLUTION (Exercice 6.4.11).

Soit $t \in \mathbb{R}$ telle que $f(t) \neq g(t)$. Pour tout $c > 0$, il existe $s \in [t - c, t + c]$ tel que $f(s) = g(s)$ (sinon, l'ensemble $\{t; f(t) \neq g(t)\}$ contiendrait tout l'intervalle $[t - c, t + c]$, qui n'est pas de mesure de Lebesgue nulle. On peut donc construire une suite $\{t_n\}_{n \geq 1}$ convergeant vers t et telle que $f(t_n) = g(t_n)$ pour tout $n \geq 1$. Faisant tendre n vers l'infini, on obtient $f(t) = g(t)$, une contradiction.

SOLUTION (Exercice 6.4.12).

On applique le Théorème 6.1.5 avec le π -système formé par les rectangles $A_1 \times A_2$, $A_1 \in \mathcal{X}_1$, $A_2 \in \mathcal{X}_2$.

SOLUTION (Exercice 6.4.13).

On observera que dans le cas des mesures de comptage, "presque-partout" est équivalent à "partout" puisque pour une telle mesure, le seul ensemble de mesure 0 est l'ensemble vide. Appliqué au produit de deux mesures de comptage sur \mathbb{Z} , le théorème de Tonelli-Fubini traite de l'interversion des sommations : Soit $\{a_{k,n}\}_{k,n \in \mathbb{Z}}$ une suite doublement indexée de réels. Alors, si cette suite est absolument sommable, c'est-à-dire :

$$\sum_{k,n \in \mathbb{Z}} |a_{k,n}| < \infty,$$

la somme $\sum_{k,n \in \mathbb{Z}} a_{k,n}$ est bien définie, pour chaque $n \in \mathbb{Z}$

$$\sum_{k \in \mathbb{Z}} |a_{k,n}| < \infty$$

et

$$\sum_{k,n \in \mathbb{Z}} a_{k,n} = \sum_{n \in \mathbb{Z}} \left(\sum_{k \in \mathbb{Z}} a_{k,n} \right).$$

Si les termes de la suite doublement indexée sont non-négatifs, on peut dans tous les cas intervertir l'ordre des intégrations.

SOLUTION (Exercice 6.4.14).

Soit I un intervalle fini de \mathbb{R} . Observons que $\int_I e^{2i\pi x} dx = 0$ si et seulement si la longueur de I est un entier. Soit maintenant $I \times J$ un rectangle fini. Il a la propriété (A) si et seulement si $\int \int_{I \times J} e^{2i\pi(x+y)} dx dy = \int_I e^{2i\pi x} dx \times \int_J e^{2i\pi y} dy = 0$. (C'est ici qu'intervient Fubini.) Mais

$$\begin{aligned} \int \int_{\Delta} e^{2i\pi(x+y)} dx dy &= \int \int_{\cup_{n=1}^K \Delta_n} e^{2i\pi(x+y)} dx dy \\ &= \sum_{n=1}^K \int \int_{\Delta_n} e^{2i\pi(x+y)} dx dy = 0, \end{aligned}$$

puisque les Δ_n forment partition de Δ et ont tous la propriété (A).

SOLUTION (Exercice 6.4.15).

Si la fonction $t \rightarrow f(t)$ est intégrable, il en est de même de la fonction $t \rightarrow f(t)e^{-2i\pi\nu t}$.
Donc la transformée de Fourier de f est bien définie. Aussi

$$\begin{aligned} |\hat{f}(\nu)| &= \left| \int_{\mathbb{R}} f(t) e^{-2i\pi\nu t} dt \right| \\ &\leq \int_{\mathbb{R}} |f(t) e^{-2i\pi\nu t}| dt \\ &= \int_{\mathbb{R}} |f(t)| dt < \infty. \end{aligned}$$

Aussi, pour tout $h \in \mathbb{R}$,

$$\begin{aligned} |\hat{f}(\nu + h) - \hat{f}(\nu)| &= \left| \int_{\mathbb{R}} f(t) (e^{-2i\pi(\nu+h)t} - e^{-2i\pi\nu t}) dt \right| \\ &\leq \int_{\mathbb{R}} |f(t)| |e^{-2i\pi(\nu+h)t} - e^{-2i\pi\nu t}| dt \\ &= \int_{\mathbb{R}} |f(t)| |e^{-2i\pi h t} - 1| dt. \end{aligned}$$

La dernière intégrale est indépendante de ν . Elle tend vers 0 quand h tend vers 0, par convergence dominée puisque l'intégrande tend vers 0 et est bornée en valeur absolue par la fonction intégrable $2|f|$.

SOLUTION (Exercice 6.4.16).

1. On applique le théorème de Tonelli :

$$\begin{aligned} \int_{\mathbb{R}} \int_{\mathbb{R}} |f(t-s)g(s)| dt ds &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} |f(t-s)| dt \right) |g(s)| ds \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} |f(u)| du \right) |g(s)| ds \\ &= \left(\int_{\mathbb{R}} |f(u)| du \right) \int_{\mathbb{R}} |g(s)| ds < \infty. \end{aligned}$$

2. On applique (g) du Théorème 6.1.8.

3. L'intégrabilité de $f * g$ a été prouvée en 2. En changeant l'ordre d'intégration, on obtient :

$$\begin{aligned}
 \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(t-s)g(s) ds \right) e^{-2i\pi\nu t} dt &= \int_{\mathbb{R}} \int_{\mathbb{R}} f(t-s)e^{-2i\pi\nu(t-s)}g(s)e^{-2i\pi\nu s} ds dt \\
 &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(t-s)e^{-2i\pi\nu(t-s)} dt \right) g(s)e^{-2i\pi\nu s} ds \\
 &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(t-s)e^{-2i\pi\nu x} dx \right) g(s)e^{-2i\pi\nu s} ds \\
 &= \left(\int_{\mathbb{R}} f(t-s)e^{-2i\pi\nu x} dx \right) \left(\int_{\mathbb{R}} g(s)e^{-2i\pi\nu s} ds \right) \\
 &= \hat{f}(\nu)\hat{g}(\nu).
 \end{aligned}$$

L'application du théorème de Fubini est justifiée par le fait que la valeur absolue de l'intégrande $f(t-s)e^{-2i\pi\nu(t-s)}g(s)e^{-2i\pi\nu s}$ est intégrable par rapport à la mesure de Lebesgue sur \mathbb{R}^2 , comme il a été démontré en 1.

