Cours de probabilités
Licence de Physique

B. CESSAC
Université de Nice.

Plan du Cours :

Préambule

L'objectif de ce cours est de donner à un étudiant de Licence de Physique les bases de théorie des probabilités qui lui seront par la suite nécessaires pour aborder des cours de Physique plus avancés (notamment Physique statistique, phénomènes de transport,

Ce n'est cependant pas le seul objectif. En effet, il convient de faire les deux remarques suivantes. En premier lieu, le champ d'applications de la théorie des probabilités n'est pas limité à la seule physique : biologie, écologie, économie, géologie, sciences de l'ingénieur, etc.. sont autant de domaines où cette théorie est un outil fondamental d'analyse et de modélisation. En une époque où le développement de l'interdisciplinarité s'avère de plus en plus nécessaire, il peut être utile pour un jeune physicien, futur actif sur le marché du travail, de savoir que les méthodes qu'il voit en cours peuvent lui ouvrir des horizons plus vastes que la seule Physique. Dans cet esprit on s'efforcera de choisir des exemples et exercices dans des domaines variés (avec toutefois une prédominance d'exemples physiques). Par ailleurs, la théorie des probabilités est une branche des mathématiques et la littérature mathématicienne, dans ce domaine comme dans d'autres, est riche d'outils, méthodes d'analyse et théorèmes, mais aussi concepts, qui peuvent constituer une ~~mine~~ mine d'or pour un physicien qui n'a pas peur d'y pénétrer. Il serait dommage de se priver de cette potentialité. On s'efforcera donc de donner des définitions mathématiques (donc précises) des objets et ~~on~~ on se réfèrera souvent à la littérature mathématique, en donnant cependant toujours des illustrations concrètes. Il s'agit encore une fois de permettre aux étudiants de pénétrer la littérature mathématique de base sans être effrayé par des mots tels que tribu, borélien, mesurable, espérance conditionnelle, etc. Il ne s'agit pas de snobisme intellectuel : le snobisme consiste à mon sens plutôt à dédaigner la littérature mathématicienne en se disant que les physiciens peuvent s'en passer. Il s'agit plutôt de faire un cours d'introduction avec une ouverture d'esprit aussi large que possible.

Néanmoins, comme le temps est limité et que nous sommes des physiciens, on ne donnera en général pas les démonstrations, en renvoyant la littérature existante et en essayant plutôt d'illustrer les résultats par des exemples dont certains pourront, pourquoi pas, être simulés sur ordinateur. Cela fera ainsi un excellent complément au cours de simulation numérique.

Le cours comprendra ~~2~~ 3 chapitres

→ Elements de base de la théorie des probabilités

→ Convergences et théorèmes limites en théorie des probabilités.

Chapitre 1

Eléments de base de la théorie des probabilités

I] Introduction

La théorie des probabilités est une branche des mathématiques dont l'objet est l'étude des lois qui régissent les phénomènes aléatoires. Le terme aléatoire s'oppose, dans le langage courant, aux termes déterministe, reproductible, certain etc... Au cours de répétitions multiples de la même expérience ou essai (plus généralement, épreuve) un phénomène aléatoire se déroule chaque fois d'une façon légèrement différente. Comme exemples immédiats on peut citer : le jet d'une pièce de monnaie ou d'un dé; la roulette au casino ou les boules au loto, etc. mais aussi la pesée sur une balance, la mesure de l'aimantation d'un métal magnétisé, les effets d'un traitement de telle maladie sur un ensemble de malades, les pannes dans un réseau électrique, etc.... Le hasard ou aléa est bien sûr inhérent à tout phénomène, mais dans certains cas celui-ci peut être négligé : on extrait alors des lois déterministes décrivant ces phénomènes (mécanique par exemple). Dans d'autres cas cet aléa ne peut être négligé. Malgré tout, on peut encore extraire des lois statistiques propres à un grand nombre d'expériences (ou réalisations) de ce phénomène. Il n'y a donc pas opposition entre la notion d'aléa et la notion de loi. On verra même plus loin que l'évolution d'un phénomène aléatoire est souvent régie par des lois déterministes dès lors qu'on parle de l'évolution de grandeurs moyennes ou de distributions de probabilité.

D'un point de vue historique

RUBRIQUE: # PROBABILITÉS

PROBABILITÉS

La résolution des jeux est à l'origine du calcul des probabilités

Au départ, on s'appliquait à calculer
ü la quantité de cas favorables
ü parmi tous les cas possibles

L'improbable à toutes les chances de se produire
- Aristote

APPROCHE PAR LES JEUX

Date	Nom	Événement
années 1500	**Jérôme Cardan**	§ Écrit un traité *"de Ludo Aleae"* relatif au jeu de dés, au jacquet (ou backgammon)
vers 1600	Galilée	§ Étudie les dés: nombre de manière de faire un total donné avec 3 dés
1654	**Chevalier Méré**	§ Pose de nombreux problèmes à Pascal ü double six ü parties interrompues - pari de Pascal
1654	**Blaise Pascal** 1623-1662	§ Résout les énigmes que lui pose le chevalier Méré § Introduit le concept de probabilités ü rapport du nombre de cas favorables à un joueur sur ü le total des cas possibles Loi de Pascal - Pensées Le degré d'excitation qu'éprouve un joueur

	Doute…	§ Le sens et la définition mathématique de *probabilité* fait l'objet de discussions acharnées. § Nombre de mathématicien, comme d'Alembert, se méfiaient de cette branche, ü car ils considéraient ses concepts de base comme fumeux et ü ses méthodes comme beaucoup moins rigoureuses que celles de la géométrie § La science donnait la possibilité aux hommes de connaître les lois de l'Univers, rien ne pouvait être dû au hasard § Pour Laplace les *phénomènes aléatoires* ne l'étaient qu'en apparence ü C'est leur complexité qui empêchait d'en trouver l'explication ü Les probabilités sont une aide avant d'en savoir davantage sur le phénomène étudié
Fin des années 1700	***Condorcet***	§ Il pense que le calcul des probabilités peut s'appliquer à l'étude des phénomènes économiques et sociaux ü Il défend l'idée d'une « mathématique sociale » et ü Il considère le calcul des probabilités comme une branche des mathématiques à part entière
Années 1800	***Statistiques***	§ Naissance d'une nouvelle profession: l'actuariat § Organisation d'une nouvelle branche du savoir, la statistique, ü dont la base mathématique théorique est le calcul des probabilités
Fin des années 1800	***Extension***	§ Utilisation en médecine et en biologie § Et aussi, à l'hérédité § Développement de la mécanique statistique et de la théorie cinétique de la matière
	Encore des doutes…	Probability begins and ends with probability On a déjà montré qu'aucune connaissance des probabilités nous aide à savoir quelles sont les conclusions qui sont justes, et qu'il n'y a aucune relation directe entre la vérité d'une proposition et sa probabilité. Les probabilités commencent et finissent avec les probabilités - John Maynard Keynes - 1883-1946 - Économiste
	Tchebychev 1821-1894	§ Utilise les travaux d Markov § Enrichit la théorie et la rend plus rigoureuse
	Francis Galton 1822-1911 ***Karl Pearson*** 1857-1936	§ Application de la statistique à l'étude des caractères héréditaires dans le prolongement des idées de Charles Darwin § Ils fondent la biométrie ü C'est le début de la consécration des probabilités comme discipline mathématique
1933	***Kolmogorov***	§ Met en place une axiomatique ü qui est largement acceptée
XX^e siècle	***Mécanique quantique***	§ Grande consommatrice de probabilité

√@romath le goût des maths

PAGES
Arbres et dénombrements

- Les cubes diaboliques
  - 7 962 624 possibiltés
  - 89 944...
- Si on examinait les cubes...
  - Vers les solutions
  - Enfin...

DOSSIERS

- Sites favoris
- Seconde générale
  - Activités
  - Fiches de cours
  - Animations Flash
  - Evaluation
  - Math et Physique
  - Calculatrice
- Premières L
  - Sujets de révision
  - Liens utiles
  - Activités d'ouverture
  - BACCALAUREAT 2001
- Option Math 1eL et TL
- Premières S
- Terminales S
  - Présentation
  - **QCMs**
- Terminales ES
- Excursion dans les graphes
- Suite Story
- Mathématiques et économie
- Les listes
- Arbres et dénombrements

02535

Arbres et dénombrements

Elie

Probabilités, ou théorie des probabilités

Branche des mathématiques qui s'attache à mesurer ou à déterminer quantitativement la probabilité qu'a un événement ou une expérience d'aboutir à un résultat donné. Cette théorie est fondée sur l'étude des permutations et des combinaisons. Elle constitue la base tous les travaux en statistiques.

Bref historique

On attribue en général à **Blaise Pascal** et à **Pierre de Fermat** l'invention au XVIIe siè d'une première théorie des probabilités appliquée aux jeux de hasard, même si Jérôme Cardan s'était déjà penché sur la question dès le XVIe siècle.

Cinquante ans plus tard, dans son ouvrage posthume Ars conjectandi (1713), **Jacques Be** systématisa le calcul des probabilités, en énonçant des théorèmes prometteurs tels que l'additivité des probabilités.

Avec les travaux de **Darwin** et du statisticien **Quételet** , la vision probabiliste du monde s'affirma encore davantage, englobant tous les domaines de la science (voir Statistique, mécanique). Aujourd'hui, les probabilités possèdent un vaste champ d'application, allant conception des ordinateurs à l'étude de l'engorgement des aérodromes.

Pascal, Blaise (1623-1662), mathématicien, physicien, théologie mystique, philosophe, moraliste et polémiste français du XVIIe siècle. L'étendue des domaines d'intérêt et du génie de Pascal est impressionna : inventeur de la machine à calculer concepteur des premiers transports commun en France, artisan de l'assèchement des marais poitevins, polémiste brillant contre les jésuites dans les Provinciales, apologiste de foi chrétienne avec les fragments rassemblés sous le titre de Pensées, fut également l'un des plus brillants prosateurs de la langue française et l'une des plus grandes figures du X siècle français.
Il conçut en 1654 un triangle, appel depuis triangle de Pascal utile à de nombreux calculs arithmétiques. Il travailla ensuite sur les probabilités partir de deux problèmes de jeu et t de géométriser le hasard. Il travailla sur l'infini mathématique (voir Infinitésimal, calcul) et mit au poin méthode d'induction en mathématique. Il est également à l'origine des méthodes combinatoir

schwab de

Les probabilités

Isabelle CUCHET

Cette branche des mathématiques consiste à déterminer la probabilité qu'a un événement d'aboutir à un résultat donné. Les Français Blaise Pascal (1623-1662) et Pierre de Fermat (1601-1665) ont été les premiers à se pencher sur ce domaine des sciences. Aujourd'hui, cette théorie est utilisée par tous les statisticiens.

En quoi consistent les probabilités ? Prenons un exemple. Un homme joue aux dés avec ses amis. Il gagnera la partie s'il sort un six. Or il existe six façons possibles de retomber pour le dé. Parmi ces six cas, un seul cas favorable (le six) aboutira au gain de la partie par l'homme. Le rapport 1/6 représente la probabilité pour que l'homme gagne la partie.

Plus généralement, les mathématiciens appellent E l'événement dont ils veulent déterminer la probabilité. Ils passent en revue tous les cas qui peuvent réellement se produire. Il en existe un nombre n. Parmi ces n cas, seul un nombre f d'entre eux conduira à l'événement E. La probabilité que E ait lieu est donc égale à f/n. Une probabilité est un nombre compris entre 0 et 1. La probabilité 0 indique que l'événement n'a aucune chance de se produire. La probabilité 1 indique que l'événement se produit à chaque fois.

La théorie des probabilités est à la base des études statistiques. Par exemple, sachant que la probabilité d'obtenir un total de 7 en lançant deux dés est de 1/6, on peut en déduire que si deux dés sont lancés de façon aléatoire un très grand nombre de fois, environ un sixième des résultats sera égal à 7.

Les probabilités mathématiques sont largement utilisées en physique quantique, pour déterminer par exemple le lieu où se trouvent les électrons autour d'un noyau atomique, en sciences sociales, pour les recensements de population, ainsi que dans l'industrie, le commerce ou le marché boursier.

Création :07/02/2002
Dernière modification : 07/02/2002

Retour au sommaire
Liste des grands dossiers

Comme on l'a vu le champ d'applications de la théorie des probabilités est vaste et ne se limite certainement pas à la Physique. Voici quelques exemples dont la plupart seront traités en cours ou en travaux dirigés.

1) <u>Lois de probabilités usuelles</u>

<u>Loi binomiale</u> : Marche aléatoire, tirages de Bernoulli

<u>Loi de Poisson</u> : Processus de désintégration

<u>Loi de Gauss</u> :

<u>Lois de puissances</u> : Tremblements de terre, Loi de Zipf, Criticalité auto organisée

2) <u>Statistiques des systèmes de grande taille</u>

* Théorie de l'information, entropie, grandes déviations
* Loi des grands nombres
* Physique statistique
* Statistiques ; sondages.

3) <u>Systèmes désordonnés</u>

* verres de spins
* Quasi cristaux
* réseaux de neurones

4) <u>Processus aléatoires</u>

* Mouvement Brownien
* Processus de diffusion, équation de la chaleur
* Equations de transport
* Localisation d'Anderson, persistence

* Processus de Markov
* Diffusion sur une variété (applications en pharmacologie)
* Processus de désintégration
* Réactions en chaîne.
* Décharge neuronale.

5) Systèmes dynamiques chaotiques

* Chaos
* fractals
* théorie ergodique
* De la dynamique microscopique à la physique statistique)

II] Axiomatique de base de la théorie des probabilités

1) Espace probabilisé

1. 1) Univers

L'un des premiers concepts de la théorie des probabilités est celui d'événement aléatoire ou événement fortuit. On donne ce nom à tout fait qui peut ou peut ne pas se produire à la suite d'une expérience à issue aléatoire.

L'ensemble de tous les événements est appelé univers des possibles, ou univers fortuit. Cet ensemble peut être muni d'une structure plus ou moins complexe, être fini ou infini, dénombrable ou non dénombrable, etc...
On note traditionnellement cet ensemble Ω et on note $\omega \in \Omega$ un événement.
Voici quelques exemples d'univers

α] Tirage à pile ou face - tirage de Bernoulli

Les événements sont donc pile ou face, et l'univers $\Omega = \{\text{pile}, \text{face}\}$
Le tirage à pile ou face est l'archétype de situations où le résultat d'une expérience est binaire : blanc ou noir, vrai ou faux, oui ou non, haut ou bas, etc... Il est commode de représenter de tels cas par un codage binaire 0 ou 1. On parle alors de tirage de Bernoulli. Il s'agit d'un exemple fondamental car il est à la fois extrêmement simple (c'est le cas - 2 événements - le plus simple tout en étant non trivial, en théorie des probabilités) mais il permet aussi, comme on le verra tout au long de ce cours, de saisir des notions relativement élaborées, ce qui fixe ensuite un cadre à l'intuition si l'on souhaite aborder des exemples plus complexes.

β] Univers fini.

Dans le cas des jeux de hasard (jeu de dé, jeux de cartes, loto, tiercé, loterie) mais aussi dans d'autres cas (

), l'univers Ω

est fini. Par exemple l'univers correspondant au jet d'un dé à 6 faces est $\Omega = \{1, 2, 3, 4, 5, 6\}$, celui correspondant au tirage d'une carte dans un jeu de 52 est $\Omega = \{1♥, 1♦, \text{etc.} \ldots, R♣\}$

Exercice : Ecrire l'univers correspondant au tirage de 2 dés, au tiercé, quartet, quintet (20 chevaux), au loto. NB : l'usage des pointillés et du "etc" peut-être judicieux.

Dans le cas d'un univers fini le calcul des probabilités se ramène la plupart du temps à un problème de combinatoire. On rappelle des éléments de base du calcul combinatoire.

Nombre de permutations de N objets : $N!$

(N possibilités pour le premier, N-1 pour le second, etc...)

Nombre de combinaisons de k objets parmi N :

Dans une combinaison les objets sont indiscernables. Par exemple, si l'on veut placer 2 boules parmi 5 et qu'on représente en rouge les 2 boules placées et en bleu les 3 autres, les configurations suivantes sont équivalentes (comptent pour une configuration, au final)

$$\underset{(r1)(r2)(b1)(b1)(b3)}{\bullet\bullet\bullet\bullet\bullet} \Leftrightarrow \overset{(b1)\,b2\ b3}{\underset{(r2)(r1)}{\bullet\bullet\bullet\bullet\bullet}} \Leftrightarrow \underset{r_1 r_2\ b_2 b_1 b_3}{\bullet\bullet\bullet\bullet\bullet} \Leftrightarrow \bullet\bullet\bullet\bullet\bullet \Leftrightarrow \bullet\bullet\bullet\bullet\bullet \ \text{etc} \ldots$$

Les boules sont bien entendu des objets physiques distincts. C'est ce qu'on matérialise par les notations entre parenthèses pour les boules. Mais, dans le calcul du nombre de combinaisons, on ne les distingue pas. Dans le cas représenté on a donc $2!\,3!$ configurations équivalentes. Plus généralement, pour placer k boules parmi N on a $k!(N-k)!$ configurations équivalentes. Comme on a en tout $N!$ configurations le nombre de configurations distinctes (non équivalentes) est :

$$\boxed{C_N^k = \frac{N!}{k!(N-k)!}} \qquad (1.1)$$

Nombre d'arrangements : Ici on place toujours k boules parmi N mais les k boules placées (et seulement elles) sont discernables. On n'a donc plus que $(N-k)!$ configurations équivalentes et le nombre d'arrangements est :

$$\boxed{A_N^k = \frac{N!}{(N-k)!}} \qquad (1.2)$$

Nombre de bijections $\mathcal{B} \circlearrowleft$

Si $\text{card}(\mathcal{B}) = N$ alors

le nombre de bijection est $N!$ ~~(1.3)~~

Nombre de parties de $\mathcal{B}$.

Soit $P(\mathcal{B})$ l'ensemble des parties de $\mathcal{B}$, on a :

$$\text{card}\, P(\mathcal{B}) = 2^{\text{card}\, \mathcal{B}}$$

* Par récurrence $\text{Card}\, P(\emptyset) = 1$; $\text{Card}\, F = n+1$; $E = F \setminus \{e\}$; $P(F) = A \cup B$; $e \in A$; $e \notin B$

$\text{Card}\, A = 2^n$ ($A = \{$partie de $E \cup e\}$) ; $\text{card}\, B = 2^n$ ($B = P(E)$) $\Rightarrow \text{Card}\, P(F) = 2^{n+1}$ *

Coefficient multinomial

Placer n_1 objets de type 1, n_2 objets de type 2 ... n_k objets de type k ...

parmi $n = n_1 + n_2 \cdots + n_k$.

$$C_n^{n_1 n_2 \cdots n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

Exercice : Montrer que $\sum_{n_1 \cdots n_K} C_N^{n_1 \cdots n_K} x_1^{n_1} x_2^{n_2} \cdots x_K^{n_K} = (x_1 + \cdots + x_K)^N$

avec $\sum_{i=1}^{K} n_i = N$. Cette formule généralise celle du binôme de Newton.

Dans le cas d'un univers fini, il y a un cas important, qui est celui d'univers produit. Considérons une expérience (tirage à pile ou face, jeu de dé, etc...) qu'on répète N fois. On appelle ω_t le résultat de la t-ième expérience et on s'intéresse à la séquence des N tirages. L'univers qui nous intéresse dans ce cas est donc composé d'événements du type :

$$[\tilde{\omega}]_N = (\omega_1, \dots \omega_N)$$

L'univers a dans ce cas une structure de produit : si Ω est l'univers d'une expérience, alors l'univers des N expériences est donc Ω^N, et le nombre de ses éléments est $|\Omega|^N$. Cet exemple est important lorsqu'on considère des phénomènes aléatoires dépendant du temps (processus aléatoires).