SOLUTION (Exercice 6.4.17).

Si $y \neq 0$, la fonction $x \rightarrow F(x, y)$ est continue et bornée sur $[0, 1]$, et elle est donc intégrable sur cet intervalle pour la mesure de Lebesgue. On a

$$\int_{[0,1]} f(x, y) dx = \left(\frac{x}{x^2 + y^2} \right)_0^1 = \frac{1}{1 + y^2}$$

Pour $y = 0$, $\int_{[0,1]} f(x, 0) dx = \int_{[0,1]} \frac{1}{x^2} dx = +\infty$. Donc,

$$\int_{[0,1]} f(x, y) dx = \frac{1}{1 + y^2}, \quad l - \text{a.e.},$$

et

$$\int_{[0,1]} \left(\int_{[0,1]} f(x, y) dx \right) dy = \int_{[0,1]} \frac{1}{1 + y^2} dy = \frac{\pi}{4}.$$

En observant que $f(x, y) = -f(y, x)$, on obtient :

$$\int_{[0,1]} \left(\int_{[0,1]} f(x, y) dy \right) dx = -\frac{\pi}{4}.$$

f n'est pas intégrable sur $[0, 1]^2$: sinon, d'après le théorème de Fubini, les deux intégrales seraient égales.

SOLUTION (Exercice 6.4.19).

$$\begin{aligned} E \left[e^{-\theta X} \right] &= E \left[e^{-\theta X} 1_{\{X=0\}} \right] + E \left[e^{-\theta X} 1_{\{X>0\}} \right] \\ &= E \left[1_{\{X=0\}} \right] + E \left[e^{-\theta X} 1_{\{X>0\}} \right] \\ &= P(X=0) + E \left[e^{-\theta X} 1_{\{X>0\}} \right]. \end{aligned}$$

La variable aléatoire $e^{-\theta X} 1_{\{X>0\}}$ est bornée uniformément en θ par une variable aléatoire (en fait, 1) et elle tend vers 0 quand $\theta \uparrow \infty$. Dans ces conditions, le théorème de convergence dominée s'applique :

$$\lim_{\theta \uparrow \infty} E \left[e^{-\theta X} 1_{\{X>0\}} \right] = 0.$$

SOLUTION (Exercice 6.4.20).

La fonction $(x, \omega) \rightarrow 1_{\{0 \leq x \leq X(\omega)\}}$ s'obtient par composition de $(x, \omega) \rightarrow (x, X(\omega))$ et de $(x, y) \rightarrow 1_{\{0 \leq x \leq y\}}$, qui sont mesurables (par exemple, la dernière fonction est du type $1_A(x, y)$ où A est un ouvert de \mathbb{R}^2). En appliquant le théorème de Tonelli à la fonction $(x, \omega) \rightarrow 1_{\{0 \leq x \leq X(\omega)\}}$ et à la mesure produit $\ell \times P$, on obtient :

$$E[X] = E \left[\int_{\mathbb{R}} 1_{\{0 \leq x \leq X\}} dx \right] = \int_{\mathbb{R}} E[1_{\{0 \leq x \leq X\}}] dx = \int_{\mathbb{R}} P(X > x) dx.$$

SOLUTION (Exercice 6.4.22).

$$\begin{aligned} P(f(X) = 0) &= E \left[1_{\{f(X)=0\}} \right] \\ &= \int_{\mathbb{R}^d} 1_{\{f(x)=0\}} f(x) dx = 0. \end{aligned}$$

SOLUTION (Exercice 6.4.23).

Le second membre est une fonction mesurable de X^2 . Il reste à démontrer que pour toute fonction mesurable φ non négative et bornée,

$$\begin{aligned} E \left[h(X) \varphi(X^2) \right] &= E \left[\left(h(\sqrt{X^2}) \frac{f_X(\sqrt{X^2})}{f_X(\sqrt{X^2}) + f_X(-\sqrt{X^2})} \varphi(X^2) \right) \right] \\ &\quad + E \left[\left(h(-\sqrt{X^2}) \frac{f_X(-\sqrt{X^2})}{f_X(\sqrt{X^2}) + f_X(-\sqrt{X^2})} \varphi(X^2) \right) \right] \end{aligned}$$

On va montrer que

$$E[h(X)1_{\{X=0\}}\varphi(X^2)] = E\left[\left(h(\sqrt{X^2})\frac{f_X(\sqrt{X^2})}{f_X(\sqrt{X^2}) + f_X(-\sqrt{X^2})}\right)\varphi(X^2)\right]$$

et

$$E[h(X)1_{\{X=0\}}\varphi(X^2)] = E\left[\left(h(-\sqrt{X^2})\frac{f_X(-\sqrt{X^2})}{f_X(\sqrt{X^2}) + f_X(-\sqrt{X^2})}\right)\varphi(X^2)\right]$$

Par exemple, pour l'avant-dernière égalité. Son membre de droite est :

$$\int_{-\infty}^{+\infty} \left(h(\sqrt{x^2})\frac{f_X(\sqrt{x^2})}{f_X(\sqrt{x^2}) + f_X(-\sqrt{x^2})}\right)\varphi(x^2)f_X(x)dx$$

ou (en décomposant le domaine d'intégration)

$$\int_0^{+\infty} (\cdots)\varphi(x^2)f_X(x)dx + \int_{-\infty}^0 (\cdots)\varphi(x^2)f_X(x)dx$$

c'est-à-dire, avec le changement ($x \rightarrow -x$) dans le second morceau,

$$\begin{aligned} \int_0^{+\infty} \left(h(\sqrt{x^2})\frac{f_X(\sqrt{x^2}) + f_X(-\sqrt{x^2})}{f_X(\sqrt{x^2}) + f_X(-\sqrt{x^2})}\right)\varphi(x^2)f_X(x)dx \\ = \int_0^{+\infty} h(\sqrt{x^2})\varphi(x^2)f_X(x)dx \\ = E[h(X)1_{X>0}\varphi(X^2)] . \end{aligned}$$

Solutions des exercices du chapitre 7

SOLUTION (Exercice 7.4.1).

Supposons d'abord que $\omega \in \limsup_n A_n$. Comme $\limsup_n A_n = \bigcap_{n=1}^{\infty} B_n$, ω appartient à tous les B_n . Donc si on fixe n'importe quel nombre N , $\omega \in B_N = \bigcup_{p=N}^{\infty} A_p$, et en particulier que $\omega \in A_{p_1}$ pour au moins un $p_1 \geq N$. On recommence avec un nouveau nombre N tel que $N \geq p_1 + 1$, et je trouverai au moins un $p_2 \geq N$, donc un $p_2 > p_1$, tel que $\omega \in A_{p_2}$; et ainsi de suite $p_1 < p_2 < p_3 \dots$ telle que $\omega \in A_{p_n}$ pour tout $n \geq 1$.

Inversement, supposons que pour un ω donné, on constate que $\omega \in A_p$ pour une infinité d'indices $p = p_1, p_2, \dots$ rangés par ordre croissant $p_1 < p_2 < p_3 < \dots$. Pour n'importe quel $n \geq 1$, ω appartient alors à $B_n = \bigcup_{p=n}^{\infty} A_p$, puisque à partir d'un certain rang les éléments de la suite $(p_j, j \geq 1)$ dépassent la valeur n . Ceci étant vrai pour tout $n \geq 1$, on en déduit que $\omega \in \bigcap_{n=1}^{\infty} B_n$, donc que $\omega \in \limsup_n A_n$.

SOLUTION (Exercice 7.4.2).

Si $\omega \in \bigcap_{n=1}^{\infty} C_n$, il existe au moins un indice N tel que $\omega \in C_N$. Or $C_N = \bigcap_{p=N}^{\infty} A_p$. Donc, $\omega \in A_p$ pour tous les $p \geq N$.

Inversement supposons que ω soit tel que l'on puisse trouver un N avec la propriété $p \geq N \Rightarrow \omega \in A_p$. On a donc $\omega \in C_N = \bigcap_{p=N}^{\infty} A_p$ et par conséquent $\omega \in \bigcup_{n=1}^{\infty} C_n$.

SOLUTION (Exercice 7.4.3).

1. Pour tout $\varepsilon > 0$, on a $P(X_n \geq \varepsilon) \leq p_n$ et donc $\lim_{n \uparrow \infty} p_n = 0$ entraîne que $X_n \xrightarrow{Pr.} 0$. Inversement si $X_n \xrightarrow{Pr.} 0$, alors en particulier $\lim_{n \uparrow \infty} P(X_n \geq \frac{1}{2}) = 0$. Or $P(X_n \geq \frac{1}{2}) = p_n$.

2. Pour tout $\varepsilon > 0$, $\sum_{n=1}^{\infty} P(X_n \geq \varepsilon) \leq \sum_{n=1}^{\infty} p_n$. Donc, si $\sum_{n=1}^{\infty} p_n < \infty$, d'après le lemme de Borel-Cantelli, $P(\limsup\{X_n \geq \varepsilon\}) = 0$, ce qui entraîne (Théorème 7.1.6) que $X_n \xrightarrow{p.s.} 0$. Inversement si $X_n \xrightarrow{p.s.} 0$, alors (Théorème 7.1.6) $P(\limsup\{X_n \geq \frac{1}{2}\}) = 0$, et donc (Borel-Cantelli inverse), $\sum_{n=1}^{\infty} p_n = \sum_{n=1}^{\infty} P(X_n \geq \frac{1}{2}) < \infty$.

SOLUTION (Exercice 7.4.4).

$$E \left[\sum_{n=1}^{\infty} 1_{A_n} \right] = \sum_{n=1}^{\infty} E[1_{A_n}] = \sum_{n=1}^{\infty} P(A_n) < \infty.$$

Donc $\sum_{n=1}^{\infty} 1_{A_n} < \infty$, P -presque sûrement. D'où le résultat.

SOLUTION (Exercice 7.4.5).

On utilise l'inégalité de Markov à la variable $\left|\frac{S_n}{n} - p\right|^4$, d'où

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^4} E\left[\left(\frac{S_n - np}{n}\right)^4\right]$$

Or $\frac{S_n}{n} - p = \frac{1}{n} \sum_{j=1}^n (X_j - p)$ et donc

$$E\left[\left(\frac{S_n - np}{n}\right)^4\right] = \frac{1}{n^4} \sum_{i,j,k,\ell=1}^n E[(X_i - p)(X_j - p)(X_k - p)(X_\ell - p)] .$$

Seuls subsistent les termes du type $E[(X_i - p)^4]$ et $E[(X_i - p)^2(X_j - p)^2]$. Or $E[(X_i - p)^4] = pq(p^3 + q^3)$ et il n'y a que n tels termes. De même $E[(X_i - p)^2(X_j - p)^2] = p^2q^2$ et il y a $3n(n-1)$ tels termes non nuls ($i \neq j$). D'où

$$E\left[\left(\frac{S_n - np}{n}\right)^4\right] = \frac{pq}{n^4}(n(p^3 + q^3) + 3pq(n^2 - n)) \leq C \frac{1}{4n^2} ,$$

pour une constante $C < \infty$.

SOLUTION (Exercice 7.4.6).

Soit A_n l'événement "il n'y a que des 1 dans le n -ème bloc". La suite $(A_n, n \geq 1)$ est une suite d'événements indépendants. D'autre part $P(A_n) = (\frac{1}{2})^{a_n} \geq \frac{1}{2n}$ et donc $\sum_{n=1}^{\infty} P(A_n) = \infty$. D'après le lemme de Borel-Cantelli (partie inverse), $P(\limsup_n A_n) = 1$. En d'autres termes, il y a une infinité de blocs ne contenant que des 1.