Exercice : Écrire l'univers correspondant à 2, puis 3, puis 4, puis N tirages de Bernoulli successifs. Pour représenter ces événements il est utile d'utiliser le codage binaire :

$$\boxed{X([\tilde{\omega}]_N) = \sum_{k=1}^{N} \frac{\omega_k}{2^k}} \qquad (1.6)$$

Ce codage a deux avantages. D'une part il permet de simplifier l'écriture des événements, (représentés par N symboles), en associant à chacun d'eux un nombre réel. Par ailleurs, l'application $X : [\tilde{\omega}]_N \to X([\tilde{\omega}]_N)$ est un exemple de variable aléatoire. Cet exemple présente comme le verra plus bas, une très grande richesse.

c) Univers dénombrable

Il existe des cas où, même s'il est possible de compter les événements, i.e. mettre en correspondance chaque événement avec un nombre entier, le nombre d'événements est infini. Par exemple l'événement : "on tire un nombre au hasard dans $[0,1]$ et celui-ci est rationnel". Cet exemple est clairement plutôt abstrait, et on peut arguer que l'infini est une idéalisation mathématique qui n'existe pas dans la nature. Outre le fait qu'il s'agit d'une remarque philosophique, l'infini est parfois bien utile pour approcher le fini quand celui-ci est très grand. On voit par exemple dans le cours de physique statistique de nombreux cas où il est plus simple de considérer que le nombre de particules, typiquement de l'ordre du nombre d'Avogadro ($\mathcal{N} = 6.\ 10^{23}$), est infini : cela simplifie en particulier les calculs. De manière plus générale, les méthodes statistiques font bien souvent appel à des passages à la limite où le nombre d'événements tend vers l'infini (loi des grands nombres, théorème central limite, théorie de grandes déviations, ...). Cela permet d'avoir une première estimation d'un résultat auquel on peut ensuite appliquer des corrections de taille finie. Notons que ce type d'approche est utilisée dans des cas d'assemblées avec un grand nombre d'éléments (population, gaz de particules, galaxie, réseau de neurones), mais aussi lorsqu'on considère un phénomène évoluant dans le temps (processus aléatoire) et qu'on considère à un régime asymptotique. Dans le premier cas on a une limite de type "thermodynamique" (le nombre d'individus tend vers l'infini) alors que dans le second c'est le temps qui tend vers l'infini.

Un exemple intéressant qui peut servir à la fois de base à un cours de physique statistique comme illustration de la limite thermodynamique, m aussi comme introduction aux processus aléatoires est de nouveau le tirage de Bernoulli. L'exemple ci-dessus est une transition vers la partie suivante puisque l'univers est en fait, non dénombrable.

Exercice : On considère une suite infinie de tirages de Bernoulli successifs. On note $\tilde{\omega} = (\omega_1 \dots \omega_t \dots)$ (séquence infinie à choix) un événement. Quel est l'univers ? Utiliser la représentation précédente et montrer que celle-ci associe à Ω l'intervalle $[0,1]$ et à chaque $\tilde{\omega}$ un nombre de $[0,1]$. La correspondance est-elle bijective ? A quel type de nombres correspondent les séquences du type $(\omega_1 \dots \omega_n, 0 \dots 0 \dots)$, les séquences périodiques, les séquences non périodiques ? En déduire que l'univers est non dénombrable, i.e. un ensemble fini à la puissance du dénombrable peut être non dénombrable.

NB : Dans le cas dénombrable, on ne peut pas directement utiliser la combinatoire mais peut utiliser des formules d'approximation lorsque $N \to +\infty$ (Stirling)

d) Univers non dénombrable

C'est typiquement le cas si les événements correspondent à la mesure de grandeurs réelles : énergie, vitesse, position, etc... L'univers peut être un intervalle ou une union d'intervalle dans $\mathbb{R}$, un domaine de $\mathbb{R}^n$, de $\mathbb{C}^n$ etc... Cela peut être aussi une surface dans $\mathbb{R}^n$ (variété) etc.

e) Autres cas :

En théorie Ω peut être virtuellement n'importe quoi. Cela peut être en particulier un espace fonctionnel : les événements sont des fonctions, un espace de probabilités i.e. on considère des probabilités aléatoires. On rencontre ce cas en physique des milieux désordonnés (verres de spin, quasicristaux, etc...). Ces cas ne sont donc pas des cas d'écoles destinés à se triturer les méninges : ils se rencontrent en physique.

1.2) Tribu

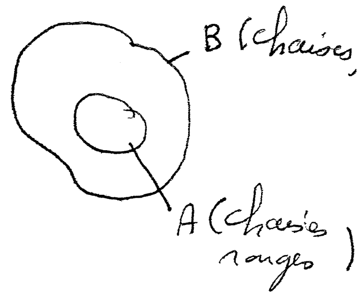
Une science n'est pas constituée d'une accumulation de faits. Pour qu'un ensemble d'expériences ou d'observations donne lieu à une explication phénoménologique ou une théorie scientifique il convient d'établir des liens logiques entre ces événements. La connaissance scientifique repose en particulier sur une structuration logique faisant intervenir les notions logiques de base telles que l'implication, l'équivalence, le ou, ~~de~~ et, la négation...

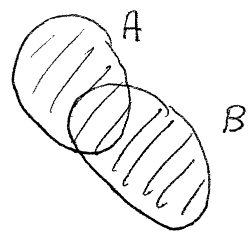
Il convient donc, en théorie des probabilités, de donner un sens à des propositions telle que : "la probabilité que l'événement A ou l'événement B soit réalisé est 0.1."

La théorie des probabilité étant formulée dans un contexte ensembliste il convient de donner un sens à ces relations logiques dans le cadre de la théorie des ensembles. On rappelle les correspondances suivantes

Logique		Théorie des ensembles
Implication	$A \Rightarrow B$	$A \subset B$
Et	A et B	$A \cap B$
Ou inclusif	A ou B	$A \cup B$
Ou exclusif	A ou bien B	$A \Delta B$ (différence symétrique)
Négation	non A	$\bar{A} = {}^{c}A$

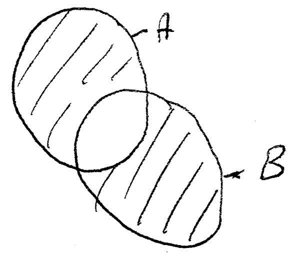
(Ex : c'est une chaise rouge $\Rightarrow$ c'est une chaise)

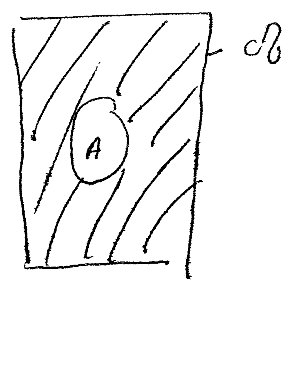




$\Leftrightarrow$ A et pas B ou B et pas A

$$A \Delta B = (A \cap \bar{B}) \cup (B \cap \bar{A})$$





Exercice : Montrer que

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$

A partir de ces relations on construit sur la base de l'univers Ω un "méta" ensemble $\mathcal{F}$, tel que :

i) $\Omega \in \mathcal{F}$

ii) Si $A \in \mathcal{F}$, $\bar{A} = \complement_{\Omega} A \in \mathcal{F}$

iii) Si $A \in \mathcal{F}$ et $B \in \mathcal{F}$, $A \cap B \in \mathcal{F}$

iV) Si A_k est une suite infinie d'événements $\bigcap_k A_k$ et $\bigcup_k A_k \in \mathcal{F}$.

L'ensemble $\mathcal{F}$ est appelé <u>σ-algèbre</u> ou <u>tribu</u>. Sa structure est conçue en sorte que si $\mathcal{F}$ contient un ensemble d'événements (resp. de propositions) elle contient toutes les combinaisons logiques créées à partir des relations logiques élémentaires citées plus haut.

$\mathcal{F}$ contient en particulier

- L'événement certain (Ω)
- L'événement impossible $(\emptyset)$

et la tribu "minimale" est $(\Omega, \emptyset)$. Mais $\mathcal{F}$ a en général une structure complexe.

Par exemple, pour le jeu de pile/face $\mathcal{F} = \emptyset, \{0\}, \{1\}, \{0, 1\}$. $\mathcal{F}$ contient donc le résultat impossible "rien", le 0, le 1 et aussi le cas $\{0$ ou $1\}$. On ne peut pas avoir 0 et 1 $(\{0\} \cap \{1\} = \emptyset)$. Considérons maintenant le jet de 2 pièces. L'univers est composé des paires ordonnées $[\omega_1 \omega_2]$ avec $\omega_i = 0, 1$. Donc $\Omega = \{[0,0], [0,1], [1,0], [1,1]\}$ et

$\mathcal{F} = \emptyset, [0,0], [0,1], [1,0], [1,1], [0,0] \cup [0,1], [0,0] \cup [1,0]$ etc.

Dans le cas où Ω est fini, $\mathcal{F}$ est composé de <u>l'ensemble des parties de Ω</u>. Dans le <u>cas général</u> ce n'<u>est pas</u> <u>l'ensemble des parties</u>. Dans le cas où Ω est $\mathbb{R}$ ou un intervalle de $\mathbb{R}$, on construit une tribu remarquable engendrée à partir des ouverts de $\mathbb{R}$. On l'appelle tribu des boréliens. Un intérêt fondamental de cette construction est qu'on une correspondance entre la tribu et la topologie. On peut alors utiliser les propriétés métriques de $\mathbb{R}$ pour faire des calculs (d'intégrales notamment)

Quelques mots sur la propriété (iv). On a vu plus haut (et on verra plus bas) qu'il est parfois utile de faire des passages à la limite. Cette propriété est une condition technique permettant de faire ces passages à la limite. Par exemple, dans la démonstration de la loi forte des grands nombres (ch 2.2) on est amené à considérer la probabilité d'une intersection infinie dénombrable d'événements ainsi qu'une union infinie. Comme (on verra plus bas) une probabilité associe à tout élément de $\mathcal{F}$ un nombre $\in [0,1]$, il convient que l'intersection et l'union dénombrable appartiennent à $\mathcal{F}$.

Exercice : Soit une suite u_n tendant vers une limite u^*. Donner la définition de $\lim_{n\to+\infty} u_n = u^*$. Supposons maintenant que $u_n(x)$ soit une fonction de x, avec $x \in [0,1]$. Pour $\varepsilon > 0$ fixé on définit les ensembles $A_n(\varepsilon) = \{x \in [0,1] \mid |u_n(x) - u^*| < \varepsilon\}$. Ecrire l'ensemble des x tels que $\exists m > 0, \forall n > m, |u_n(x) - u^*| < \varepsilon$. Comment écririez-vous l'ensemble des x tels que $\lim_{n\to+\infty} u_n(x) = u^*$? Quel est le problème ? Comment y remédier ?

* $\lim_{n\to+\infty} u_n = u^* \iff \forall \varepsilon > 0, \exists m > 0, \forall n \geq m, |u_n - u^*| < \varepsilon$.

L'ensemble des x t.q. $\exists m > 0, \forall n > m, |u_n(x) - u^*| < \varepsilon$ est $\bigcup_{m=0}^{+\infty} \bigcap_{n=m}^{+\infty} A_n(\varepsilon)$

L'ensemble des x t.q. $u_n(x) \to u^*$ pourrait s'écrire $\bigcap_{\varepsilon>0} \bigcup_{m=0}^{+\infty} \bigcap_{n=m}^{+\infty} A_n(\varepsilon)$, mais la première intersection est non dénombrable. On peut y remédier en choisissant une suite rationnelle (dénombrable donc) de ε_k ($= \frac{1}{k}$ par ex) qui tend vers 0 et utiliser la densité de $\mathbb{Q}$ ...

Exercice : Montrer que la tribu $\mathcal{F}$ correspondant à une suite infinie de tirages de Bernoulli est composée des "cylindres" $[\omega_k \dots \omega_{k+n}]$ composés des séquences dont les bits sont fixés aux valeurs $\omega_k \dots \omega_{k+n}$ entre k et $k+n$. c.a.d

$$[\omega_k \dots \omega_{k+n}] = \{\tilde{\omega} \in \{0,1\}^{\mathbb{N}} \mid \omega'_k = \omega_k ; \omega'_{k+1} = \omega_{k+1} ; \dots \omega'_{k+n} = \omega_{k+n}\}$$

Montrer que dans le codage binaire, ces cylindres correspondent à des intervalles ouverts de diamètre $\frac{1}{2^{n-k}}$. Quelle est l'image de la tribu $\mathcal{F}$ par le codage binaire ?

Voir correction ci-jointe

Incompatibilité : Une notion importante en théorie de probabilités est celle d'événements incompatibles. Deux événements A, B sont incompatibles ssi

$$\boxed{A \cap B = \emptyset} \qquad (1.7)$$

Cela signifie qu'ils ne peuvent être réalisés simultanément. Cette notion est structurelle et ne fait absolument pas intervenir de probabilité (contrairement à la notion d'indépendance avec laquelle les étudiants la confondent parfois :)

Exercice : Lemme de Borel-Cantelli

Soit A_n une suite dénombrable d'évènements. On suppose que $\sum_{n=0}^{+\infty} P(A_n) < \infty$. En déduire que $P\left[\bigcap_{m=0}^{+\infty} \bigcup_{n=m}^{+\infty} A_n\right] = 0$.

Soit $u_n(x)$ une suite de fonctions, où $x \in [0,1]$. Quelle est la définition de $u_n(x) \underset{n\to+\infty}{\longrightarrow} u^*$? Pour $\varepsilon > 0$ fixé on définit la famille d'ensemble de $[0,1]$ $A_n(\varepsilon) = \{x ; |u_n(x) - u^*| \geq \varepsilon\}$.

On suppose que pour tout ε, $\sum_{n=0}^{+\infty} P(A_n(\varepsilon)) < \infty$. En déduire que la probabilité que $\lim_{n\to+\infty} u_n(x) = u^*$ est égale 1 (i.e. $P\left[\{x ; \lim_{n\to+\infty} u_n(x) = u^*\}\right] = 1$.

On a $\bigcap_{m=0}^{+\infty} \bigcup_{n=m}^{+\infty} A_n \overset{\forall m}{\subset} \bigcup_{n=m}^{+\infty} A_n$. Donc

$$P\left[\bigcap_{m=0}^{+\infty} \bigcup_{n=m}^{+\infty} A_n\right] \leq P\left[\bigcup_{n=m}^{+\infty} A_n\right] \leq \sum_{n=m}^{+\infty} P(A_n), \quad \forall m$$

Les termes de la forme $\sum_{n=m}^{+\infty} P(A_n)$ constituent la queue de la série $\sum_{n=0}^{+\infty} P(A_n)$. Comme cette série converge, ~~parkalan~~ $\sum_{n=m}^{+\infty} P(A_n) \underset{m\to+\infty}{\longrightarrow} 0$

Comme l'inégalité ci-dessus est vraie $\forall m$, $P\left[\bigcap_{m=0}^{+\infty} \bigcup_{n=m}^{+\infty} A_n\right] = 0$.

La limite $\lim_{n\to+\infty} u_n(x) = u^*$ signifie que $\forall \varepsilon > 0$, $\exists m > 0$, $\forall n \geq m$, $|u_n(x) - u^*| < \varepsilon$. En termes ensemblistes l'ensemble des x tels que pour $\varepsilon > 0$, $\exists m > 0$, $\forall n > m$, $|u_n(x) - u^*| < \varepsilon$

est l'ensemble $\bigcup_{m=0}^{+\infty} \bigcap_{n=m}^{+\infty} \overline{A_n(\varepsilon)}$ dont le complémentaire est

$\bigcap_{m=0}^{+\infty} \bigcup_{n=m}^{+\infty} A_n(\varepsilon)$. C'est par définition, l'ensemble des x tels que $u_n(x) \not\to u^*$

En vertu du résultat précédent si $\sum P(A_n(\varepsilon)) < \infty$ alors

$P\left[\bigcap_{m=0}^{+\infty} \bigcup_{n=m}^{+\infty} A_n(\varepsilon)\right] = 0$. Ce résultat est vrai $\forall \varepsilon > 0$,

donc la probabilité de l'ensemble des x tels que $u_n(x) \not\to u^*$

est nulle. Donc $P[\{x ; \lim_{n \to +\infty} u_n(x) = u^*\}] = 1$.

1.3) Probabilité

La donnée de l'univers Ω et de la tribu $\mathcal{F}$ constitue ce qu'on appelle un espace probabilisable. A noter que toutes les notions introduites jusqu'à maintenant (et notamment la notion d'incompatibilité) ne nécessitent pas la donnée préalable d'une probabilité.

L'idée très simple derrière la notion de probabilité est d'associer à un ensemble d'événements (qui peut être une union ou une intersection de sous-ensembles de Ω) un nombre compris entre 0 et 1, qui attribue à l'événement impossible la valeur 0, à l'événement certain la valeur 1, et qui est une fonction croissante dans le sens où $A \subset B \Rightarrow P(A) \leq P(B)$.

Plus précisément une probabilité est une fonction d'ensembles

$$P : \mathcal{F} \longrightarrow [0,1]$$

telle que :

(i) $P(\emptyset) = 0$; $P(\Omega) = 1$

(ii) Si $A_1, A_2 \ldots A_k \ldots$ est une séquence (finie ou dénombrable) d'ensembles disjoints alors :

$$P\left[\bigcup_{k=1}^{+\infty} A_k\right] = \sum_{k=1}^{+\infty} P(A_k)$$

La propriété (ii) est appelée continuité monotone. De nouveau, elle est essentielle pour les passages à la limite (cf exercice sur le lemme de Borel-Cantelli)

Exercice : Montrer que les propriétés (i)(ii) entraînent :

(iii) $P(\bar{A}) = 1 - P(A)$

(iv) $A \subset B \Rightarrow P(A) \leq P(B)$

(v) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (formule des probabilités totales)

Exercice : Généraliser la formule des probabilités totales à 3, puis à 4 événements.

Autres exercices : voir feuille jointe.

Indépendance : Deux événements A, B sont indépendants ssi

$$\boxed{P(A \cap B) = P(A)\,P(B)} \qquad (1.8)$$

Cette notion est fondamentale et sera utilisée tout au long de ce cours. Contrairement à la notion d'incompatibilité, la notion d'indépendance nécessite la donnée préalable d'une probabilité de référence. En d'autres termes deux événements peuvent être indépendants par rapport à une probabilité et pas par rapport à une autre

Ex.

Exercice : Montrer que deux événements incompatibles ne sont en général pas indépendants (sauf si P(A) ou P(B) = 0). Donner un exemple de 2 événements incompatibles qui ne sont pas indépendants.

Exercice : Evenements indépendants pour une probabilité et pas pour une autre → feuille jointe.

Exercice : Peut-on affirmer que 2 jets successifs d'un dé sont indépendants ? Si oui, pourquoi n'y a t'il pas incompatibilité avec ce qui est dit plus haut ?

Définition : Le triplet $(\Omega, \mathcal{A}, P)$ est appelé espace probabilisé.

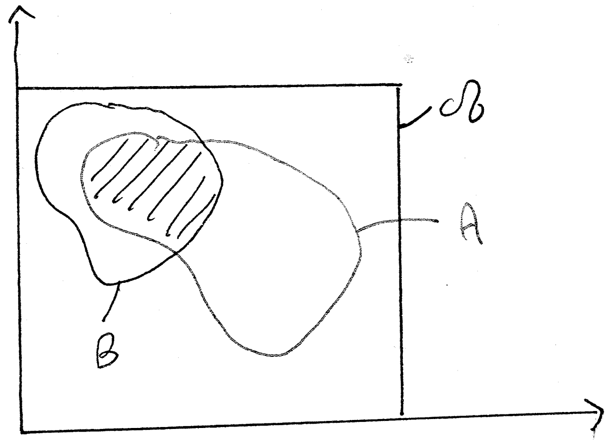
1.4) Probabilités conditionnelles

La propriété (i) attribue à l'univers Ω la probabilité maximale, 1. Plus généralement la probabilité est relative à un univers donné ou univers de référence Ω. Il est néanmoins possible de changer d'univers de référence, en prenant un sous ensemble de Ω.