SOLUTION (Exercice 7.4.7).

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires IID de répartition commune F . Pour chaque $n \geq 1$ on définit

$$Z_n = \frac{X_1 + \cdots + X_{2^n}}{\sqrt{2^n}} .$$

On montre aisément, par récurrence, à partir de l'hypothèse de l'énoncé, que

$$P(Z_n \leq x) = F(x) .$$

Aussi, d'après le théorème de la limite gaussienne,

$$\frac{Z_n}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) .$$

On a donc $Z_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma)$ et $Z_n \stackrel{\mathcal{D}}{\sim} F$, et donc $F = \mathcal{N}(0, \sigma)$.

SOLUTION (Exercice 7.4.8).

Puisque la variable aléatoire est poissonnienne

$$\phi_{X_n}(u) = E[e^{iuX_n}] = e^{\lambda_n(e^{iu}-1)}$$

où $\lambda_n = E[X_n]$. Si $X_n \xrightarrow{\mathcal{D}} X$, nécessairement

$$\lim_{n \uparrow \infty} \phi_{X_n}(u) = \phi_X(u)$$

où $\phi_X(u) = E[e^{iuX}]$. En particulier la suite $(\lambda_n, n \geq 1)$ admet une limite $\lambda \in \mathbb{R}_+$. On a donc

$$\phi_X(u) = e^{\lambda(e^{iu}-1)}.$$

Commentaire. Il se peut que $\lambda = 0$, auquel cas X est dégénérée ($P(X = 0) = 1$)

SOLUTION (Exercice 7.4.9).

A. Soit $G_n(x)$ et $H_n(x)$ les fonctions de répartition de Y_n et Z_n respectivement.

$$\begin{aligned} G_n(x) &= P(Y_n \leq x) = 1 - P(Y_n > x) = 1 - P(X_1 > x, \dots, X_n > x) \\ &= 1 - P(X_1 > x) \dots P(X_n > x) = 1 - (1 - F(x))^n \\ H_n(x) &= P(Z_n \leq x) = P(X_1 \leq x, \dots, X_n \leq x) = P(X_1 \leq x) \dots P(X_n \leq x) \\ &= F(x)^n. \end{aligned}$$

Donc $\lim_{n \uparrow \infty} G_n(x) = 0$ si $x < a$, $= 1$ si $x > a$ et $\lim_{n \uparrow \infty} H_n(x) = 0$ si $x < b$, $= 1$ si $x \geq b$, c'est-à-dire, $\lim_{n \uparrow \infty} G_n(x) = 1_{\{x \geq a\}}$ pour tout $x \neq a$, et $\lim_{n \uparrow \infty} H_n(x) = 1_{\{x \geq b\}}$ pour tout $x \in \mathbb{R}$.

B. On a

$$P(nY_n \leq x) = P\left(Y_n \leq \frac{x}{n}\right) = G_n\left(\frac{x}{n}\right) = 1 - \left(1 - F\left(\frac{x}{n}\right)\right)^n.$$

Si $F(x) = x1_{\{0 \leq x \leq 1\}}$, alors

$$P(nY_n \leq x) = \left(1 - \left(1 - \frac{x}{n}\right)^n\right) 1_{[0, n]}(x),$$

et donc

$$\lim_{n \uparrow \infty} P(nY_n \leq x) (1 - e^{-x}) 1_{[0, \infty)}(x).$$

SOLUTION (Exercice 7.4.10).

$E[X_n Y_m] - E[X Y] = E[X_n Y_m - X Y] = E[X_n(Y_m - Y)] + E[Y(X_n - X)]$ et donc, d'après l'inégalité de Schwarz :

$$|E[X_n Y_m] - E[X Y]|^2 \leq E[X_n^2] E[(Y_m - Y)^2] + E[Y^2] E[(X_n - X)^2].$$

Comme $\lim_{n \uparrow \infty} E[(Y_n - Y)^2] = \lim_{n \uparrow \infty} E[(X_n - X)^2] = 0$, il suffit de prouver que $E[X_n^2]$ reste borné, ce qui découle de l'inégalité du triangle

$$E[X_n^2]^{1/2} \leq E[(X_n - X)^2]^{1/2} + E[X^2]^{1/2}.$$

SOLUTION (Exercice 7.4.11).

La fonction caractéristique de aS_n est $e^{-|a||u|}$. La fonction caractéristique de S_n/\sqrt{n} , $e^{-\frac{|u|}{\sqrt{n}}}$, tend vers la fonction $1_{\{0\}}(u)$ qui n'est pas une fonction caractéristique car non continue. Donc S_n/\sqrt{n} ne converge pas en distribution. On n'a pas non plus la convergence en probabilité car la convergence en probabilité entraîne la convergence en distribution. Ni la convergence presque sûre, car cette dernière entraîne la convergence en probabilité et en distribution. Ni la convergence en moyenne quadratique car celle-ci entraîne la convergence en probabilité et en distribution.

SOLUTION (Exercice 7.4.13).

La fonction caractéristique de S_n/n est celle d'une distribution de Cauchy. Donc S_n/n converge en distribution. Notons que si l'on définit la moyenne de la distribution de Cauchy comme étant 0 (la distribution de Cauchy est symétrique), on pourrait être amené à penser que S_n/n tend presque sûrement vers 0 en invoquant la loi des grands nombres. Mais ça n'est pas possible, car alors S_n/n tendrait en distribution vers la distribution ayant toute sa masse en 0, or on a vu que S_n/n tend en distribution vers une distribution de Cauchy. En fait, la loi forte des grands nombres de Kolmogorov exige que X_n soit intégrable, c'est-à-dire $E|X_n| < \infty$ et il n'en est rien lorsque X_n est Cauchy. La distribution de Cauchy *n'admet pas de moyenne*, bien qu'il soit tentant de dire qu'elle est de moyenne nulle.

Solutions des exercices du chapitre 8

SOLUTION (Exercice 8.5.1).

$E = \{1, 2, 3, 4, 5, 6\}$, $p_{13} = \frac{1}{2}$, $p_{14} = \frac{1}{2}$, $p_{24} = 1$, $p_{35} = \frac{1}{2}$, $p_{36} = \frac{1}{2}$, $p_{46} = 1$, $p_{55} = 1$, $p_{66} = 1$. On a :

$$P(X_2 = 6 | X_1 \in \{3, 4\}, X_0 = 1) = \frac{P(X_2 = 6, X_1 \in \{3, 4\}, X_0 = 1)}{P(X_1 \in \{3, 4\}, X_0 = 1)}.$$

Or,

$$P(X_2 = 6, X_1 \in \{3, 4\}, X_0 = 1) = P(X_2 = 6, X_0 = 1)$$

et

$$P(X_1 \in \{3, 4\}, X_0 = 1) = P(X_0 = 1).$$

Donc

$$\begin{aligned} P(X_2 = 6 | X_1 \in \{3, 4\}, X_0 = 1) &= \frac{P(X_2 = 6, X_0 = 1)}{P(X_0 = 1)} \\ &= P(X_2 = 6 | X_0 = 1) = \frac{1}{2} \times 1 + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}. \end{aligned}$$

Des calculs similaires (ou un simple coup d'œil sur le graphe de transition) donnent :

$$P(X_2 = 6 | X_1 \in \{3, 4\}, X_0 = 2) = P(X_2 = 6 | X_0 = 2) = 1.$$

SOLUTION (Exercice 8.5.2).

$$\begin{aligned} P(X_{n+1} = j_1, \dots, X_{n+k} = j_k | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) &= \\ \frac{P(X_{n+1} = j_1, \dots, X_{n+k} = j_k, X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0)}{P(X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0)} &= \frac{A}{B}, \end{aligned}$$

où

$$\begin{aligned} A &= P(X_0 = i_0) p_{i_0 i_1} \dots p_{i_{n-1} i} p_{i j_1} p_{j_1 j_2} \dots p_{j_{k-1} j_k} \\ B &= P(X_0 = i_0) p_{i_0 i_1} \dots p_{i_{n-1} i} \end{aligned}$$

et donc

$$\frac{A}{B} = p_{i j_1} p_{j_1 j_2} \dots p_{j_{k-1} j_k}$$

Des calculs similaires donnent la même valeur pour $P(X_{n+1} = j_1, \dots, X_{n+k} = j_k | X_n = i)$.

SOLUTION (Exercice 8.5.3).

$$\begin{aligned} P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\ = P(f(i, Z_{n+1}) = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) \\ = P(f(i, Z_{n+1}) = j | X_n = i) = P(f(i, Z_1) = j | X_0 = i). \end{aligned}$$

Des calculs semblables donnent

$$P(X_{n+1} = j | X_n = i) = P(f(i, Z_1) = j | X_0 = i).$$

SOLUTION (Exercice 8.5.4).

$$X_{n+1} = \begin{cases} S - Z_{n+1} & \text{if } X_n \leq s \\ X_n - Z_{n+1} & \text{if } X_n > s. \end{cases}$$

Donc $X_{n+1} = f(X_n, Z_{n+1})$ où

$$f(i, z) = (S - z) 1_{i \leq s} + (i - z) 1_{i > s}.$$

Le résultat découle du Théorème 8.1.3.

SOLUTION (Exercice 8.5.8).

$d = 1$ (on peut aller de 1 en 1 en 5 ou 6 étapes ; or le PGCD de 5 et 6 est 1).

SOLUTION (Exercice 8.5.9).

$d_i = \text{PGCD}(A)$, où $A = \{n \geq 1, p_{ij}(n) > 0\}$. Mais comme $p_{ii} > 0$, A contient 1. Donc son PGCD est 1.

SOLUTION (Exercice 8.5.11).

On vérifie que lorsqu'une telle distribution π existe, les équations de balance détaillée sont satisfaites. C'est bien le cas, puisque pour tout i , $i \geq 0$, on a $\pi(i)p_i = \pi(i+1)q_{i+1}$. En effet :

$$\pi(0) \frac{p_1 p_2 \cdots p_{i-1}}{q_1 q_2 \cdots q_i} p_i = \pi(0) \frac{p_1 p_2 \cdots p_i}{q_1 q_2 \cdots q_{i+1}} q_{i+1}.$$

SOLUTION (Exercice 8.5.12).

1) On doit vérifier que $\pi(1)p_{12} = \pi(2)p_{21}$. En effet

$$\frac{\beta}{\alpha + \beta} \alpha = \frac{\alpha}{\alpha + \beta} \beta$$

2) Pour $n = 1$, $P(T_1 = 1|X_0 = 1) = P(X_1 = 1|X_0 = 1) = p_{11} = 1 - \alpha$. Pour $n \geq 2$,
 $P(T_1 = n|X_0 = 1) = P(X_1 = 2, \dots, X_{n-1} = 2, X_n = 1|X_0 = 1) = p_{12} \underbrace{p_{22} \dots p_{22}}_{n-2} p_{21} =$
 $\alpha(1 - \beta)^{n-2} \beta.$

SOLUTION (Exercice 8.5.15).

Notons $\{Z_n\}_{n \geq 1}$, la suite IID des numéros de caillou tirés au sort. X_{n+1} est une fonction (indépendante de n) de X_n et de Z_{n+1} . De plus, la suite $\{Z_n\}_{n \geq 1}$ est indépendante de la configuration initiale X_0 . Donc $\{X_n\}_{n \geq 0}$ est une CMH d'après le Théorème 8.1.3.