Ainsi soit $A \subset \Omega$, la probabilité P_A ou $P(\cdot \mid A)$ définie par

$$P_A(B) = P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad \forall B \in \mathcal{A}$$

est appelée probabilité conditionnelle de B sachant A. On voit que c'est bien une probabilité, mais on voit que tous les événements considérés appartiennent tous à A (puisqu'on prend l'intersection avec A). Par ailleurs $P_A(A) = \frac{P(A \cap A)}{P(A)} = 1$. L'univers de référence de P_A est donc A.



NB On utilise l'équation précédente aussi sous la forme :

$$\boxed{P(A \cap B) = P(B|A)\,P(A)} \qquad (1.9)$$

La notion de probabilité conditionnelle (comme la notion d'espérance conditionnelle qu'on verra plus loin) est extrêmement importante en théorie des probabilités. Elles interviennent dans de nombreuses situations dont on verra quelques exemples en travaux dirigés. Un exemple immédiat et simple correspond à la situation suivante. On cherche à connaître certaines propriétés du système : (Ω). On fait un ensemble d'hypothèses a priori sur Ω, en sorte que ces hypothèses soient mutuellement incompatibles et sont suffisantes pour décrire toutes les possibilités : en termes mathématiques, cela signifie qu'on se donne un ensemble complet d'événements $\{A_i\}$, c'est à dire :

$$\boxed{\begin{array}{lll} A_i \cap A_j = \emptyset & \forall i \neq j & \text{(incompatibilité mutuelle)} \\ \bigcup_i A_i = \Omega & & \text{(partition de } \Omega\text{)} \end{array}}$$

On se donne par ailleurs les probabilités a priori que les événements A_i soient réalisés ($P(A_i)$). (Si l'on n'a aucune information a priori sur ces probabilités il est naturel de prendre $P(A_i) = \frac{1}{N}$)

Muni de ces hypothèses on fait maintenant une première expérience qui donne le résultat R_1. On dispose donc maintenant d'une information supplémentaire et il convient donc de modifier les probabilités des événements A_i en conséquence. En d'autres termes, puisque le résultat R_1 a été obtenu on restreint l'univers à l'ensemble des événements correspondant à R_1. On veut donc calculer $P(A_i | R_1)$.

On a :

$$P(R_1) = P(R_1 \cap \Omega) = P(R_1 \cap \bigcup_i A_i) = P\left[\bigcup_i (R_1 \cap A_i)\right]$$
$$= \sum_i P(R_1 \cap A_i) = \sum_i P(A_i)\, P(R_1 | A_i)$$

et :

$$P(A_i | R_1) = \frac{P(A_i \cap R_1)}{P(R_1)} = \frac{P(A_i)\, P(R_1 | A_i)}{P(R_1)}$$

Ce qui nous donne finalement la formule de Bayes :

$$\boxed{P(A_i|R_1) = \frac{P(A_i)\,P(R_1|A_i)}{\sum_k P(A_k)\,P(R_1|A_k)} \qquad (1.10)}$$

Puisque qu'on connait R_1 et les A_i on peut en général calculer $P(R_1|A_i)$ et on obtient donc $P(A_i|R_1) = P_{R_1}(A_i)$. Les probabilités $P_{R_1}(A_i)$ pondèrent les hypothèses A_i ~~en~~ tenant compte de l'information acquise en faisant l'expérience R_1 et donnent en général des poids distincts ~~aux~~ A_i. On peut alors faire une nouvelle expérience R_2 qui nous permet de calculer $P(A_i|R_1, R_2) = P_{R_1}(A_i|R_2) = P_{R_1,R_2}(A_i)$ ... etc.
En multipliant les expériences on fait progressivement émerger une (ou plusieurs) hypothèses qui ont un poids statistique plus important et qui permet de faire de choisir (valider) une des hypothèses. (de façon statistique, donc en général sans certitude absolue).

Exercices : voir feuille jointe.

2) Variables aléatoires :

2.1) Définitions :

Jusqu'à présent nous avons parlé d'~~événements~~, d'univers, de probabilités de façon relativement abstraite. Cependant, dans de nombreux cas, on a finalement à traiter d'événements qui s'expriment en termes de nombres. Ces nombres sont par exemple associés à une mesure dans une expérience. Le courant électrique à un instant donné dans un circuit, la pression d'un gaz à un instant donné, l'aimantation ~~instantanée~~ d'un matériau, le pourcentage de gens qui affichent une intention de vote pour X à la date t, la croissance au second semestre 2004 ... sont des nombres, qui mesurent, certaines caractéristiques du système étudié au moment où est faite la mesure, (qui, manifestement, peut être extrêmement complexe)

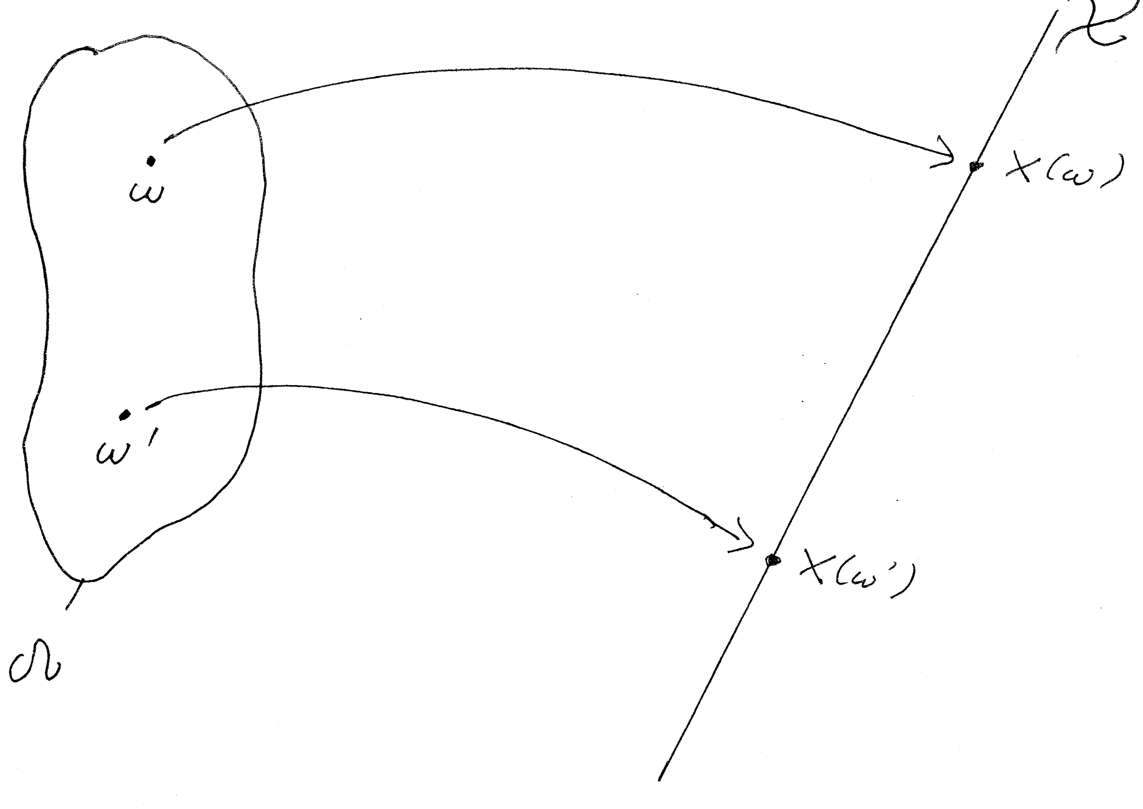
On ne connait en général donc pas le système (Ω) de façon détaillée mais on peut supposer qu'à l'instant où l'on fait la mesure il se trouve dans un "état" ω, et que le résultat mesuré est une fonction de ω.

Si, par exemple Ω représente l'ensemble des états microscopiques (*) d'un matériau magnétique, l'aimantation associe à chacun de ces états une valeur numérique, etc...

Cette image, certes imprécise et naïve a pour seul but de motiver la définition mathématique d'une variable aléatoire.

Une variable aléatoire est une fonction mesurable d'un espace probabilisé $(\Omega, \mathcal{F}, P)$ dans un espace de nombres ($\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{C}$ ...).

(*) Il y a ici un "flou artistique" sur la notion d'état, qui est cependant suffisant pour introduire de façon intuitive les notions abstraites qui suivent.



Si par exemple Ω est l'ensemble des microétats d'un gaz et X la pression, $X(\omega)$ mesure cette pression lorsque le gaz est dans l'état ω. Comme ω fluctue au cours du temps, $X(\omega)$ fluctue aussi, de façon aléatoire. On n'a pas accès à Ω, on a seulement accès ~~aux~~ à l'ensemble des valeurs que prend X.

Le concept de variable aléatoire permet donc de passer de l'espace abstrait Ω à un espace beaucoup plus accessible, un espace de nombres noté $\mathcal{X}$.

Ce passage nécessite de transporter la structure probabiliste de Ω sur $\mathcal{X}$. En effet, on veut pouvoir donner un sens à des expressions comme $\text{Prob}[X \in [x, x+dx[]$, $\text{Prob}[X \in A]$, $\text{Prob}[X = 2]$ ~~mais~~ aussi $\text{Prob}[X = 2 \text{ ou } X = 3]$, $\text{Prob}[X \geq 3 \text{ et } X \leq 4]$ etc... A partir de la probabilité sur Ω on construit une probabilité sur $\mathcal{X}$, P_X, appelée probabilité induite par P ou probabilité de X sous P, donnée par :

(2.1) $$\boxed{P_X(A) = P[\{\omega \mid X(\omega) \in A\}] \; ; \forall A \in \mathcal{B}}$$

$\mathcal{B}$ est la tribu qu'on a sur $\mathcal{X}$; c'est la tribu des boréliens)

La définition de P_X nécessite donc que tout $A \in \mathcal{B}$ ait un antécédent dans $\mathcal{F}$, c'est à dire que l'image réciproque de tout $A \in \mathcal{B}$ appartient à $\mathcal{F}$. C'est le sens du mot <u>mesurable</u> dans la définition d'une variable aléatoire. C'est une condition technique qu'on peut omettre lorsqu'on fait des calculs mais qu'il faut néanmoins avoir à l'esprit lorsqu'on souhaite construire des modèles probabilistes par exemple.

<u>En résumé</u> :

Une variable aléatoire X est une fonction mesurable d'un espace probabilisé $(\Omega, \mathcal{F}, P)$ dans un espace de nombres, probabilisé $(\mathcal{X}, \mathcal{B}, P_X)$. La loi P_X ou loi de X sous P est donnée par :

$$P_X(A) = P[\{\omega \mid X(\omega) \in A\}], \quad \forall A \in \mathcal{B}.$$

On appelle <u>réalisation</u> d'une variable aléatoire la valeur qu'elle prend pour un ω donné : c'est donc un <u>nombre</u> alors que X est une <u>fonction</u>. On note en général les variables aléatoires avec une <u>majuscule</u> et les réalisations avec une <u>minuscule</u>. Ainsi $P_X(x) = P[\{\omega \in \Omega \mid X(\omega) = x\}]$ est la probabilité que la variable aléatoire X prenne la valeur x.

<u>NB</u> : On oublie alors en général $\Omega, \mathcal{F}, P$ et l'<u>espace de référence</u> devient $(\mathcal{X}, \mathcal{B}, P_X)$

2.2) <u>Exemples</u>

Selon que $\mathcal{X}$ est discret ou continu, la variable aléatoire X est dite <u>discrète</u> ou <u>continue</u>.

a) <u>$\mathcal{X}$ est fini</u>

Dans ce cas la variable aléatoire X prend un <u>nombre</u> fini de valeurs $x_1, x_2 \dots x_k$. Bien souvent le calcul de P_X se ramène à un calcul de combinatoire où l'on compte <u>le nombre</u> de cas favorables et le nombre de cas possible (|X|) et la probabilité est donnée par le rapport $\dfrac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}$.

Exemples :

* Tirage de Bernoulli :

On associe au tirage à pile ou face la variable aléatoire X telle X(pile) = 0, X(face) = 1, avec probabilités

$$P(X=0) = \quad ; \quad P(X=1) = \quad , \qquad p+q=1$$

Pour un tirage non biaisé (pièce non truquée $p = q = \frac{1}{2}$)

* Jeu de dé :

Pour un dé à N faces on associe à chaque face la variable aléatoire $X(\text{face } i) = i$, avec probabilité p_i $\left(\sum_{i=1}^{N} p_i = 1\right)$. Pour un dé non truqué tous les p_i sont égaux à $\frac{1}{N}$ et la loi P_X est dite uniforme.

* N tirages de Bernoulli indépendants.

On fait N tirages indépendants et on note $X_t \in \{0,1\}$ la variable aléatoire correspondant au tirage t. On définit une nouvelle variable aléatoire :

$$S_N = \sum_{t=1}^{N} X_t$$

S_N est parfois appelé variable de comptage puisqu'elle compte le nombre de 1.

Celle-ci prend donc ses valeurs sur l'ensemble fini $\{k\}_{k=0}^{N}$ où k correspond au nombre de fois où 1 apparait dans une séquence de tirage donnée. On va calculer la loi de probabilité de S_N

On a donc : $P[S_N = k]$ = Proba [1 apparaisse k fois dans séquence aléatoire $X_1 - X_N$]

Il y a différentes façons de faire ce calcul. En voici une. On répertorie d'abord toutes les séquences favorables. Par exemple pour $N=5$ et $k=2$ les séquences favorables sont

11 000 ; 10100 ; 10010 ; 10001 ; 01100 etc ~

La probabilité cherchée est la probabilité de l'union de ces séquences et comme celles-ci sont incompatibles (1 tirage ne peut pas valoir à la fois 0 et 1) c'est la somme, sur toutes les configurations favorables de la probabilité de chaque configuration. Chaque configuration contient k 1 et $N-k$ 0. Les tirages étant indépendants la probabilité de chaque séquence favorable est $p^k q^{N-k}$.

Combien y a t'il de séquences favorables ? Il s'agit de placer k 1 (indiscernables) parmi N bits. On a donc C_N^k configurations favorables. La loi cherchée est donc :

$$\boxed{P[S_N = k] = C_N^k p^k q^{N-k}} \qquad (2.2)$$

Cette loi, très importante est appelée <u>loi binomiale</u>. Elle se rencontre sitôt que l'on fait <u>N tirages indépendants</u> d'une variable aléatoire <u>binaire</u> et que l'on souhaite calculer la probabilité qu'une valeur apparaisse k fois dans la séquence.

<u>Exemple</u> : marche aléatoire → voir feuille jointe.

* N tirages d'un dé à K faces. (*)

Dans l'esprit de l'exemple précédent on fait maintenant N tirages indépendants d'un dé à K faces où la face i a la probabilité p_i avec $\sum_{i=1}^{K} p_i = 1$. On cherche à connaître la probabilité que la face i apparaisse n_i fois, avec $\sum_{i=1}^{K} n_i = N$. De façon tout à fait analogue on repèrerait les configurations favorables. Chaque configuration a la probabilité $p_1^{n_1} \cdots p_K^{n_K}$ et le nombre de configurations favorables est donné par le coefficient multinomial $C_N^{n_1 \cdots n_K} = \frac{N!}{n_1!\, n_2! \cdots n_K!}$. La loi cherchée est donc

$$\boxed{P[n_1, n_2 \cdots n_K] = C_N^{n_1 \cdots n_K}\, p_1^{n_1} \cdots p_K^{n_K}} \qquad (2.3)$$

Cette loi, qui généralise la loi binomiale, est appelée <u>loi multinomiale</u>.

<u>Exercice</u>: Vérifier que $P[n_1 \cdots n_K]$ est bien normalisée.

(*) On sort ici du cadre fixé dans cette section, (on calcule une proba conjointe) mais il s'agit d'une extension immédiate du cas précédent.

b) X est dénombrable

Ici la variable aléatoire prend une infinité de valeurs discrètes. Dans ce cas, bien sûr, on ne peut plus calculer la probabilité par la formule cas favorables / cas possibles puisque le nombre de cas possibles est infini. On peut néanmoins parfois utiliser celle-ci en faisant ensuite un passage à la limite.

Exemples:

* Loi de Poisson

La variable aléatoire $X = 0, 1 \ldots$ ~~est~~ de Poisson de paramètre $\lambda > 0$

si :

$$\boxed{P[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}} \qquad (2.4)$$

Cette loi se rencontre notamment dans les processus de désintégration (exercice), les processus à "décharge" (neurones, systèmes à seuils ...)

Elle apparaît aussi comme limite de la binomiale. (exercice)

$$C_{t-1}^{k-1}\, p^{k-1}\, q^{t-k} = \text{Prob}[\ k-1\ \text{`1' à } t-1\ ]$$

Dès lors, $P[\tau_k = t] = \text{Prob}[\ k-1\ \text{`1' à } t-1 \text{ et `1' au temps } t\ ]$

Les tirages étant indépendants :

$$P[\tau_k = t] = \text{Prob}[\ k-1\ \text{`1' à } t-1\ ]\ P[\ \text{`1' au temps } t\ ]$$

$$\Rightarrow \quad P[\tau_k = t] = C_{t-1}^{k-1}\, p^{k}\, q^{t-k} \qquad \boxed{t \geq k}$$

NB : Si l'on change l'origine des temps $t' = t-k$ alors τ'_k est le premier temps supérieur à k tel qu'on ait le `1'. La loi de cette variable aléatoire est :

$$\boxed{P[\tau'_k = n] = C_{n+k-1}^{k-1}\, p^{k}\, q^{n}} \qquad (2.6)$$

appelée loi binomiale négative ou loi de Pascal.

c) X est continue (X est non dénombrable)

Dans ce cas il y a peu de sens à calculer la probabilité $P[X=x]$ (on verra de nombreux cas où cette probabilité est nulle). On peut par contre calculer la probabilité que X appartienne à un intervalle donné. Un rôle particulier est donné aux probabilités du type $P[X \leq x] = P[X \in]-\infty, x[]$.
On appelle fonction de répartition de X, la fonction

$$\boxed{F_X(x) = P[X \leq x]} \quad (2.7)$$

F_X possède les propriétés suivantes, qui se déduisent aisément des propriétés de base des probabilités (exercice):

i) F_X est croissante

ii) $F_X(x) \underset{x \to -\infty}{\longrightarrow} 0$; $F_X(x) \underset{x \to +\infty}{\longrightarrow} 1$

iii) $P[X \in [a,b]] = F_X(b) - F_X(a)$

iv) $P[X > x] = 1 - F_X(x)$

Exercices: Tracer la fonction de répartition du jeu de dé, de la loi uniforme..

La propriété (iii) est particulièrement intéressante puisque:

$$\boxed{P[X \in [x, x+dx]] = F_X(x+dx) - F_X(x)} \quad (2.8)$$

Lorsque F_X est dérivable (ce qui n'est pas toujours le cas - cf infra) on appelle densité de X la fonction

$$\boxed{\rho_X(x) = \frac{dF_X(x)}{dx}} \quad (2.9)$$

en sorte que $P[X \in [x, x+dx]] \sim \rho_X(x)dx$.

Remarques
i) une densité est toujours ≥ 0
ii) [illegible] $\int \rho(x)dx = 1$

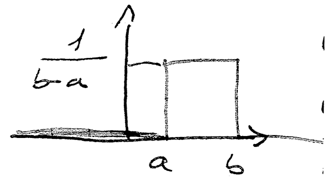
· Les v.a continues usuelles ont une densité et leur loi est donnée en terme de leur densité.