On identifie l'état $S_{i_1} \dots S_{i_M}$ à la suite $I = (i_1, \dots, i_M)$. La matrice de transition a pour termes non nuls

$$p_{II} = \alpha_{i_1}, \quad p_{II_l} = \alpha_{i_l}$$

où $I = (i_1, \dots, i_M)$, $I_l = (i_1, \dots, i_l, i_{l-1}, i_{l+1}, \dots, i_M)$. Elle est irréductible (il suffit de montrer que dans le graphe de transition on peut trouver un chemin orienté de probabilité menant, pour tout $l \neq k$, de $I = (i_1, \dots, i_l, \dots, i_k, i_M)$ à $J = (i_1, \dots, i_k, \dots, i_l, i_M)$). L'espace d'état est fini, et donc la chaîne est positive récurrente. Soit $\pi = \{\pi(I)\}$ la probabilité $\pi((i_1, \dots, i_l)) = C \alpha_{i_1}^M \alpha_{i_2}^{M-1} \dots \alpha_{i_M}$. On vérifie les équations de balance détaillée

$$\pi(I) p_{II_l} = \pi(I_l) p_{I_l I}, \quad \text{pour tout } I, \text{ tout } l \geq 2$$

Mais

$$p_{II_l} = \alpha_{i_l}, \quad p_{I_l I} = \alpha_{i_{l-1}},$$

et donc

$$\left(C \alpha_{i_1}^M \alpha_{i_2}^{M-1} \dots \alpha_{i_{l-1}}^{M-l} \alpha_{i_l}^{M-l-1} \dots \alpha_{i_M} \right) \alpha_{i_l} = \left(C \alpha_{i_1}^M \alpha_{i_2}^{M-1} \dots \alpha_{i_l}^{M-l} \alpha_{i_{l-1}}^{M-l-1} \dots \alpha_{i_M} \right) \alpha_{i_{l-1}}.$$

SOLUTION (Exercice 8.5.16).

La probabilité de transition de (i, k) à (j, ℓ) en n étapes est $p_{ij}(n) p_{k\ell}(n)$. Comme \mathbf{P} est irréductible et *apériodique*, d'après le Théorème 8.1.10, pour toutes paires (i, j) et (k, ℓ) , il existe m tel que $n \geq m$ entraîne $p_{ij}(n) p_{k\ell}(n) > 0$. La chaîne produit est donc irréductible. Contre-exemple : la marche aléatoire symétrique à une dimension.

SOLUTION (Exercice 8.5.17).

$$\begin{aligned}
 & P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1 \in A_1, \dots, Z_k \in A_k) \\
 &= \sum_{m=0}^k P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1 \in A_1, \dots, Z_k \in A_k, \tau = m) \\
 &\quad + P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1 \in A_1, \dots, Z_k \in A_k, \tau > k) \\
 &= \sum_{m=0}^k P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_m^1 \in A_m, \tau = m, Z_{m+1}^2 \in A_{m+1}, \dots, Z_k^2 \in A_k) \\
 &\quad + P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_k^1 \in A_k, \tau > k).
 \end{aligned}$$

Comme $\{\tau = m\}$ est indépendant de $Z_{m+1}^2 \in A_{m+1}, \dots, Z_k^2 \in A_k$ ($k \geq m$),

$$\begin{aligned}
 &= \sum_{m=0}^k P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_m^1 \in A_m, \tau = m) P(Z_{m+1}^2 \in A_{m+1}, \dots, Z_k^2 \in A_k) \\
 &\quad + P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_k^1 \in A_k, \tau > k) \\
 &= \sum_{m=0}^k P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_m^1 \in A_m, \tau = m) P(Z_{m+1}^1 \in A_{m+1}, \dots, Z_k^1 \in A_k) \\
 &\quad + P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_k^1 \in A_k, \tau > k) \\
 &= \sum_{m=0}^k P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_m^1 \in A_m, \tau = m, Z_{m+1}^1 \in A_{m+1}, \dots, Z_k^1 \in A_k) \\
 &\quad + P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_k^1 \in A_k, \tau > k) \\
 &= P(X_0^1 \in C_1, X_0^2 \in C_2, Z_1^1 \in A_1, \dots, Z_k^1 \in A_k).
 \end{aligned}$$

SOLUTION (Exercice 8.5.18).

Ceci ne concerne que les distributions de $\{X_n^1\}_{n \geq 0}$ et $\{X_n^2\}_{n \geq 0}$, et donc on peut faire l'hypothèse d'une représentation

$$X_{n+1}^\ell = f(X_n^\ell, Z_{n+1}^\ell) \quad (\ell = 1, 2),$$

où $X_0^1, X_0^2, Z_n^1, Z_n^2$ ($n \geq 1$) satisfont les conditions de l'Exercice 8.5.17. On vérifie que pour τ , la condition de l'Exercice 8.5.17 est satisfaite. Avec $\{Z_n\}_{n \geq 1}$ comme dans l'Exercice 8.5.17, on a

$$X_{n+1} = f(X_n, Z_{n+1}),$$

ce qui prouve le résultat annoncé.

SOLUTION (Exercice 8.5.19).

(a) La question concerne la distribution de $\{X_n\}_{n \geq 0}$ et on peut donc supposer qu'on a la représentation

$$X_{n+1} = f(X_n, Z_{n+1})$$

où $\{Z_n\}_{n \geq 1}$ est IID, indépendante de X_0 . On a dans ce cas

$$Y_{n+1} = \begin{pmatrix} X_{n+1} \\ \vdots \\ X_{n+L} \\ X_{n+L+1} \end{pmatrix} = \begin{pmatrix} X_{n+1} \\ \vdots \\ X_{n+L} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ f(X_{n+L}, Z_{n+L+1}) \end{pmatrix}$$

c'est-à-dire

$$Y_{n+1} = \Phi(Y_n, Z_{n+L+1})$$

pour une certaine fonction Φ . Avec $\tilde{Z}_n = Z_{n+L}$, $n \geq 1$,

$$Y_{n+1} = \Phi(Y_n, \tilde{Z}_{n+1})$$

D'après le Théorème 8.1.3, $\{Y_n\}_{n \geq 0}$ est une CMH si $\{\tilde{Z}_n\}_{n \geq 1}$ est IID, indépendante de Y_0 .

La propriété IID découle de la propriété IID de $\{Z_n\}_{n \geq 1}$. Mais $Y_0 = (X_0, \dots, X_{n+L})$ est une fonction de $(X_0, Z_1, \dots, Z_{n+L})$ et est indépendante de Z_{n+L+1} .

Avec $I = (i_0, \dots, i_L)$ et $J = (i_1, \dots, i_L, i_{L+1})$, on a

$$p_{IJ} = p_{i_{L-1}i_L}$$

(b) On doit montrer que pour tout

$$I = (i_0, \dots, i_L) \in F$$

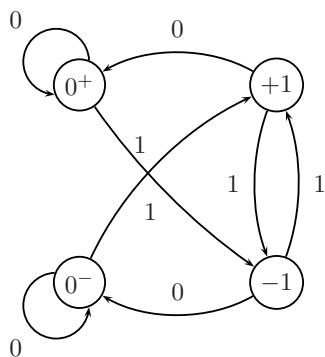
il y a un chemin de probabilité positive vers

$$J = (j_0, \dots, j_L) \in F$$

Ceci est vrai car il existe un chemin de probabilité positive de i_L à j_0 , disons $i_L, k_1, k_2, \dots, k_r, j_0$ tel que $p_{i_L k_1} p_{k_1 k_2} \dots p_{k_r j_0} > 0$ puisque $\{X_n\}_{n \geq 0}$ est irréductible.

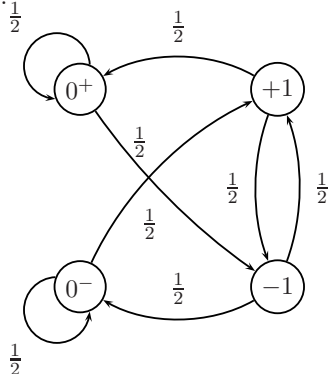
(c) $\sigma = \{\sigma(i_0, \dots, i_L)\}$, $(i_0, \dots, i_L) \in F$ est une distribution stationnaire de $\{Y_n\}_{n \geq 0}$ si et seulement si, lorsque $Y_0 \stackrel{\mathcal{D}}{\sim} \sigma$, alors $Y_1 \stackrel{\mathcal{D}}{\sim} \sigma$. Soit $\sigma(i_0, \dots, i_L) = \pi(i_0 i_1) p_{i_0} \dots p_{i_{L-1} i_L}$. Si $Y_0 \stackrel{\mathcal{D}}{\sim} \sigma$, alors $X_0 \stackrel{\mathcal{D}}{\sim} \sigma$, et la chaîne $\{X_n\}_{n \geq 0}$ est stationnaire, et en particulier X_0, \dots, X_L et X_1, \dots, X_{L+1} ont la même distribution.

SOLUTION (Exercice 8.5.22).



l'automate

Le graphe de transition :



La matrice de transition :

$$\mathbf{P} = \begin{pmatrix} 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix}$$

et pour $n \geq 2$,

$$\mathbf{P}^n = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

La probabilité stationnaire

$$\pi^T = \frac{1}{4} (1, 1, 1, 1) .$$

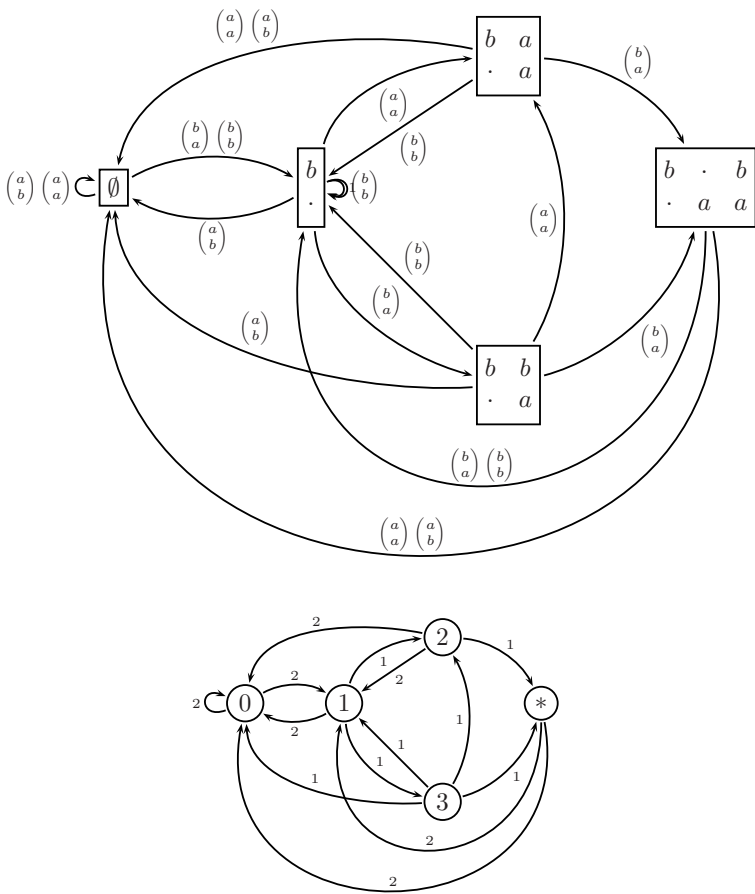
Pour $n \geq 2$

$$E[Y_n] = 0,$$

et

$$\text{cov}(X_n, X_{n+k}) = \begin{cases} +\frac{1}{2} & \text{si } k = 0; \\ -\frac{1}{4} & \text{si } k = \pm 1; \\ 0 & \text{autrement} \end{cases}$$

SOLUTION (Exercice 8.5.24).