Exemples :

a) Loi uniforme :

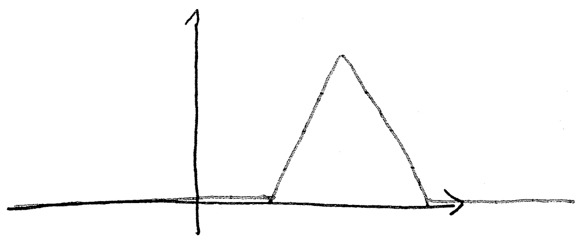
X est une variable uniforme sur l'intervalle $[a,b]$ si

$$(2.10) \qquad \boxed{\rho_X(x) = \begin{cases} \dfrac{1}{b-a} & ; x \in [a,b] \\ 0 & ; x \notin [a,b] \end{cases}}$$



b) Loi triangulaire (loi de Simpson)

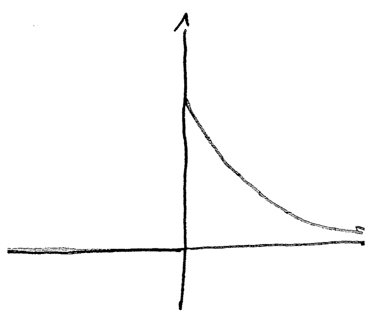
$$(2.11) \qquad \boxed{\rho_X(x) = \begin{cases} \dfrac{2}{b-a} - \dfrac{2}{(b-a)^2}|a+b-2x| & ; x \in [a,b] \\ 0 & x \notin [a,b] \end{cases}}$$



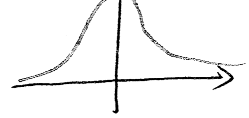
c) Loi exponentielle

X a une loi exponentielle de paramètre λ si

$$(2.12) \qquad \boxed{\rho_X(x) = \begin{cases} \lambda e^{-\lambda x} & ; \quad x \geq 0 \\ 0 & ; \quad x < 0 \end{cases}}$$



d) Loi normale ou loi de Gauss



$$\boxed{\rho_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp -\frac{1}{2}\frac{(x-m)^2}{\sigma^2}} \quad (2.13)$$

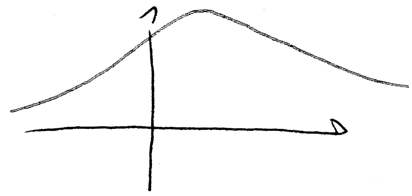
$x \in]-\infty, +\infty[$

m est la moyenne de X, et σ son écart-type (voir plus bas)

e) Loi de Cauchy

X admet une loi de Cauchy de paramètres (α, λ) si

$$\boxed{\rho_X(x) = \frac{1}{\pi}\,\frac{1}{\lambda^2 + (x-\alpha)^2}} \quad (2.14)$$



Contre exemple : On donne maintenant un exemple de variable aléatoire qui n'a pas de densité (voir annexe jointe)

En résumé, une loi de probabilité peut-être purement ponctuelle (cas d'une v.a. discrète), absolument continue (cas d'une v.a continue à densité), ou singulière continue (distribution fractale)

2.3) Caractéristiques des lois de probabilités :

A partir de la loi de probabilité de X on définit un certain nombre de quantités importantes.

a) Moments

On appelle moyenne de X ou espérance de X la quantité notée traditionnellement $\overline{X}$ ou $\langle x \rangle$ ou $E[X]$, est définie, lorsqu'elle existe, par :

Cas discret :

$$\boxed{E[X] = \sum_k x_k \, P(x_k)} \qquad (2.15)$$

où x_k sont les réalisations possible de X. Il peut y en avoir une infinité et dans ce cas l'espérance est donnée par une série

Cas absolument continu

$$\boxed{E[X] = \int x \, \rho_X(x) \, dx} \qquad (2.16)$$

où l'intégrale est effectuée sur le support de P_X (un point x est dans le support de P_X si pour tout ouvert O contenant x, $P_X[O] > 0$)

Cas singulier continu : (voir annexe)

Propriété : i) L'espérance est linéaire $E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$

ii) L'espérance d'une constante est cette constante

Plus généralement on définit de la même façon la moyenne d'une fonction de X, $f(x)$:

cas discret
$$\boxed{E[f(x)] = \sum_k f(x_k)\, P_X(x_k)} \quad (2.17)$$
continu
$$\boxed{E[f(x)] = \int f(x)\, \rho_X(x)\, dx}$$

Un cas particulier de l'équation précédente permet de définir les moments d'ordre k

$$(2.18) \quad \boxed{M_k = \langle x^k \rangle = E[X^k]} \qquad k \in \mathbb{N}.$$

La moyenne est le moment d'ordre 1. Le moment d'ordre 0 est la normalisation $\int \rho(x)\,dx = 1$. A partir des moments on construit quelques quantités fondamentales :

Variance
$$V_X = \langle x^2 \rangle - \langle x \rangle^2 \quad (2.18)$$
Ecart type
$$\sigma_X = \sqrt{V_X}$$

Exercice : Montrer que $V_X = E\left[(X - E(X))^2\right]$, i.e. la variance mesure l'écart moyen à la moyenne. En déduire que $V_X \geq 0$. Que peut-on dire d'une v.a. de variance nulle ?

Coefficient d'asymétrie

$$\boxed{a = \frac{E\left[(X - \langle x \rangle)^3\right]}{\sigma_X^3}} \quad (2.19)$$

mesure l'asymétrie.
nul si ρ_X symétrique

Coefficient d'aplatissement :

$$\boxed{A = \frac{E\left[(X - \langle x \rangle)^4\right]}{\sigma_X^4} - 3} \quad (2.20)$$

Comparaison de l'aplatissement par rapport à la loi normale ($A = 0$)
$A \not\geq 0$ plus aplatie
$A > 0$ moins "

On définit , les moments centrés en retranchant sa moyenne à la v.a X (X est alors dite centrée)

$$\rightarrow \quad \boxed{E\left[(X-\langle X\rangle)^k\right]} \qquad (2.21)$$

Exercice : Calculer les moments d'ordre k pour toutes les lois vues : ; . Montrer en particulier que.

Loi binomiale $\quad \boxed{E[X] = Np \; ; \quad V_X = Npq} \qquad (2.22)$

Loi de Poisson $\quad \boxed{E[X] = V_X = \lambda} \qquad (2.23)$

Distribution de Gauss $\quad \boxed{E[X] = m \; ; \quad V_X = \sigma^2} \qquad (2.24)$

Que peut-on dire des moments de la loi de Cauchy ?

b) Fonctions génératrices

Les moments ainsi que d'autres quantités peuvent être calculés à partir de fonctions appelées fonctions génératrices.

→ Fonction caractéristique

$$\boxed{\varphi_X(t) = E[e^{itX}]} \qquad (2.25)$$

Dans le cas continu cette fonction s'écrit donc

$$\boxed{\varphi_X(t) = \int_{-\infty}^{+\infty} e^{itx} \rho_X(x)\,dx} \qquad (2.26)$$

et c'est donc la transformée de Fourier de ρ_X. Dans le cas discret c'est une série de Fourier.

Si l'on développe e^{itX} en série entière on a :

$$\varphi_X(t) = E\left[\sum_{k=0}^{+\infty} \frac{(itX)^k}{k!}\right] = \sum_{k=0}^{+\infty} \frac{(it)^k}{k!} E[X^k]$$

$$\Rightarrow \quad \varphi_X(t) = \sum_{k=0}^{+\infty} \frac{(it)^k}{k!} m_k$$

Par ailleurs si $\varphi_X(t)$ est développable en série entière :

$$\varphi_X(t) = \sum_{k=0}^{+\infty} \frac{t^k}{k!} \left.\frac{d^k\varphi_X}{dt^k}\right|_{t=0}$$

Par identification on en déduit :

$$\boxed{m_k = \frac{1}{i^k} \left.\frac{d^k\varphi_X}{dt^k}\right|_{t=0}} \qquad (2.27)$$

On a en particulier :

$$m_0 = \varphi_x(0) = 1$$

$$m_1 = E[X] = \frac{1}{i} \left.\frac{d\varphi_x}{dt}\right]_{t=0}$$

$$m_2 = E[X^2] = - \left.\frac{d^2\varphi_x}{dt^2}\right]_{t=0}$$

De manière générale, les moments sont générés en dérivant $\varphi_x(t)$ ou alors en développant φ_x en série entière et en identifiant les termes de la série.

Exercice : Calculer les fonctions caractéristiques des lois suivantes : Tirage de Bernoulli ; jeu de dé ; loi binomiale ; loi de Poisson ; loi uniforme, exponentielle, Gaussienne, de Cauchy.

→ Fonction génératrice

La fonction génératrice est

$$\boxed{Z_x(t) = E[e^{tX}]} \quad (2.28)$$

C'est donc l'analogue de $\varphi_x(t)$ mais avec une exponentielle réelle. On a des propriétés analogues pour la génération des moments :

$$\boxed{m_k = \left.\frac{d^k Z_x}{dt^k}\right]_{t=0}} \quad (2.29)$$

Un exemple de fonction génératrice est la fonction de partition (cf cours de Physique Statistique).

→ Fonction log-génératrice

Soit $\boxed{F_x(t) = \log Z_x(t)}$. On a :

$$\left.\frac{dF_X}{dt}\right|_{t=0} = \frac{1}{Z_X(0)} \left.\frac{dZ_X}{dt}\right|_{t=0} = E[X]$$

$$\left.\frac{d^2F_X}{dt^2}\right|_{t=0} = -\frac{1}{Z_X^2(0)}\left(\left.\frac{dZ_X}{dt}\right|_{t=0}\right)^2 + \frac{1}{Z_X(0)}\left.\frac{d^2Z_X}{dt^2}\right|_{t=0}$$

$$= E[X^2] - E[X]^2 = V_X$$

Plus généralement, la dérivée <u>k-ième</u> de F_X génère le cumulant d'<u>ordre k</u>.

<u>Exercice</u> : Écrire les cumulants d'ordre 3 et 4. Les lier aux coefficients d'asymétrie et d'aplatissement.

Un exemple de fonction log-génératrice est l'<u>énergie libre</u>. A noter que dans certains cas, $Z_X(t)$ peut s'annuler pour ~~cet~~ $t=0$. Dans ce cas l'~~énergie libre~~ a une singularité qui correspond à la non existence de moments d'ordre supérieur à un ordre donné.

<u>Exercice</u> : Loi de Puissance tronquée → voir feuille jointe.

c) <u>Entropie</u>

Entropie

I) Généralités

Remarque : Il existe différentes définitions de l'entropie, non toutes équivalentes.

1) Entropie thermodynamique (Clausius) : $dS = \frac{\delta Q}{T}$

2) Entropie statistique (Boltzmann-Gibbs) : $S = -k \log \Omega$

3) Entropie de Shannon (Information) : $S = -\sum_i p_i \log p_i$

+ entropie de Kullback, entropie de Kolmogorov-Sinaï, entropie topologique, etc...

Idée : En général, l'idée d'entropie est associée à la mesure du désordre d'un système, ou de l'incertitude qu'on a sur le résultat d'une expérience.

Exemples

(Thermodynamique) : → transfert d'~~une~~ chaleur d'une source chaude (+ désordonnée) vers source froide (plus ordonnée)

→ Second principe. Entropie (désordre) max. à l'équilibre

Théorie de l'information : Incertitude (entropie) avant le résultat d'une expérience → quantité d'information reçue une fois l'expérience faite

$$\Rightarrow \boxed{S = -I}$$

Théorie de la complexité : systèmes dynamiques

Dans un ordinateur les nombres sont codés sur N bits. Par exemple un entier > 0 est codé comme : $\sum_{k=0} s_k 2^k$ où $s_k \in \{0,1\}$

le nombre de séquences distinguables croît avec N comme

$$2^N = e^{N \log 2}$$

$\Rightarrow$ croissance exponentielle à vitesse $\log 2$ $\rightarrow$ mesure de la complexité. Dans le contexte des systèmes dynamiques ~~[illegible]~~ ce taux de croissance est aussi appelé <u>entropie topologique</u>

$\rightarrow$ taux de croissance exponentielle des séquences distinguables.

Théorie de l'information

On va se focaliser sur la définition de Shannon. L'entropie mesure l'incertitude sur un événement (résultat d'expérience) avant que l'expérience soit faite ou encore la quantité d'information reçue une fois l'expérience faite

1) Ex : Entropie d'un seul événement

On considère un événement A de probabilité $p(A)$. On veut mesurer la quantité d'information obtenue quand l'événement se produit c. a. d. définir une fonction H à valeurs dans $\mathbb{R}$ et ne dépendant que de $p(A)$ (Il serait plus difficile de définir une quantité dépendant directement de l'événement). Essentiellement, quelle(s) propriétés minimales doit satisfaire H ? la suivante : l'information obtenue par 2 expériences successives correspondant à des événements <u>indépendants</u> doit s'ajouter. C. à d. où A et B sont 2 événements indépendants :

$$H(P(A\cap B)) = H(P(A)\,P(B)) = H(P(A)) + H(P(B))$$

On montre alors qu'il n'existe qu'une famille de fonctions satisfaisant cette propriété, la famille $\lambda \log x$. $\lambda \in \mathbb{R}$.

Def : On appelle <u>entropie de Wiener</u> la quantité S, définie pour un seul événement, et telle que :

$$\boxed{S(p) = -\log p}$$

2) Entropie de Shannon et axiomes de Kinchin

On considère maintenant une expérience dont les résultats possibles sont les événements $A_1 \dots A_n$, de probabilité $p_1 \dots p_n$. On mesure l'information manquante par une fonction $S \equiv S(p_1 \dots p_n)$ qui satisfait aux axiomes suivants (Kinchin) :

1) <u>Symétrie</u> $S(\dots p_i \dots p_j \dots) = S(\dots p_j \dots p_i \dots)$

2) <u>certitude</u> $S(1 \dots 0) = 0$

3) $S(0, p_1 \dots p_n) = S(p_1 \dots p_n)$ un événement certain n'ajoute rien

4) <u>incertitude max</u> si équiproba $S\left(\frac{1}{n} \dots \frac{1}{n}\right) \geq S(p_1 \dots p_n)$

5) $S \geq 0$ et S fonction continue des p_i.

le 6ème axiome généralise la propriété d'additivité définie ci-dessus.

On a 2 expériences α, β dont les résultats possibles sont

$A_1 \dots A_n$; $B_1 \dots B_\ell$

On fait l'expérience α on obtient A_i. Quelle est l'incertitude ~~dan~~ le résultat de β sachant que A est arrivée ?

Par définition, c'est une fonction des proba conditionnelles $P(B_j | A_i)$

et c'est en fait $S(P(B_1|A_i), \dots S(B_\ell|A_i))$

avec : $P(B_j|A_i) = \dfrac{P(B_j \cap A_i)}{P(A_i)}$

L'incertitude totale sur les 2 expériences est mesurée par l'entropie

$S(P(A_1 \cap B_1), \dots P(A_i \cap B_j) \dots)$. Contrairement au cas précédent (1 événement) les événements ne sont pas supposés indépendants et l'entropie globale n'est pas la somme des 2 entropies. La généralisation est :

$$6) \quad S(\{P(A_i \cap B_j)\}_{i,j}) = S(\{P(A_i)\}_{i=1}^{n}) + \sum_{i=1}^{n} P(A_i) \, \underbrace{S(\{P(B_j|A_i)\})}$$

$S(\{P(A_i \cap B_j)\}_{i,j})$: entropie globale

$S(\{P(A_i)\}_{i=1}^{n})$: incertitude sur la première exp.

$P(A_i)$: proba. d'avoir le résultat A_i

$S(\{P(B_j|A_i)\})$: entropie cond. sachant que A_i est le résultat de la première exp.

Th : Il existe une unique fonction satisfaisant 1-6, l'entropie de Shannon :

$$\boxed{S(p_1 \dots p_n) = -k \sum_{i=1}^{n} p_i \log p_i}$$

(définie à une constante près)

C'est aussi l'entropie de Wiener moyenne.

Ex 1 On tire à pile ou face. A pile on associe 0, à face 1. On obtient ainsi une chaîne de bits aléatoires. On suppose que

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

L'entropie de Shannon est alors :

$$\boxed{S = -p \log p - (1-p) \log (1-p)}$$

Rq : L'entropie topologique (mesure de la complexité) est $\log 2 = S_h$

$$\Rightarrow \quad \boxed{S \leq S_h}$$

avec égalité ssi $p = 1/2$ (S_h est l'entropie du cas équiprobable)

$\rightarrow$ les 2 définitions ne coïncident pas.

Ex 2

On considère un alphabet $\mathcal{A} = \{a_1 \ldots a_d\}$ contenant d symboles. Connaissant les n premières lettres d'un mot, on cherche à connaître l'incertitude sur la lettre qui va suivre (mesure de la prédictibilité ou redondance du langage). Soit

$$P(a \mid a_1 \cdots a_n)$$

la proba que ce symbole soit a sachant que les n premiers sont $a_1 \ldots a_n$.

2.4] Espérance conditionnelle

On a défini plus haut la notion de probabilité conditionnelle. On a vu dans différents exemples qu'elle permettait de restreindre l'univers des possibles en "conditionnant" les événements considérés par un sous-ensemble d'événements particuliers correspondant, par exemple, à une hypothèse, au résultat d'une expérience, etc... Une notion corollaire (et en fait plus générale) et la notion d'espérance ~~c~~onditionnelle. Il existe en effet une définition très générale de l'espérance conditionnelle qui permet de retrouver la probabilité conditionnelle comme cas particulier. Cependant, dans un souci ~~pe~~ pédagogique on va commencer par des exemples simples où l'espérance conditionnelle se déduit de la probabilité conditionnelle.

Commençons par un exemple. Considérons un dé à 6 faces non truqué et X la variable aléatoire qui associe à chaque face un numéro $\{1..6\}$. La valeur moyenne de X est

$$E[X] = \sum_{i=1}^{6} x_i \, P_X(x_i) = \frac{1}{6} \sum_{i=1}^{6} i = \frac{7}{2}$$

On considère maintenant l'événement $A = \{X = 2 \text{ ou } X = 3\}$. La probabilité conditionnelle de X sachant A est :

$$P[X = x_i \mid A] = \frac{P_X[\{X = x_i\} \cap A]}{P_X(A)}$$

C'est à dire qu'on considère les événements tels que $X = x_i$ (il n'y en a qu'un ici) et prend l'intersection avec A. On étend donc naturellement la notion de probabilité conditionnelle vue plus haut au cas de variables aléatoires. La condition $\{X = x\} \cap A$ sélectionne 2 événements de probabilité non nulle : $\{2\}$, $\{3\}$ dont la probabilité est $1/6$. Par ailleurs

$P_X(A) = \frac{1}{3}$. On a donc finalement :

$$P[X=1|A] = P[X=4|A] = P[X=5|A] = P[X=6|A] = 0$$

et $\quad P[X=2|A] = P[X=3|A] = 1/2$

Cette probabilité est bien sûr normalisée. A partir de cette probabilité il est naturelle de définir une espérance, l'espérance conditionnelle de X sachant A. Dans le cas discret :

$$\boxed{E[X|A] = \sum x_i \, P_X(x_i|A)} \qquad (2.4.1)$$

Ainsi dans l'exemple traité : $E[X|A] = 2\times\frac{1}{2} + 3\times\frac{1}{2} = 5/2$

C'est la valeur moyenne de X lorsqu'on restreint l'univers en A, avec comme probabilité conditionnelle $P[\cancel{X}|A]$.

On peut aussi écrire :

$$\boxed{E[X|A] = \frac{1}{P(A)} \sum_{x_i \in A} x_i \, P_X(x_i)} \qquad (2.4.2)$$

puisque $P_X(\{X=x_i\} \cap A)$ est non nulle seulement si $x_i \in A$.

Pour une variable aléatoire continue, on procède de façon similaire, à ceci près que les événements du type $\{X=x\}$ sont de probabilité nulle. On utilise alors la fonction de répartition conditionnelle :

$$\boxed{F_X(x|A) = \frac{P[\{X<x\} \cap A]}{P(A)}} \qquad (2.4.3)$$

et si F_X est absolument continue on appelle densité conditionnelle la quantité :

$$\boxed{\rho_X(x|A) = \frac{dF_X}{dx}(\cdot|A)} \qquad (2.4.4)$$

L'espérance conditionnelle est :

$$\boxed{E[X|A] = \int x \, F_X(dx|A) \underset{F_X <<}{=} \int x \, \rho_X(x|A)\, dx} \qquad (2.5.5)$$

On peut également l'écrire sous la forme

$$\boxed{E[X|A] = \frac{1}{P(A)} \int_A X(\omega)\, P(d\omega)} \qquad (2.5.6)$$

<u>Exemple</u> : X est une v.a. uniforme sur $[0,1]$ et $A = \{\omega \mid X(\omega) \in [\frac{1}{4}, \frac{1}{2}]\}$

On a donc : $P(A) = 1/4$ et $E[X|A] = \frac{1}{(\frac{1}{4})} \int_{1/4}^{1/2} x\, dx = 4 \left[\frac{x^2}{2}\right]_{1/4}^{1/2}$

$\Rightarrow \quad E[X|A] = 3/8.$

Supposons maintenant qu'au lieu de conditionner par un événement, on ait un ensemble complet d'événements $E_1, \dots E_k$ (qui constituent donc une partition de Ω et qui sont 2 à 2 disjoints). Comme pour la formule de Bayes ces événements peuvent par exemple correspondre à un ensemble d'hy[pothèses], théories, conditions etc… qui permettent de "plaquer" une structure sur Ω. Cela peut aussi correspondre à un ensemble de conditions qui déterminent le résultat d'une expérience. Par exemple, on fait le jeu suivant. On tire un dé à 6 faces. Si le résultat est 1 ou 2 on avance de 3 cases, si c'est 3 ou 4 on recule de 4, si c'est 5 ou 6 on reste sur place. Soit $E_1 = \{1,2\}$, $E_2 = \{3,4\}$, $E_3 = \{5,6\}$ et Y la variable aléatoire correspondant au nombre de cases dont on a sauté. On a alors : $P[Y=3] = \frac{1}{3}$; $P[Y=-4] = \frac{1}{3}$; $P[Y=0] = 1/3$ et

$E[Y] = 3 \times \frac{1}{3} + \frac{1}{3} \times -4 + \frac{1}{3} \times 0 = -\frac{1}{3}$. Mais on peut également définir l'espérance conditionnelle de Y par rapport à chacun des év[énements]

E_n

L'objectif de cet exemple

nelle par rapport à un ensemble d'évènements, préalable à la notion d'espérance conditionnelle d'une variable aléatoire par rapport à une autre.

Soit donc une tribu engendrée par un ensemble complet d'évènements $E_1 \dots E_k$ (i.e. $\mathcal{F}$ est la plus petite tribu contenant les E_ℓ), on appelle espérance conditionnelle de Y par rapport à $\mathcal{F}$ la variable aléatoire, notée $E[Y|\mathcal{F}]$ telle que :

$$\boxed{\begin{aligned} E[Y|\mathcal{F}] &= E[Y/E_k] \quad \text{si } \omega \in E_\ell \\ \Leftrightarrow \quad E[Y|\mathcal{F}](\omega) &= E[Y/E_\ell]\, \chi_{E_\ell}(\omega) \end{aligned}} \qquad (2.4.7)$$

où $\chi_{E_\ell}(\omega)$ est la fonction indicatrice de E_ℓ (elle vaut 1 si $\omega \in E_\ell$ et 0 autrement)

<u>Exemple</u> : On tire un premier dé, soit X le résultat. Si $X \in E_1 = \{1,$ alors on retire le dé et on avance du nombre de case indiqué. Si $X \in E_2 = \{$ on retire le dé et on recule du nombre de case indiqué. Enfin si $X \in E_3 = \{5,$ On reste sur place. Soit Y le nombre de cases dont on s'est déplacé.

On a donc : $E[Y/E_1] = 7/2$; $E[Y/E_2] = -7/2$; $E[Y/E_3] = 0$

Soit $z_i = E[Y/E_i]$. On note $Z = E[Y|\mathcal{F}]$ pour simplifier.

La loi de Z est par définition :

$$\boxed{P[Z = z_i] = P[\{\omega \in E_i\}] = P(E_i)} \qquad (2.4.8)$$

et sa valeur moyenne est

$$E[Z] = \sum_\ell z_\ell \, P[Z = z_\ell] = \sum_\ell P(E_\ell)\, E[Y/E_\ell]$$

En vertu de (2.4, 6)

$$E[\cancel{Z}] := \sum_k \int_{E_k} Y(\omega) P(d\omega) = \int_{\cup E_k} Y(\omega) P(d\omega) = E[Y]$$

où on a utilisé que $\mathcal{F}$ est complet. On vient donc de démontrer la propriété importante

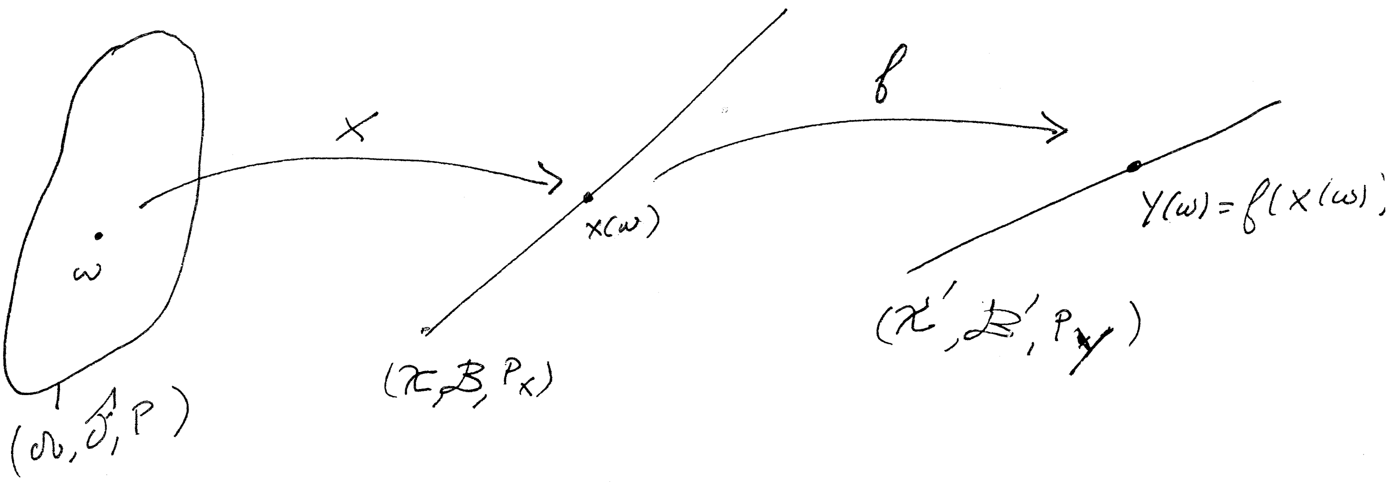
$$\boxed{E\left[E[Y|\mathcal{F}]\right] = E[Y]} \qquad (2.4.9)$$

Soit maintenant X une v.a. discrète prenant un nombre fini de valeurs, $x_1, x_2 - x_N$. On appelle $E_i = \{\omega \mid X(\omega) = x_i\}$ et $\mathcal{F}_X$ la tribu engendrée par les E_i. Soit Y une autre v.a. telle que $\mathcal{F}_X \subset \mathcal{F}_Y$. Alors on appelle espérance conditionnelle de $Y|X$ la v.a. $E[\cancel{X}|\mathcal{F}]$

On a vu plus haut qu'une variable aléatoire X était une fonction d'un espace probabilisé $(\Omega, \mathcal{T}, P)$ dans un autre espace probabilisé $(\mathcal{X}, \mathcal{B}, P_X)$ où $\mathcal{X}$ est un ensemble de nombres (par exemple $\mathcal{X} \subset \mathbb{N}$ ou $\mathcal{X} \subset \mathbb{R}$). On a également vu que la loi de la variable aléatoire X, P_X, est induite par P par la relation :

$$\boxed{P_X(A) = P[\{\omega \mid X(\omega) \in A\}]} \qquad (2.5.1)$$

On se place maintenant dans le cas où X est continue ($\mathcal{X} \subset \mathbb{R}$). Dans ce cas $\mathcal{B}$ est la tribu des boréliens. Considérons maintenant une fonction $f : \mathcal{X} \to \mathcal{X}' \subset \mathbb{R}$, dérivable et soit la variable $Y = f(X)$.



Si f est continue elle transporte l'espace probabilisé $(\mathcal{X}, \mathcal{B}, P_X)$ dans l'espace probabilisé $(\mathcal{X}', \mathcal{B}', P_Y)$ et Y est une nouvelle variable aléatoire dont la loi est :

$$P_Y(A') = P[\{\omega \mid f(X(\omega)) = Y(\omega) \in A'\}]$$

$$\boxed{P_Y(A') = P_X[\{x \mid f(x) \in A'\}]} \qquad (2.5.2)$$

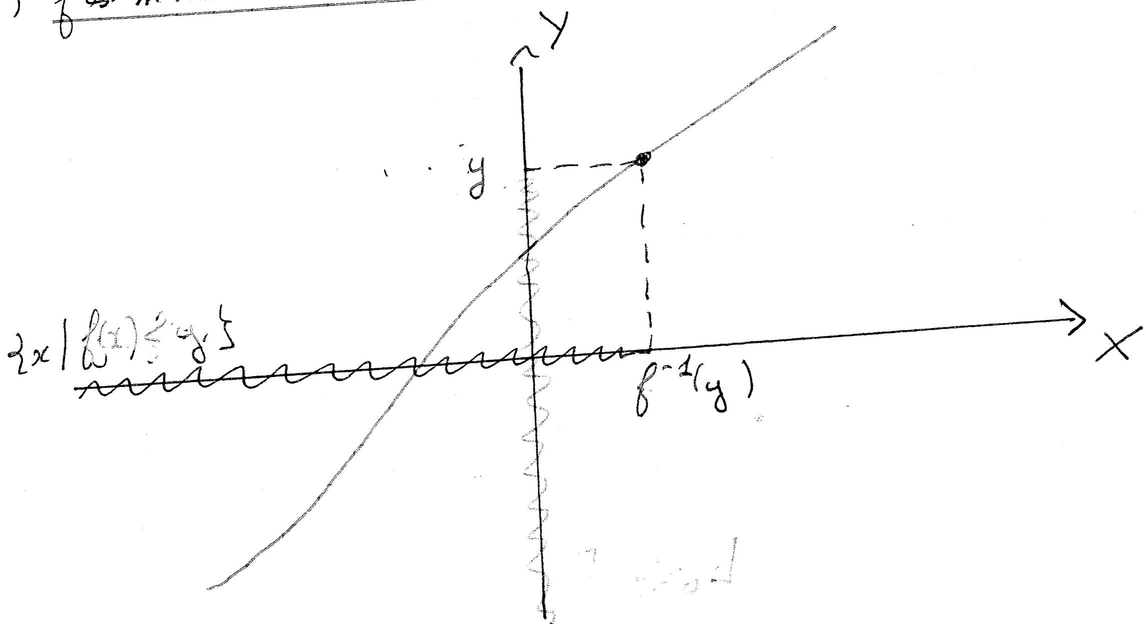
Dans la deuxième égalité on a "oublié" la référence explicite à

et on a pris comme espace de référence $\mathcal{X}$, l'espace des valeurs de X.
A partir de la définition de P_Y, on peut calculer la fonction de répartition de Y, F_Y :

$$\boxed{F_Y(y) = P_X[\{x \mid Y(x) < y\}] = P_X[\{x \mid f(x) < y\}]} \quad (2.5.3)$$

On va maintenant considérer plusieurs cas :

i) f est monotone croissante :



Dans ce cas $f(x) < y \Leftrightarrow x < f^{-1}(y)$, donc :

$$F_Y(y) = P_X[x < f^{-1}(y)] = F_X(f^{-1}(y))$$

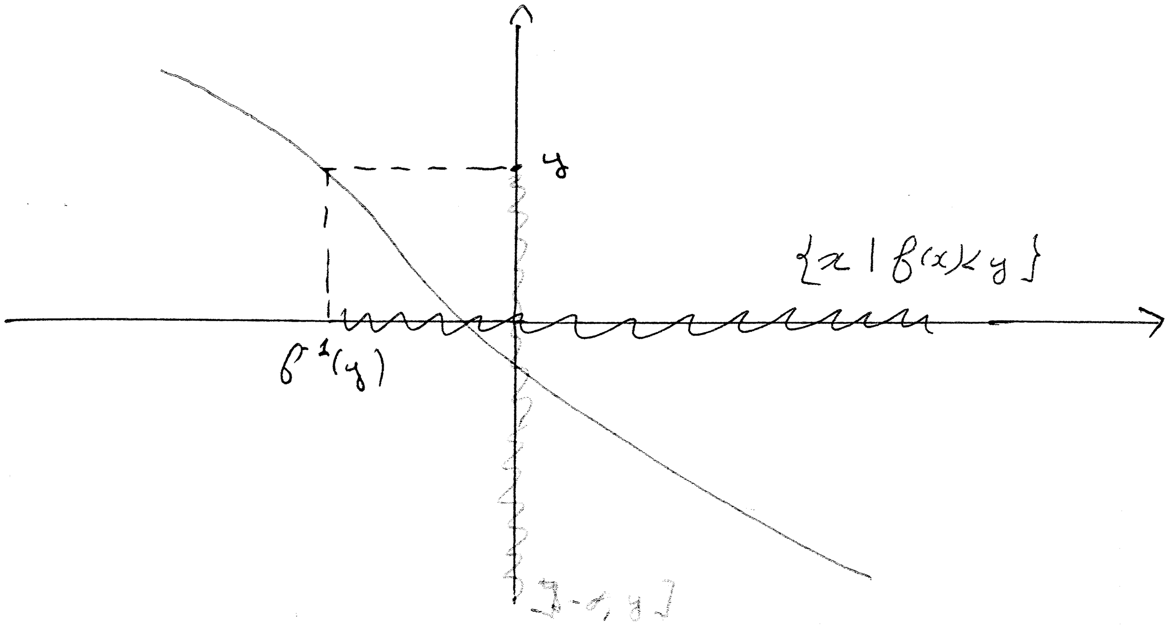
La densité de Y est :

$$\rho_Y(y) = \frac{dF_Y}{dy} = \frac{d}{dy} F_X(f^{-1}(y)) = F_X'(f^{-1}(y)) \frac{d}{dy} f^{-1}(y)$$

$$\boxed{\rho_Y(y) = \frac{\rho_X(f^{-1}(y))}{f'(f^{-1}(y))}} \quad (2.5.4)$$

qui nous donne donc explicitement ρ_Y en fonction de ρ_X et f.

ii) f est monotone décroissante



Dans ce cas $f(x) < y \Rightarrow x > f^{-1}(y)$ donc

$$F_Y(y) = P_X[x > f^{-1}(y)] = 1 - P[x \leq f^{-1}(y)] = 1 - F_X[f^{-1}(y)]$$

donc : $$\rho_Y(y) = F'_Y(y) = -\frac{d}{dy} F_X[f^{-1}(y)] = -\frac{\rho_X(f^{-1}(y))}{f'(f^{-1}(y))}$$

mais f est décroissante donc $-f' = |f'|$. On peut alors regrouper le cas i, ii dans une même équation :

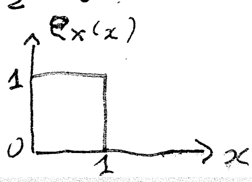
$$\boxed{\rho_Y(y) = \frac{\rho_X(f^{-1}(y))}{|f'(f^{-1}(y))|}} \qquad (2.5.5)$$

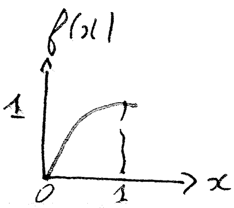
Exemple :

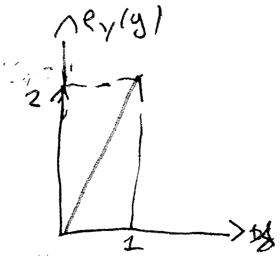
i) X uniforme sur $[0,1]$; $Y = \sqrt{x} = f(x) \in [0,1]$; $f^{-1}(y) = y^2$

$f'(x) = \frac{1}{2} x^{-1/2}$

$$\rho_Y(y) = \frac{\rho_X(y^2)}{\frac{1}{2}(y^2)^{-1/2}} = 2y$$

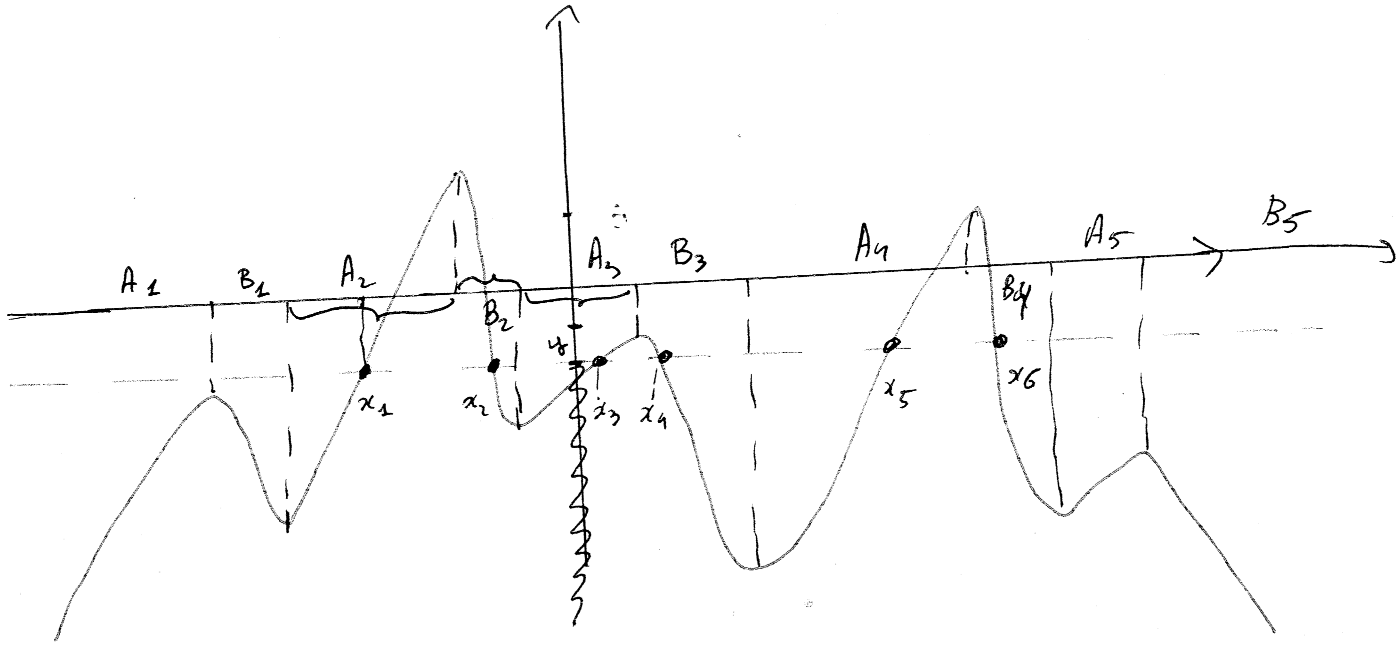






iii) f est quelconque

Dans le cas général, le support de f peut être décomposé en domaines A_k où elle est croissante alternant avec des domaines B_k où elle est décroissante (NB : on peut avoir une infinité (dénombrable) de tels domaines)



Par définition, sur chacun des segments A_k ou B_k, y a au plus une préimage notée x_j. On est alors ramené au cas précédent sur chacun des segments. Plus précisément :

$$\{x \mid f(x) < y\} = \bigcup_k \left[\left(A_k \cap \{x \mid f_{A_k}(x) < y\}\right) \cup \left(B_k \cap \{x \mid f_{B_k}(x) < y\}\right)\right]$$

où f_{A_k} est la restriction de f à A_k. Quand on calcule la probabilité de $\{x \mid f(x) < y\}$ on a affaire à une union de domaines disjoints donc

$$P[\{x \mid f(x) < y\}] = \sum_k \left(P[A_k \cap f_{A_k}(x) < y] + P[B_k \cap \{x \mid f_{B_k}(x) < y\}] \right)$$

qui est une somme sur des intervalles qui nous ramènent aux cas (i), (ii).
Plus précisément, soit l'intersection $C \cap f_C(x) < y$ (où $C = A_k, B_k$) est vide et elle ne contribue pas à la somme, soit l'intersection est non vide, dans ce cas il y a un et un seul $x_i \in C$ tel que $f(x_i) = y$. Dans ce cas si C

est un intervalle A_ℓ (où f est croissante) alors $A_\ell \cap \{x \mid f_{A_\ell}(x) < y\} = \{x \in A_\ell, x \leq x_j\}$. Si c'est un B_ℓ alors $B_\ell \cap \{x \mid f_{B_\ell}(x) < y\} = \{x \in B_\ell, x \geq x_j\}$. Si on note $A_\ell = [a_\ell^-, a_\ell^+]$, $B_\ell = [b_\ell^-, b_\ell^+]$ on a finalement

$$P[f(X) < y] = \sum_{\substack{x_j \\ f(x_j)=y}} \left(P[X \in [a_\ell^-, x_j]]\chi(x_j \in A_\ell) + P[X \in [x_j, b_\ell^+]]\chi(\ldots \right.$$

où χ est la fonction indicatrice : elle vaut 1 si l'événement est réalisé et 0 autrement. Autrement dit, dans la somme précédente on a un seul terme non nul pour chaque x_j. Enfin, par définition de la fonction de répartition, $P[X \in [a_\ell^-, x_j]] = F_X(x_j) - F_X(a_\ell^-)$ et $P[X \in [x_j, b_\ell^+]] = F_X(b_\ell^+) - F_X(x_j)$. Finalement donc :

$$F_Y(y) = \sum_{\substack{x_j \\ f(x_j)=y}} \left(F_X(x_j) - F_X(a_\ell^-)\right)\chi(x_j \in A_\ell) + \left(F_X(b_\ell^+) - F_X(x_j)\right)\chi(x_j \in B_\ell)$$

et en dérivant par rapport à y.

$$\boxed{\rho_Y(y) = \sum_{\substack{x_j \\ f(x_j)=y}} \frac{\rho_X(x_j)}{|f'(x_j)|}} \qquad (2.5.6)$$

Exercice : Vérifier qu'on retrouve bien les cas ici.