(unité=1/2)

$$\pi(*) = \frac{9}{178}.$$

Index

- écart-type, 63
- élément aléatoire, 164
- événement, 3

- algèbre, 9
- analyse à un pas, 209
- apériodique
 - chaîne, 220
 - matrice de transition, 220

- balance
 - détaillée (équations de), 217
 - globale (équations de), 213
- Beppo Levi (théorème de), 156
- Bernoulli (vecteur aléatoire de), 36
- Bernstein
 - approximation polynomiale de, 48
- borélien (ensemble), 146
- borélienne (fonction), 147
 - simple, 148
- Borel
 - loi forte des grands nombres de, 181
- Borel–Cantelli, 178
- branchement (processus de), 54
- Buffon
 - l’aiguille de, 186

- chaîne de Markov, 203
 - récurrente, 224
 - positive, 225
 - réversible, 217
 - transiente, 224
- Chernoff, 184
- classes
 - cycliques, 221
 - de communication, 219

- CMH, 204
- codage de source, 133
- code
 - préfixe, 130
 - de Huffman, 134
 - uniquement déchiffrable, 130
- communication
 - classes de, 219
- continuité séquentielle
 - des mesures, 149
- convergence
 - dominée, 156, 163, 186
 - en distribution, 191
 - critère de la fonction caractéristique, 193
 - en loi, 191
 - monotone, 156, 163, 186
 - presque sûre, 180
 - presque-sûre
 - critère de —, 180
- corrélation (coefficient de), 85
- couplage, 235
- covariance (matrice de), 63, 86
- Crámer, 185
- cycliques
 - classes, 221

- déviation (grandes —), 184
- densité de probabilité, 61
 - conditionnelle, 105
 - marginale, 62
- distribution
 - d’un élément aléatoire, 164
- distribution stationnaire, 213

- entropie, 125
- ergodique
 - chaîne, 235
 - théorème — pour les CMH, 238
- espérance, 40, 161
 - conditionnelle, 107
- espace
 - mesuré, 148
 - mesurable, 146
- fonction
 - borélienne, 147
 - simple, 148
 - caractéristique, 68, 70, 193
 - de répartition, 150, 165, 192
 - génératrice, 49
 - mesurable, 147
- fonction génératrice, 49
 - d'une somme aléatoire, 50, 73
- formule
 - de Poincaré, 31
 - du changement de variable, 165
 - du produit des espérances, 46, 167
- Fubini, 158
- Gibbs, 126
 - échantillonneur de, 246
- grands nombres
 - loi faible des, 47
 - loi forte des, 177
- Huffman
 - code de, 134
- inégalité
 - de Chebyshev, 162
 - de Chernoff, 184
 - de Gibbs, 126
 - de Jensen, 163
 - de Kraft, 131
 - de Markov, 162
- indépendance
 - d'événements, 21
 - de variables aléatoires, 21
 - de vecteurs aléatoires, 69
- intégrale
 - de Lebesgue, 145
 - de Lebesgue–Stieltjes, 160
- intervalle de confiance, 194
- irréductible
 - chaîne, 219
 - graphe de transition, 219
 - matrice de transition, 219
- Ising, 246
- khi-deux
 - variable aléatoire du, 67
- Kolmogorov
 - loi forte des grands nombres de, 177
- Kraft, 131
- Lebesgue
 - mesure de, 149
 - théorème de, 156
- loi forte des grands nombres
 - d'Émile Borel, 178, 181
 - de Kolmogorov, 177, 182
 - grandes déviations à la, 184
- marche aléatoire, 206, 209
 - sur \mathbb{Z} , 206, 220
 - sur un graphe, 218
- Markov
 - chaîne de, 203
 - propriété de, 204
- matrice
 - de transition, 204
 - en n étapes, 205
 - irréductible, 219
 - stochastique, 204
- MCMC, 244
- mesurable
 - espace, 146
 - fonction, 147
- mesure, 148
 - de Lebesgue, 149
 - de probabilité, 150
 - de Radon, 150
 - finie, 150

- image, 156
- localement finie, 150
- produit, 158
- sigma-finie, 150
- Metropolis
 - algorithme d'échantillonnage de, 245
- Monte Carlo, 244
- moyenne, 63
- moyenne quadratique
 - convergence en, 196
- négligeable (ensemble), 151
- Neyman-Pearson
 - critère de, 120
 - test de, 120
- période, 220
- Paul Lévy
 - critère de la fonction caractéristique de, 193
- Poisson
 - événements rares de, 39
 - variable aléatoire de, 40
- presque
 - partout, 42, 151
 - sûrement, 43, 84
- probabilité conditionnelle, 19
- processus stochastique, 203
- quantité d'information, 125
- récurrent
 - état, 222
 - nul, 222
 - positif, 222
- régression linéaire, 88
- réversible
 - chaîne de Markov, 217
- sigma-additivité, 148
- sous-, 149
- simple
 - fonction borélienne, 148
- théorème
 - de Beppo Levi, 156
 - d'approximation, 148
 - de Fubini, 158
 - de la loi gaussienne limite, 191
 - de Lebesgue, 156
 - de Paul Lévy, 193
 - de Tonelli, 158
 - des conditionnements successifs, 111
 - du produit des espérances, 46, 167
- Tonelli, 158
- transformée de Crámer, 185
- transition
 - graphe de, 204
 - matrice de, 204
 - probabilité de, 204
- tribu, 11, 146
 - engendrée, 146, 161
 - grossière, 146
 - produit, 158
 - triviale, 146
- variable aléatoire, 5, 12, 161
 - du khi-deux, 67
 - binomiale, 37
 - de Cauchy, 67
 - de loi gamma, 65
 - de Poisson, 40
 - discrète, 35
 - exponentielle, 64
 - géométrique, 57
 - gaussienne, 64
 - au sens large, 92
 - réduite, 92
 - standard, 92
 - intégrable, 162
- variance, 63
- vecteur aléatoire, 36, 61, 161
 - gaussien, 92
 - multinomial, 38

From the same author :

Point Processes and Queues : Martingale Dynamics (1981)

ISBN 978-0-387-90536-5

An Introduction to Probabilistic Modeling (1994) ISBN 978-0-387-96460-7

Mathematical Principles of Signal Processing · Fourier and Wavelet Analysis (2002)

ISBN 978-0-387-95338-0

Markov Chains · Gibbs Fields, Monte Carlo Simulation, and Queues (2001)

ISBN 978-0-387-98509-1,

Elements of Queueing Theory · Palm Martingale Calculus and Stochastic Recurrences
(2003) ISBN 978-3-540-66088-0

See also : www.bluepinko.fr