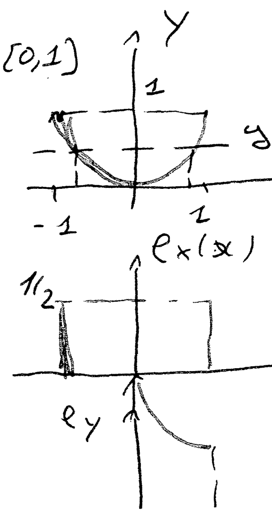
Exercice : X uniforme sur $[-1, 1]$, $Y = X^2$; $f(x) \in [0,1]$

Pour $y \in [0, 1]$ on a 2 préimages $x = \pm\sqrt{y}$

$f'(x) = 2x \Rightarrow$

$$\rho_Y(y) = \frac{1}{2}\frac{1}{|2\times -\sqrt{y}|} + \frac{1}{2}\frac{1}{(2\times\sqrt{y})}$$

$$\rho_Y(y) = \frac{1}{2\sqrt{y}} \longrightarrow$$ Vérifier la normalisation



Remarques

1) Si on veut calculer les moments de Y (ou plus généralement l'espérance d'une fonction $\varphi(Y)$) on peut utiliser soit la loi de Y, soit la loi de X puisque (théorème de la mesure image)

$$\boxed{E[\varphi(Y)] = \int_y \varphi(y)\,\rho_Y(y)\,dy = \int_x \varphi(y(x))\,\rho_X(x)\,dx} \quad (2.5.7)$$

Il peut être parfois plus avantageux d'utiliser la loi de X (le calcul est plus simple).

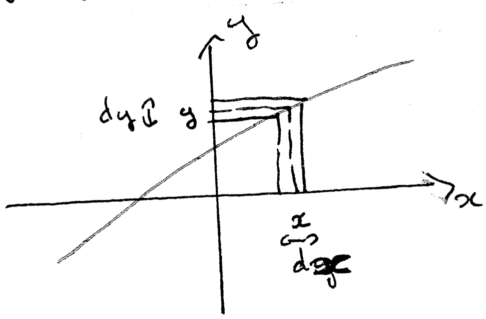
Exemple : X uniforme sur $[0, 2\pi]$, $Y = \cos(X) \Rightarrow \langle Y^k \rangle = \langle \cos^k(X) \rangle$

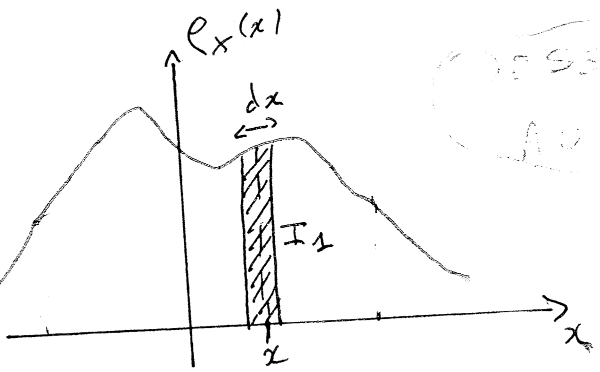
2) Dans le cas où f est monotone, la transformation $Y = f(X)$ est . Dans ce cas, on a :

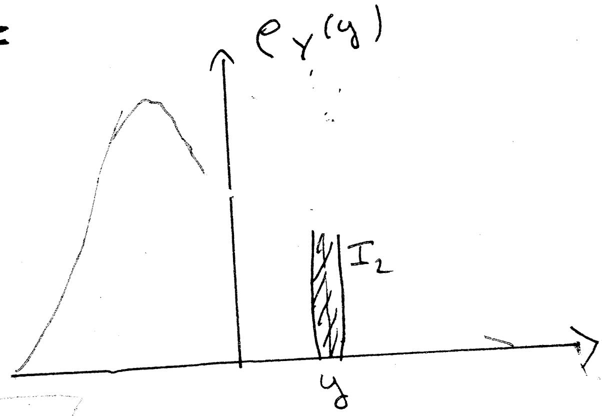
$$\boxed{\rho_X(x)\,dx = \rho_Y(y)\,dy}$$

Comme $\rho_X(x)\,dx$ est la probabilité que $X \in [x, x+dx]$, cette égalité traduit la conservation de la probabilité locale par chgt de variable.
Elle se déduit en fait trivialement de la relation $P[Y \in [y, y+dy]]$
$= P[\{X \mid f(x) \in [y, y+dy[\,] = P_X[f^{-1}([y, y+dy))] \underset{\text{inversibilité}}{=} P_X[x, x+dx]$

avec $y = f(x)$.







$\boxed{I_1 = I_2}$

3) On appelle opérateur de Perron-Frobenius l'opérateur qui associe à la densité ρ_X la densité ρ_Y par l'application f ($Y = f(X)$),

On a donc

$$\boxed{P_f \rho(y) = \sum_{x_j : f(x_j) = y} \frac{\rho(x_j)}{|f'(x_j)|}} \qquad (2.5.8)$$

On l'écrit aussi parfois sous la forme

$$\boxed{P_f \rho(y) = \int \delta(y - f(x))\, \rho(x)\, dx} \qquad (2.5.9)$$

où δ est la distribution de Dirac. On intègre ainsi sur les x qui ont pour image y, avec le poids ρ. L'emploi de la distribution δ est délicat puisque l'intégrale n'est pas simplement la somme $\Sigma \rho(x_j)$ comme on pourrait le penser au premier abord. Elle fait en plus intervenir la dérivée de f aux points x_j.

Exercice : Gaussienne sur la parabole (ex.)

L'opérateur de Perron-Frobenius est particulièrement important dans le cas suivant. On considère un système dynamique donné par $\boxed{x_{t+1} = f(x_t)}$. Dans de nombreux cas (cf infra) l'évolution est chaotique, c'est à dire qu'il n'est pas possible de prédire l'évolution d'une condition initiale après un certain temps sauf si l'on connaissait celle-ci avec une précision infinie, ce qui est concrètement impossible. On préfère alors décrire les propriétés statistiques de ce système. La description de systèmes dynamiques en termes statistiques porte le nom de théorie ergodique et c'est elle qui fait notamment le lien entre la description microscopique, dynamique de systèmes en physique et leur description macroscopique en termes d'observables moyennes (physique statistique).
Dans ce contexte, on est amené à se poser le problème suivant. On part

d'une distribution aléatoire de conditions initiales de densité ρ

on cherche à connaître l'évolution de cette densité par la dynamique.

(penser par exemple à l'évolution d'une goutte de Martini dans un verre de Vodka, lorsqu'on mélange à la cuillère et non au shaker) le processus associé est chaotique).

L'évolution au bout du temps t est donnée dans notre cas par l'itéré t-ième de f, $f^t = \underbrace{f \circ f \circ \cdots f}_{t}$ et la densité image est

$$\rho_t = Pf^t \rho$$

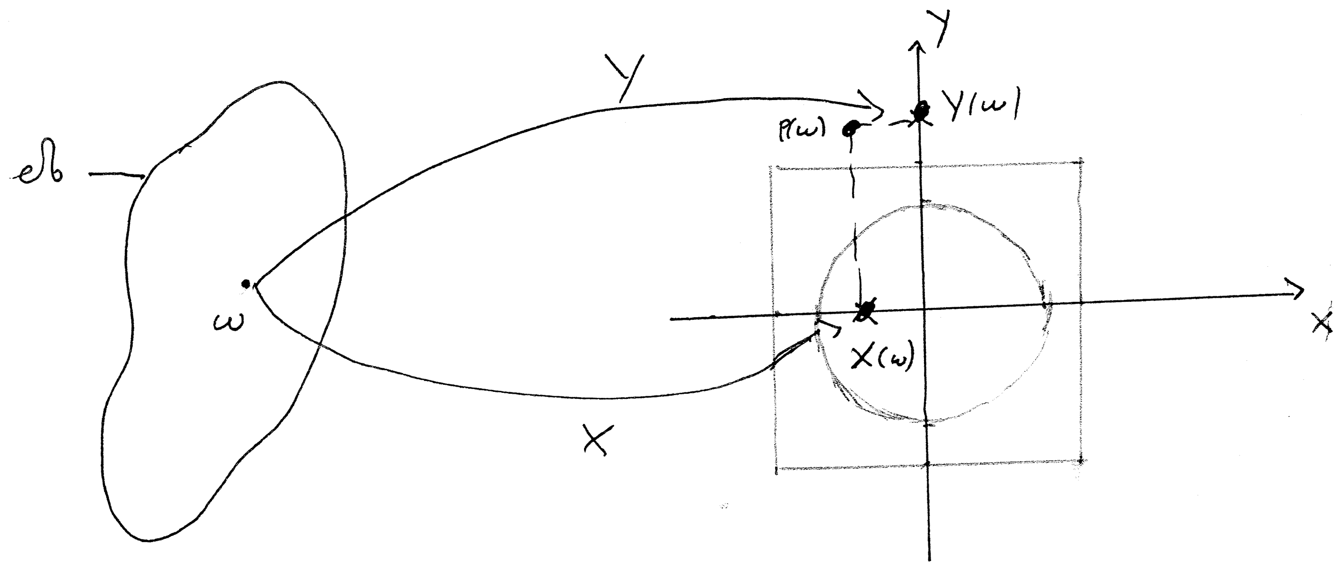
telle que :

$$\boxed{\rho_t(y) = \int \delta(y - f^t(x))\rho(x)\,dx}$$

<u>Exemple</u> : Multiplication par 2 mod 1 → voir feuille jointe

2.6) Systèmes de variables aléatoires

On s'est pour l'instant intéressé au cas d'une variable aléatoire. Cependant, dans de nombreux cas on a à traiter plusieurs variables aléatoires simultanément. Si par exemple on considère le point d'impact d'un projectile sur une cible, ce point a deux coordonnées x, y et il est donc paramétré par la donnée simultanée de 2 variables aléatoires X, Y



La position d'une particule à l'instant t requiert la donnée de 3 coordonnées d'espace X_t, Y_t, Z_t qui sont autant de variables aléatoires qu'il faut connaître simultanément. L'état d'un gaz (classique) est paramétré, au temps t, par 6N variables aléatoires, N étant le nombre de particules, et chaque particule étant déterminée par 3 coordonnées d'espace, et 3 de vitesse (ou d'impulsion)....

Comme pour le cas d'une variable aléatoire, on va considérer que les N variables aléatoires $X_1 \dots X_N$ qu'on considère simultanément, sont des fonctions d'un même état ω, appartenant à un espace probabilisé Ω. Cela nous permet de définir sans difficulté la notion de probabilité conjointe, sans introduire de notion vraiment nouvelle.

Considérons donc un ensemble de variables aléatoires $X_1 \dots X_N$ ou, de manière équivalente, un vecteur aléatoire $\vec{X} = \{X_i\}_{i=1}^{N}$. Chaque

X_i est donc une fonction d'un espace probabilisé $(\Omega, \mathcal{F}, P)$ dans un espace probabilisé $(\mathcal{X}_i, \mathcal{B}_i, P_{X_i})$. Soient $A_1 .. A_N$ un ensemble d'évènements $(A_i \in \mathcal{B}_i)$, la probabilité conjointe de $X_1 .. X_N$ est

$$\boxed{P_{\vec{X}}(A_1, .., A_N) = P[\{\omega \mid X_1(\omega) \in A_1, X_2(\omega) \in A_2, \ldots, X_N(\omega) \in A_N\}]} \quad (2.6.1)$$

NB : On ne discutera pas ici les problèmes de mesurabilité – il faut que l'image réciproque du domaine $A_1 \times A_2 .. \times A_N$ par $\vec{X}$ soit dans $\mathcal{F}$.

Un cas particulier immédiat mais néanmoins intéressant se produit lorsque les variables aléatoires sont indépendantes, i.e. les évènements $\{\omega \mid X_i(\omega) \in A_i\} = X_i^{-1}(A_i)$ sont 2 à 2 indépendants, $\forall$ les A_i. Dans ce cas en effet :

$$P_{\vec{X}}(A_1 .. A_N) = P[\{\omega \mid X_1(\omega) \in A_1\}]\, P[\{\omega \mid X_2(\omega) \in A_2\}] \ldots P[\{\omega \mid X_N(\omega) \in A_N\}]$$

Donc, dans ce cas :

$$\boxed{P_{\vec{X}}(A_1, .. A_N) = P_{X_1}(A_1) \ldots P_{X_N}(A_N)} \quad (2.6.2)$$

La probabilité conjointe est le produit des probabilités P_{X_i}, appelés, dans ce contexte probabilités marginales.

Pour simplifier, on va maintenant introduire quelques notions fondamentales dans le cas de 2 variables X, Y et on généralisera par la suite. On va considérer par ailleurs le cas de v.a continues. On appelle fonction de répartition conjointe la fonction :

$$\boxed{F_{XY}(x, y) = P[X < x, Y < y]} \quad (2.6.3)$$

La généralisation à N variables est bien sûr :

$$\boxed{F_{\vec{X}}(x_1, \dots x_N) = P[X_1 < x_1, \dots X_N < x_N]} \quad (2.6.4)$$

Par rapport à la définition donnée plus haut de la probabilité conjointe, la fonction de répartition dépend des événements A_i du type $]-\infty, x_i]$ (les événements $\{\omega \mid \vec{X}(\omega) \in]-\infty, x_1] \times]-\infty, x_2] \times \dots]-\infty, x_N]$ sont appelés rectangles).

La densité conjointe de X, Y est obtenue (lorsqu'elle existe) par dérivation de la fonction de répartition

$$\boxed{\rho_{XY}(x,y) = \frac{\partial^2 F_{XY}}{\partial x \partial y}(x,y)} \quad (2.6.5)$$

et la généralisation à N variables est :

$$\boxed{\rho_{\vec{X}}(\vec{x}) = \frac{\partial^N F_{\vec{X}}}{\partial x_1 \dots \partial x_N}(\vec{x})} \quad (2.6.6)$$

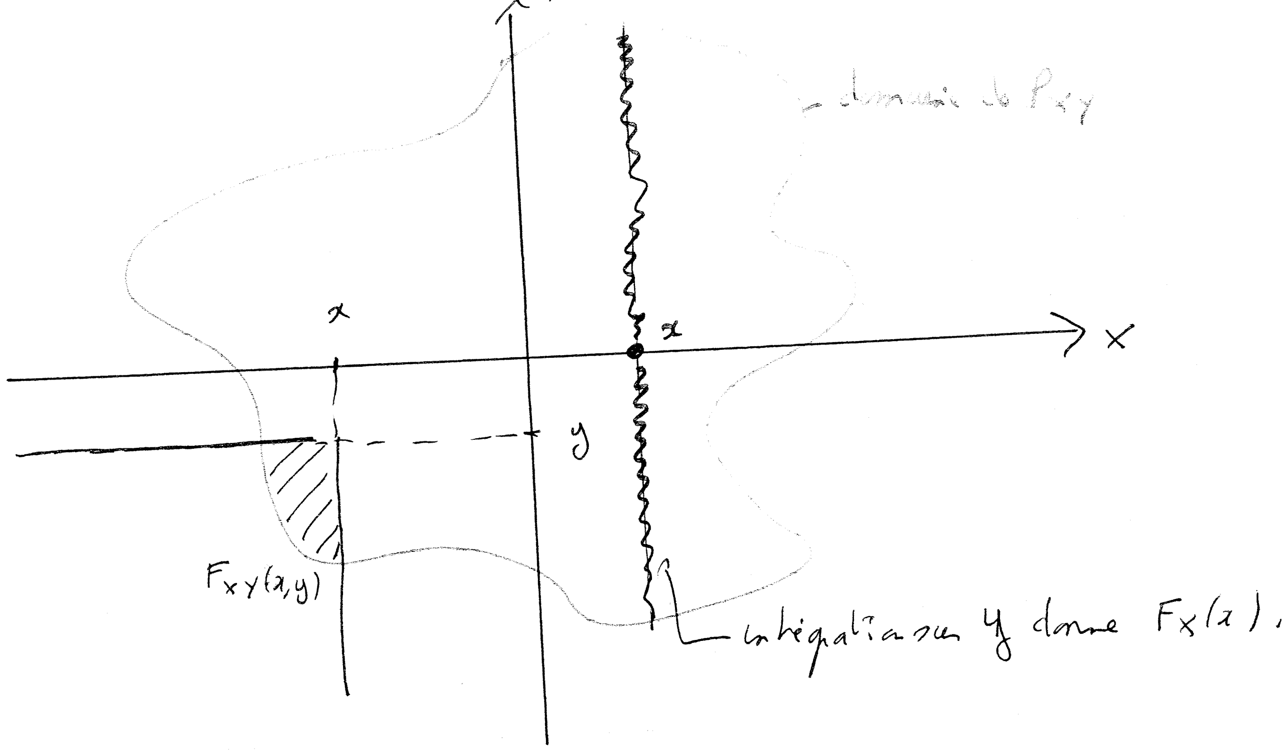
X, Y étant 2 v.a, ont chacune une loi de probabilité P_X, P_Y, de fonction de répartition F_X, F_Y. On a :

On obtient la de X (resp. Y) par intégration de la de probabilité conjointe par rapport à y (resp. x). Par cette raison, les lois P_X, P_Y sont appelés lois marginales.

Le dessin suivant illustre la situation.

$\rho_X(x)dx = P[X \in [x, x+dx[\,] = P[X \in [x, x+dx[, Y \in]-\infty, +\infty[\,]$

$= \left(\int_{-\infty}^{+\infty} \rho_{XY}(x,y)\,dy\right) dx$



Si l'on a plus de 2 variables (N) on a bien entendu N marginales à 1 variable, mais on a aussi des "marginales" à 2, 3 .. $N-1$ variables $P(X_1, X_2)$, $P(X_1, X_3)$, $P(X_1, X_2, X_4)$ etc...

Dans le cas de 2 v.a. indépendantes on a

$$\boxed{F_{XY}(x,y) = P[X<x]\,P[Y<y] = F_X(x)\,F_Y(y)} \qquad (2.6.7)$$

La fonction de répartition conjointe est le produit des fonctions de répartition.
Il en va de même pour la densité conjointe :

$$\boxed{\rho_{XY}(x,y) = \frac{d^2}{dx\,dy} F_{XY}(x,y) = \frac{d^2}{dx\,dy}\left(F_X(x) F_Y(y)\right) = \frac{d}{dx}F_X \frac{d}{dy}F_Y}$$

Soit :

$$\boxed{\rho_{XY}(x,y) = \rho_X(x)\,\rho_Y(y)} \qquad (2.6.8)$$

On utilisera par la suite cette propriété comme définition de l'indépendance

Déf : N v.a. sont indépendantes si leur densité conjointe est le produit des densités.

de la même façon [illegible]

définir l'espérance d'une fonction $f(X,Y)$ (resp. $f(X_1,X_2 - X_N)$), par

$$\boxed{E[f(X,Y)] = \int_{\mathbb{R}^2} f(x,y)\,\rho_{XY}(x,y)\,dx\,dy} \qquad (2.6.9)$$

Quelques cas particuliers importants

i) <u>Fonction de corrélation</u>:

$$\boxed{C(X,Y) = E[XY] - E[X]E[Y]} \qquad (2.6.10)$$

La fonction de corrélation généralise en quelque sorte la variance puisque $C(X,Y) = V_X$. <u>On dit que 2 v.a. sont non corrélées si $C(X,Y)=0$</u>

<u>Prop</u> : Deux v.a indépendantes sont non corrélées (mais la réciproque est fausse)

Lorsqu'on a plus de 2 variables, on peut calculer la corrélation de toutes les paires i, j (y compris $i=j$). On note $C_{ij} = C(X_i,X_j)$. On range les C_{ij} dans une matrice appelée <u>matrice de variance-covariance</u>. Cette matrice, C, a les propriétés suivantes

* La matrice C est symétrique
* Elle est donc diagonalisable dans une base orthogonale et ses valeurs propres sont réelles.
* $C_{ii} = V_{X_i}$

Les valeurs propres de C jouent un rôle fondamental dans le cas des <u>distributions gaussiennes</u> (cf TD).

ii) Fonctions génératrices

Soit $\vec{t}$ le vecteur $t_1 \dots t_N$ et $\vec{t}\vec{X} = t_1 X_1 + t_2 X_2 + \dots + t_N X_N$

On définit la fonction génératrice de $\vec{X}$ par :

$$\boxed{\varphi_{\vec{X}}(\vec{t}) = E\left[e^{i\vec{t}\vec{X}}\right]} \qquad (2.6.11)$$

Prop :

$$E\left[X_1^{n_1} X_2^{n_2} \dots X_N^{n_N}\right] = \frac{1}{i^{\Sigma n_i}} \frac{d^{n_1+n_2+\dots n_N}}{dt_1^{n_1} dt_2^{n_2} \dots dt_N^{n_N}} \varphi_{\vec{X}}(\vec{t})\Big|_{\vec{t}=0}$$

On définit de même la fonction génératrice

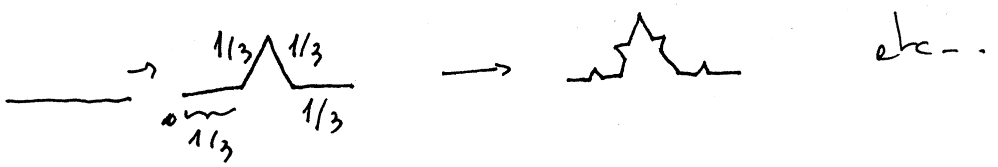
$$G_{\vec{X}}(\vec{t}) = E\left[e^{\vec{t}\vec{X}}\right]$$

Exemple important : La fonction de partition est la fonction génératrice de la loi conjointe appelée distribution de Gibbs. Pour un modèle d'Ising par exemple $\vec{t}$ joue le rôle d'un champ magnétique local et en dérivant par rapport aux champs locaux on obtient l'aimantation locale (dérivée d'ordre 1) ainsi que la susceptibilité (ordre 2) qui donne la réponse d'un spin à la perturbation d'un autre spin par un champ local infinitésimal. La susceptibilité est directement donnée par la fonction de corrélation correspondante, par le théorème de fluctuation - réponse. (cf TD)

Objets fractals

objets ayant une dimension non entière + propriétés d'invariance d'échelle.

Ex 1 : Courbe de Von Koch



Ex 2 : Ensemble de Cantor triadique

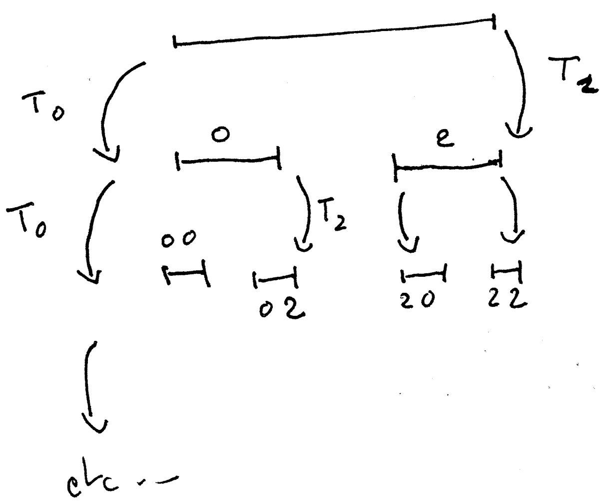
On se donne 2 applications $T_0(x) = \frac{1}{3}x$; $x \in [0,1]$

$T_2(x) = \frac{1}{3}x + \frac{2}{3}$; $x \in [0,1]$

On part de l'intervalle $I = [0,1]$

On calcule $\mathcal{T}(I) = T_0(I) \cup T_2(I)$

et les itérés de $\mathcal{T}$.



Ex 3 : Système de fonctions itérées (IFS)

1, 2 sont des exemples de IFS.

On se donne une famille de N contractions $T_i : \mathbb{R}^d \circlearrowleft$

c.a.d. $\sup_x \frac{\|T_i x\|}{\|x\|} < 1$

On calcule $\mathcal{T}(B) = \bigcup_{i=1}^{N} T_i(B)$ où B est un ensemble compact

On montre que $\mathcal{T}^n(B) \longrightarrow A$ où A est appelé <u>attracteur</u>

c'est un ensemble fractal.

<u>Ex</u> : Lorsque $n \to +\infty$, K_n courbe de Von Koch à l'étape n tend vers A, attracteur de Von Koch, qui est un fractal.

<u>Propriétés des fractales</u>

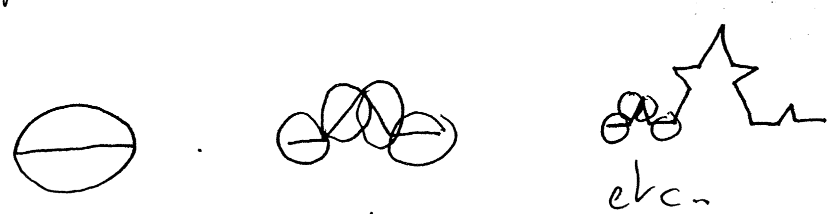
1) <u>Dimension non entière</u>

Comment "mesurer" un fractal ? par exemple l'ensemble de Von Koch.

Mesurons sa longueur. A l'étape 0 le segment est de longueur 1. A l'étape 1 on a 4 segments de longueur $\frac{1}{3}$, à l'étape 2 on a 16 segments de longueur $\frac{1}{9}$ etc — A l'étape n on a 4^n segments de longueur $\frac{1}{3^n}$ soit une longueur $\left(\frac{4}{3}\right)^n \xrightarrow[n \to +\infty]{} +\infty$.

La longueur est infinie !

Qu'elle est la surface de cet ensemble ? On recouvre les segments par des boules de diamètre $1/3^n$.



$n = 0$ $n = 1$ etc.

A l'étape n on a $\underline{4^n}$ boules de diamètre $\left(\frac{1}{3^n}\right)^2$ donc de surface $\frac{1}{4}\frac{\pi}{3^{2n}}$. La surface totale est donc $\frac{\pi}{4}\left(\frac{4}{3^2}\right)^n \to 0$

La surface est donc nulle !

On a donc un objet de longueur infinie et de surface nulle. Comment le mesurer ? $\longrightarrow$ On généralise le concept de longueur, surface, volume à une dimension non entière. On recouvre l'ensemble par des boules de diamètre $\frac{1}{3^n}$ et on définit la <u>mesure de Hausdorf</u> par:

$$\mu_\delta(K_n) = \sum_{k=0}^{n} \frac{1}{(3^n)^\delta} = \left(\frac{4}{3^\delta}\right)^n$$

où K_n est l'ensemble à l'étape n. Pour $\delta = 1$ on a la longueur, $\delta = 2 \to$ surface, $\delta = 3$ volume etc.. mais ici on considère que δ est réel. On montre que cette fonction est <u>croissante</u> avec δ. On appelle <u>dimension de Hausdorf</u> l'<u>unique</u> nombre d_H tel que

$$\mu_\delta(K_n) \to +\infty \qquad \delta < d_H$$

$$\mu_\delta(K_n) \to 0 \qquad \delta > d_H$$

$$0 < \mu_\delta(K_n) < +\infty \quad \forall n \qquad \delta = d_H$$

On voit ainsi que $\mu_\delta(K_n) = \exp n(\log 4 - \delta \log 3)$

tend vers une valeur finie (différente de 0) ssi

$$\boxed{\delta = d_H = \frac{\log 4}{\log 3}}$$

C'est la <u>dimension de Hausdorff</u> de l'ensemble de Von Koch.

Elle est <u>non entière</u>.

<u>Ex</u> : Montrer que la dimension de l'ensemble de Cantor est $d_H = \frac{\log 2}{\log 3}$

i.e. c'est un objet intermédiaire entre un segment et un ensemble dénombrable de points.

2) <u>Autosimilarité</u>

La définition d'un système de fonctions itérées implique que $T(A) = A$, c'est à dire que A est composé de copies de lui-même contractées par chacun des T_i. Il est donc "auto-similaire" puisqu'un zoom redonne la même structure.

3) <u>La fractale</u>

La mesure de Hausdorff permet de définir une probabilité.

Si B est un sous ensemble de A, A étant l'ensemble de Von Koch, Cantor etc.

alors on définit la probabilité de B comme.

$$P(B) = \frac{\mu_{d_H}(B)}{\mu_{d_H}(A)}$$

en sorte que $P(A) = 1$. Cette probabilité est autosimilaire. En effet, si l'on considère une petite boule de diamètre ε, $\mathcal{B}(x,\varepsilon)$ où $x \in A$, la probabilité de cette boule est :

$$P(\mathcal{B}(x,\varepsilon)) \sim \varepsilon^{d_H} \quad \text{par définition.}$$

Si l'on dilate la boule d'un facteur λ on a :

$$P(\mathcal{B}(x,\lambda\varepsilon)) \sim \lambda^{d_H} \varepsilon^{d_H} \sim \lambda^{d_H} P(B(x,\varepsilon))$$

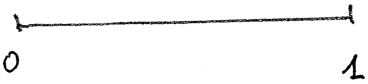
3) Forme de la fonction de répartition.

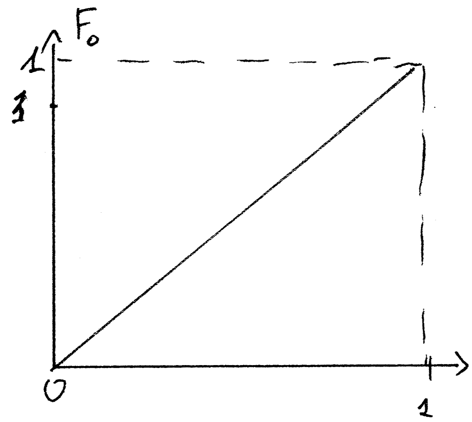
Qu'elle est l'allure de P ? Regardons par exemple la fonction de répartition de P, dans le cas de l'ensemble de Cantor. On a $F(x) = P([0,x[)$ ($=$ Proba $[X < x]$ où X est un point tiré au hasard). On va construire F itérativement. On part de la loi uniforme sur $[0,1] \Rightarrow F_0(x) = x$. A l'étape 1 on a des points seulement dans $[0,1/3] \cup [\frac{2}{3};1]$. La longueur totale est donc $2/3$. Pour définir une probabilité on doit diviser par ce facteur (normalisation) pour que $p([0,1/3] \cup [2/3,1]) = 1$. A cette étape $F_1(x) = \frac{3}{2}x$, $x \in [0,$

$F_1(x) = \frac{1}{2}$ si $x \in [1/3, 2/3[$ (pas de points $\Rightarrow F$ constante) et $F_1(x) = \frac{3x}{2} - \frac{1}{2}$

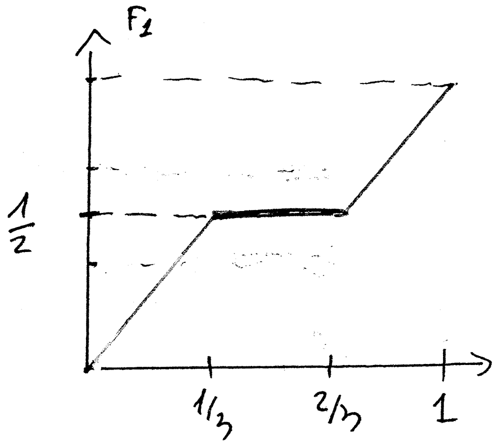
si $x \in [2/3, 1]$

0

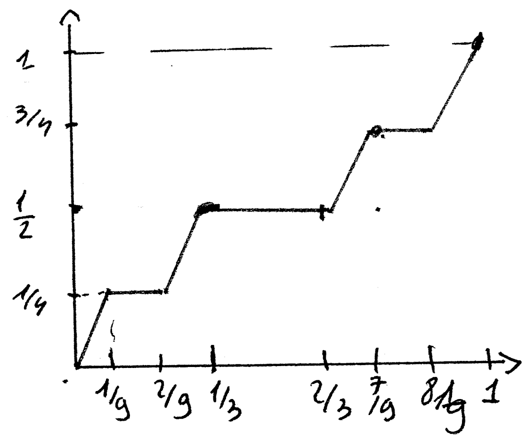




n = 1



n = 2



etc..

Que devient la courbe limite ? Est-elle régulière ? dérivable ?

Peut-on définir la densité ?

C'est une courbe fractale.

Chapitre 2

Convergences et théorèmes limites en théorie des probabilités.

I] Introduction

Comme on l'a évoqué dans le chapitre précédent, on a souvent à faire en théorie des probabilités, des passages à la limite. C'est bien sûr le cas en statistiques lorsqu'on essaie d'estimer la moyenne réelle d'une quanti[té] en faisant la moyenne empirique sur un grand nombre d'échantillons. La loi des grands nombres (cf infra) garantit, sous des conditions mathématiques précises, que la moyenne estimée "converge" (dans un s[ens] précisé plus bas) vers la moyenne réelle. On peut également avancer [une] estimation des corrections lorsque le nombre d'échantillons est très grand. Ce type de limite est aussi important en physique statistique. Il appara[ît] notamment dans la notion de limite thermodynamique : on fait tend[re] le nombre de "particules" vers l'infini selon des modalités précises qu[i] assurent l'existence d'un macroétat (état de Gibbs) dans la l[imite] ainsi que l'existence de grandeurs telles que l'énergie par particule, l'énergie libre par particule, etc... On fait aussi implicitement ap[pel] à un théorème limite (la loi des grands nombres) quand on rempla[ce] l'énergie instantanée (qui est une variable aléatoire) par sa valeur moyenne, ou lorsqu'on calcule la distribution des fluctuations autour de la valeur moyenne (théorème central limite). On approxime ainsi le système fini à très grand nombre de particules par un système caractérisé par des grandeurs qu'on obtient par passage à la limite

Un autre type de limite existe lorsqu'on s'intéresse à des régimes stationnaires. Ici c'est le temps qui tend vers l'infini. Un système préparé dans un état initial évolue vers un état asymp[totique] (à l'équilibre ou hors équilibre) lorsque $t \to +\infty$. On redesc[...]

de ce type de limite dans le chapitre suivant, consacré aux processus aléatoires.

Dans le présent chapitre on s'intéresse aux différents types de convergence et on énonce quelques théorèmes fondamentaux. Auparavant, nous allons commencer par un exemple simple mais néanmoins riche et très utile pour guider l'intuition dans les parties suivantes, plus abstraites.

1) Un exemple simple

On reprend l'exemple du tirage de Bernoulli. On considère donc une suite de N v.a. notées $X_1, \dots X_N$, prenant des valeurs binaires. Elles peuvent représenter une suite de tirages à pile ou face, une suite de bits aléatoires, mais aussi une chaîne de spins d'Ising sans interaction. Pour fixer les idées on prend $X_i \in \{0, 1\}$, $\forall i$.

On s'intéresse maintenant à la moyenne empirique :

$$(1.1) \quad \boxed{m_N = \frac{1}{N} \sum_{i=1}^{N} X_i}$$

(m_N mesure la fréquence empirique des 1 dans la chaîne)

m_N est une variable aléatoire. Dans le cas où $X_i \in \{0, 1\}$ elle prend les valeurs $\frac{k}{N}$; $k = 0$ à N, où k est le nombre de 1 apparaissant dans la séquence. Notons que si X_i représente un spin d'Ising m_N correspond à l'aimantation par spin.

Comme on l'a vu au chapitre précédent on a 2^N configurations possibles correspondant aux différentes réalisations de $\{X_1 \dots X_N\}$. On note $\Omega = \{0, 1\}^N$ l'espace de ces configurations et ω une configuration. On cherche maintenant le nombre de configurations ω qui donnent la valeur $\frac{k}{N}$ à $m_N(\omega)$. Ce sont les configurations qui ont k '1' et $N-k$ '0'. Leur nombre est donc $\underline{C_N^k}$

On remarque alors la propriété étonnante suivante. La proportion des configurations donnant la valeur $\frac{k}{N}$ pour m_N est donc $\frac{C_N^k}{2^N} = g_N(k)$.

Lorsque $N \to +\infty$ la formule de Stirling nous donne

$$\log g_N(k) \sim -(N-k)\log\left[1-\frac{k}{N}\right] - k\log\left(\frac{k}{N}\right) - \frac{1}{2}\log\left[2\pi k\left(1-\frac{k}{N}\right)\right]$$

$$- N\log 2$$

C_N^k est maximum pour $k = \frac{N}{2}$, donc $g_N(k)$ est maximum pour cette valeur. On a :

$$\log g_N\left(\frac{N}{2}\right) = +\frac{N}{2}\log(2) + \frac{N}{2}\log 2 - \frac{1}{2}\log\left[\frac{\pi N}{2}\right] - N\log 2$$

$$= \log\left[\frac{2}{\pi N}\right]$$

Donc, $\boxed{g_N\left(\frac{N}{2}\right) = \sqrt{\frac{2}{\pi N}} \to 0}$ (1.2)

Par conséquent, pour tout k, ~~la~~ la proportion des configurations telles que $m_N = \frac{k}{N}$ tend vers 0 quand $N \to +\infty$.

On remarque que ce résultat est purement combinatoire et qu'il ne fait nulle part intervenir la loi de probabilité de X_i (on ne l'a d'ailleurs pas encore explicitée).

On se donne maintenant la loi de probabilité de X_i. Ce sont des v.a.i.i.d avec $P[X_i = 0] = q$; $P[X_i = 1] = p$.

La probabilité que $m_N = k$ est :

$$P[m_N = k] = C_N^k\, p^k\, q^{N-k} = P_N(k)$$

$\boxed{\left(\begin{array}{c}\text{et bien sûr}\\ E[m_N] = p\end{array}\right)}$

On cherche maintenant à connaître le comportement de $P_N(k)$ quand $N \to +\infty$.
En utilisant de nouveau la formule de Stirling on obtient :

$$\log P_N(k) = \log C_N^k + k\log p + (N-k)\log q$$

$$= -(N-k)\log\left(1-\frac{k}{N}\right) - k\log\left(\frac{k}{N}\right) + k\log p + (N-k)\log q + \sigma_N(k),$$

avec $\sigma_N(k) = -\frac{1}{2}\log\left[2\pi k\left(1-\frac{k}{N}\right)\right]$

Finalement :

$$\boxed{\log P_N(k) = -(N-k)\log\left[\frac{1-k/N}{1-p}\right] - k\log\left[\frac{k/N}{p}\right] + \sigma_N(k)} \quad (1.3)$$

On introduit la fonction :

$$\boxed{I_p(z) = (1-z)\log\left[\frac{1-z}{1-p}\right] + z\log\left[\frac{z}{p}\right]} \quad (1.4)$$

On a donc :

$$\boxed{P_N(k) = \exp -N I_p(k/N) \times R_N(k)} \quad (1.5)$$

avec $\boxed{R_N(k) = \exp \sigma_N(k) = \frac{1}{\sqrt{2\pi k(1-k/N)}}}$ (1.6)

On étudie maintenant la fonction $I_p(z)$, $z \in [0,1]$.

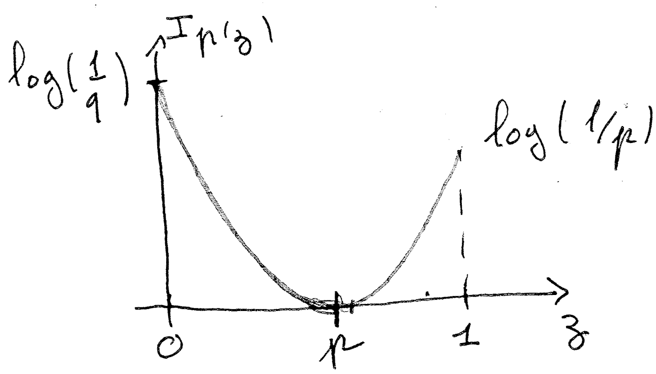
i) $I_p(z)$ est convexe pour $z \in [0,1]$

* $I_p'(z) = \log\left(\frac{z}{p}\right) + 1 - \log\left(\frac{1-z}{1-p}\right) - 1 = \log\left(\frac{z}{1-z}\frac{1-p}{p}\right)$

$I_p''(z) = \frac{1}{z} + \frac{1}{1-z} \geq 0$ *

ii) $I_p(z)$ est positive et atteint son minimum, 0, en $z = p$.

* $I_p'(p) = 0$; $I_p(p) = 0$; $I_p(0) = -\log(1-p) \geq 0$



Donc, quand $N \to +\infty$ $P_k(N) \to 0$ exponentiellement vite, sauf "où $\frac{k}{N} \to p$. Cela nous suggère que lorsque $N \to +\infty$ les configurations telles que $\frac{k}{N} \to p$ ont une probabilité 1 et toutes les autres une probabilité nulle".

On va rendre cette idée plus précise mathématiquement. On calcule la probabilité $\text{Prob}[|m_N - p| > \varepsilon]$ pour $\varepsilon > 0$. Par définition c'est $P[\{\omega \,|\, |m_N(\omega) - p| > \varepsilon]$ c'est à dire qu'on estime la probabilité de l'ensemble des configurations ω telles que $|m_N(\omega) - p| > \varepsilon$.

On a :

$$\text{Prob}[|m_N - p| > \varepsilon] = \sum_{k \in A_N(\varepsilon)} P_N(k) \quad \text{avec } \boxed{A_N(\varepsilon) = \{k \,|\, |m_N - p| > \varepsilon\}}$$

On :

$$\max_{k \in A_N(\varepsilon)} P_N(k) \leq \sum_{k \in A_N(\varepsilon)} P_N(k) \leq N \max_{k \in A_N(\varepsilon)} P_N(k)$$

En prenant le log et en divisant par N ~~et~~ en utilisant l'équation (1.5), on a (en notant k^* le point où $P_N(k)$ atteint son max sur A_N)

$$\underbrace{\frac{1}{N} \log \kappa_N(k^*)}_{\to 0} - I_p\left(\frac{k}{N}\right) \leq \frac{1}{N} \log \text{Prob}[|m_N - p| > \varepsilon] \leq \frac{\log N}{N} + \frac{1}{N} \log \kappa_N(k^*) - I_p\left(\frac{k^*}{N}\right)$$

Donc, asymptotiquement (Propriété de grande déviation).

$$(1.7) \quad \boxed{\text{Prob}[|m_N - p| > \varepsilon] \sim \exp - N I_p\left(\frac{k^*}{N}\right) = \exp - N \max_{k \in A_N(\varepsilon)} I_p\left(\frac{k}{N}\right)}$$

mais, par définition si $k \in A_N(\varepsilon)$, $\frac{k}{N} \neq p$, donc $I_p\left(\frac{k}{N}\right) \to 0$

On vient de démontrer, dans cet exemple particulier, la loi faible des grands nombres.

Loi faible des grands nombres

$$\boxed{\forall \varepsilon > 0, \quad \lim_{N \to +\infty} \mathrm{Prob}\left[|m_N - p| > \varepsilon \right] = 0.} \qquad (1.8)$$

Clairement, m_N est la fréquence des 1 apparaissant dans la séquence ω. On s'attend donc à ce que soit un estimateur de la fréquence réelle, p, c'est à dire que $m_N \to p$ quand $N \to +\infty$. Ce n'est toutefois pas ce que l'on vient de montrer ici. On a simplement montré que la probabilité que m_N s'écarte de plus de ε de p tend vers zéro. Ce type de convergence s'appelle la convergence en loi. On y reviendra dans la section suivante. La convergence en loi n'implique cependant pas que $\lim_{N \to +\infty} m_N = p$.

Tout d'abord il existe une infinité de séquences qui ne convergent pas vers p. En fait, pour tout $\alpha \in [0,1]$, on peut trouver une infinité de séquences ω telles que $m_N(\omega) \to \alpha$. Ainsi les séquences du type $\overline{01}$, $\overline{10}$, $\overline{0011}$, $\overline{1100}$ etc sont telles que $m_N(\omega) \to \frac{1}{2}$. De même les séquences $\overline{001}$, $\overline{010}$, $\overline{100}$ sont telles que $m_N(\omega) \to \frac{1}{3}$. Plus généralement, il suffit de faire en sorte que la fréquence empirique approche α quand $N \to +\infty$. Mais, par ailleurs, il existe une infinité de séquences qui ne convergent pas. Prendre par exemple une séquence constituée de K '0', puis K^K '1', puis K^{K^K} '0' etc... En conclusion il y a "très peu" de séquences qui convergent vers p (cf estimation (1.2) du début). On va néanmoins montrer maintenant que l'ensemble des séquences telles que $m_N(\omega) \to p$ est de probabilité 1.

Cela constitue la loi forte des grands nombres.

Loi forte des grands nombres

$$\boxed{P\left[\left\{\omega \mid \lim_{N\to+\infty} m_N(\omega) = p\right\}\right] = 1} \qquad (1.9)$$

Convergence presque-sûre

La démonstration utilise la propriété de décroissance exponentielle (ou propriété de grande déviation) et le lemme de Borel Cantelli vu en TD. (A faire en TD).

Il n'y a pas de paradoxe. Le résultat nous montre simplement que la probabilité (sur l'espace des configurations) ne "voit" que certaines configurations, celles justement telles que $m_N(\omega) \to p$ quand $N \to \infty$. Cependant ces configurations sont une "toute petite partie" de l'ensemble total des configurations. Par ailleurs, en changeant de proba, on voit d'autres configurations. C'est en fait exactement ce que l'on a en physique statistique lorsqu'on change la température T. En changeant celle-ci, la valeur moyenne observée pour l'aimantation, l'énergie etc.. change. Cela signifie simplement que la dynamique microscopique explore un nombre restreint de configurations celles qui sont typiques pour la probabilité de Gibbs à la température T.

On s'intéresse maintenant aux fluctuations de taille finie. En effet la loi des grands nombres nous renseigne sur la valeur limite que prend P-presque sûrement m_N, mais elle ne nous dit rien sur les fluctuations autour de la valeur asymptotique (p) lorsque N est grand, mais fini. Pour estimer ces fluctuations on va de nouveau utiliser la fonction $I_p(z)$. Pour cela, on va faire un développement limité de $I_p(z)$. L'idée est que puisque la proba est de type $\exp - N I_p$ les fluctuations sont essentiellement dominées par les valeurs de $m_N(\omega)$ qui sont proches de p.

On a :

$$I_p(z) = I_p(p) + (p-z) I'_p(p) + \frac{(p-z)^2}{2} I''_p(p) + o(p-z)^3$$

Or, $I_p(p) = I'_p(p) = 0$, donc pourvu que $I''_p(p) \neq 0$ les fluctuations sont dominées par la dérivée seconde de I_p.

On a : $I''_p(z) = \frac{1}{z} + \frac{1}{1-z} \Rightarrow I''_p(p) = \frac{1}{p} + \frac{1}{q} = \frac{p+q}{pq} = \frac{1}{pq}$

On remarque que pq est la variance de la loi de Bernoulli. Ce n'est pas un hasard.

On a donc, en vertu de (1.5)

$$P_N(k) = \text{Prob}\left[m_N = \frac{k}{N}\right] = \frac{1}{\sqrt{2\pi k\left(1-\frac{k}{N}\right)}} \cdot \exp - N I_p\left(\frac{k}{N}\right)$$

qui devient, au voisinage de p, et pour N grand

$$P[m_N = z] \simeq \frac{1}{\sqrt{2\pi N z(1-z)}} \exp - \frac{N}{pq}(p-z)^2$$

avec $z = \frac{k}{N}$. Au voisinage de p, $\frac{1}{\sqrt{z(1-z)}} \sim \frac{1}{\sqrt{pq}}$, donc :

$$\boxed{P[m_N = z] \sim \frac{1}{\sqrt{2\pi N pq}} \exp - \frac{N}{pq} \frac{(p-z)^2}{2}} \quad (1.10)$$

On peut réécrire le terme dans l'exponentielle comme : $\left(\frac{\sqrt{N}(z-p)}{\sqrt{pq}}\right)^2 = \left[N\left(\frac{z-p}{\sqrt{Npq}}\right)\right]^2$. Cela invite à définir une nouvelle variable aléatoire :

$$\boxed{Y_N = \frac{\sum_{i=1}^{N} X_i - Np}{\sqrt{Npq}} = N\left[\frac{m_N - p}{\sqrt{Npq}}\right]} \quad (1.11)$$

C'est une v.a centrée $E[Y_N] = \frac{N}{\sqrt{Npq}}\left(E[m_N] - p\right) = 0$ et si $m_N = p + \delta y$, $Y_N = \delta y$.

Donc :

$$\boxed{P_N[y_N = y] \simeq \frac{1}{\sqrt{2\pi}} \exp - \frac{y^2}{2}} \qquad (1.12)$$

avec $Var(Y_N) = \sigma_N^2 = Npq$ qui est la variance de la loi binomiale

On vient de démontrer le théorème central limite dans un cas particulier le théorème central limite. On l'énoncera de façon précise dans la section suivante. Son interprétation est donc que les fluctuations autour de la moyenne sont caractérisées par une distribution Gaussienne (1.10). Par ailleurs, en renormalisant la variable ΣX_i correctement (notez le facteur d'échelle en $N^{-1/2}$), la nouvelle variable est une Gaussienne centrée réduite

Rq : Il est tout à fait fondamental ici que la dérivée seconde de I_p soit non nulle. En effet, si $I_p''(p) = 0$ il faut aller à l'ordre supérieur et le comportement sera différent (on aura notamment un facteur d'échelle $\neq 1/2$). On a remarqué que $\boxed{I_p''(p) = \frac{1}{Var(X_i)}}$. C'est un résultat général comme on va le voir dans un instant. Il signifie ici que le théorème central limite nécessite que $Var(X_i) < \infty$.

Interprétation de la fonction I_p

On peut réécrire I_p comme :

$$I_p(z) = (1-z)\log(1-z) + z\log z - (1-z)\log(1-p) - z\log p$$

Or $z \in [0,1]$ est une fréquence et donc moralement une probabilité. $I_p(z)$ a alors la forme d'une entropie. Pour le voir, prenons d'abord $p = 1/2$. On a alors

$$\boxed{I_{\frac{1}{2}}(z) = -S(z) + \log 2} \qquad (1.13)$$

où $S(z) = -z\log z - (1-z)\log(1-z)$ est l'entropie de la distribution de proba $P[X_i = 1] = z$. Dans le cas général $I_p(z)$ a la forme d'une entropie relative ou entropie de Kullback. On rappelle que l'entropie relative entre 2 distributions de probabilité P, Q est (cf) :

$$\boxed{I[P,Q] = -\sum_{i=1}^{n} P_i \log\left(\frac{P_i}{Q_i}\right)} \qquad (1.14)$$

Ici $I_p(z)$ est l'entropie relative entre la distribution binaire de proba z et la distribution réelle, de probabilité p. Elle mesure en quelque sorte la distance (même si elle n'a pas toutes les propriétés d'une distance) entre ces 2 probabilités et elle est nulle si $z = p$.

On voit donc ici le rôle tout particulier que joue l'entropie dans la propriété de grande déviation (1.7) qui conduit aux lois des grands nombres, ainsi que dans le théorème central limite, où sa dérivée seconde contrôle la probabilité des fluctuations au voisinage de la moyenne. Ce résultat est à relier à l'exercice où l'on a montré que les fluctuations d'énergie étaient liées à la capacité calorifique.

2) Convergences

a) Convergence presque sûre.

On considère maintenant le cas général où l'on souhaite étudier les propriétés asymptotiques d'une suite de variables aléatoires $\{X_i\}_{i=1}^{n}$. Remarquons tout d'abord que les variables aléatoires étant des fonctions on parle bien ici de convergences de fonctions et non de convergence de nombres. On sait qu'il y a, en analyse fonctionnelle, de nombreuses notions de convergence. On va les retrouver ici, parfois sous un autre nom.

La notion la plus immédiate de convergence en analyse fonctionnelle est la convergence simple : $X_n \longrightarrow X^*$ simplement, si $X_n(\omega) \to X^*(\omega)$ $\forall \omega \in \Omega$. Cette notion est de peu d'utilité en théorie des probabilités si on l'utilise telle qu'elle. En effet demander la convergence pour tous les $\omega \in \Omega$ est beaucoup ~~trop~~ fort et cette propriété n'est en général jamais satisfaite.

On peut par contre demander que $X_n(\omega) \to X^*(\omega)$ pour un ensemble de P probabilité 1. On a vu dans l'exemple précédent qu'on pouvait très bien avoir cette propriété même si l'ensemble sur lequel la suite converge est une toute petite partie de Ω. Ce type de convergence s'appelle la convergence presque-sûre.

Définition : Une suite de v.a. $\{X_n\}$ converge P-presque sûrement vers une v.a. X^* si :

$$P\left[\{\omega \mid \lim_{n\to+\infty} X_n(\omega) = X^*(\omega)\}\right] = 1 \qquad (2.1)$$

On notera la référence à la probabilité P. On a ainsi vu dans l'exemple précédent qu'une suite pouvait converger P_1 presque sûrement vers une v.a (ou un nombre) α_1 et P_2 presque sûrement vers une autre v.a (ou nombre) α_2. C'est simplement dû au fait que

P_1 et P_2 ne "voient" pas les mêmes ensembles. Dans ce cas on dit que les 2 probabilités sont étrangères.

Def : 2 distributions de probabilité P_1, P_2 sont absolument continues ssi tout ensemble de P_1 probabilité nulle est de P_2 probabilité nulle.

Def : 2 mesures sont étrangères si elles ne sont pas absolument continues

Prop : Si une suite $X_n \xrightarrow{P_1 p.s} X^*$ et si $P_1 << P_2$ alors $X_n \xrightarrow{P_2 p.s} X^*$

Un exemple important de convergence presque-sûre est la loi forte des grands nombres.

Soit $\{X_i\}$ une suite de v.a. i.i.d., de moyenne $E(X_i) = m < \infty$.

Alors la variable aléatoire :

$$Y_N = \frac{1}{N} \sum_{i=1}^{N} X_i \xrightarrow[P p.s]{} m \qquad (2.2)$$

NB : Ici on a donc convergence vers un nombre.

Un autre exemple important, très utile en physique statistique et qui généralise en quelque sorte la loi des grands nombres est le théorème ergodique. On le présentera en détails dans le chapitre suivant.

Convergence presque sûre et lemme de Borel-Cantelli

Pour démontrer la convergence presque-sûre on utilise bien souvent le lemme de Borel Cantelli :

Soit $\{A_n\}_{n=1}^{+\infty}$ une suite d'événements, telle que :

i) $\sum_{n=1}^{+\infty} P(A_n) < +\infty$ alors $P\left[\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right] = 0 \qquad (2.3)$

ii) $\sum_{n=1}^{+\infty} P(A_n) = +\infty$ alors $P\left[\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right] = 1 \qquad (2.4)$

NB : On appelle parfois l'ensemble $\bigcap_{n=1}^{+\infty} \bigcup_{\ell=n}^{+\infty} A_\ell$ la limite supérieure de la suite d'événements $\lim A_n$. Cette limite correspond à l'événement une infinité d'événements A_ℓ a lieu.

Supposons qu'on ait une propriété de grande déviation comme dans l'ex précédent i.e. $P[|Y_N - X^*| \geq \varepsilon] \sim \exp - N I(\varepsilon)$ où $I(\varepsilon)$ est positive et nulle en $\varepsilon = 0$. Alors en posant $A_N(\varepsilon) := \{\omega \mid |Y_N - X^*| \geq \varepsilon\}$ l'ensemble $\{\omega \mid \lim_{N \to +\infty} Y_N = X^*(\omega)\}$ est donné par la condition.

$\forall \varepsilon > 0$, $\exists N_0 > 0$, $\forall N > N_0$ $|Y_N(\omega) - X^*(\omega)| < \varepsilon$. En termes ensemblistes pour ε fixé, c'est l'ensemble $\bigcup_{N_0=1}^{+\infty} \bigcap_{N=N_0}^{+\infty} \overline{A_N}(\varepsilon)$, dont le complémentaire (l'ensemble des ω pour lesquels la limite n'est pas $X^*(\omega)$ ou n'existe pas) est $\bigcap_{N=1}^{+\infty} \bigcup_{N=N_0}^{+\infty} A_N(\varepsilon)$. Comme $P[A_N(\varepsilon)] \sim \exp - N I(\varepsilon)$ la série $\sum P[$ converge donc $\lim \left[\bigcap_{N=1}^{+\infty} \bigcup_{N=N_0}^{+\infty} A_N(\varepsilon) \right] = 0$, $\forall \varepsilon > 0$. On en déduit que l'ensemble des ω pour lesquels la limite est X^* est de probabilité 1.

b) Convergence en probabilité

La convergence presque sûre est la plus forte des convergences qu'on rencontre en théorie des probabilités (avec la convergence L^2, cf infra). Mais parfois, on n'a pas cette convergence, bien qu'on ait une convergence dans un sens plus faible.

On dit que $X_n \to X^*$ en probabilité si :

$$\boxed{\forall \varepsilon > 0 \quad \lim_{n \to +\infty} P\left[\,|X_n - X^*| > \varepsilon\,\right] = 0.} \qquad (2.4)$$

Le sens est le suivant. Pour chaque ~~n~~ on regarde l'ensemble des ω tels que $|X_n - X^*| > \varepsilon$. La probabilité de cet ensemble tend vers 0.

Ex 1 :

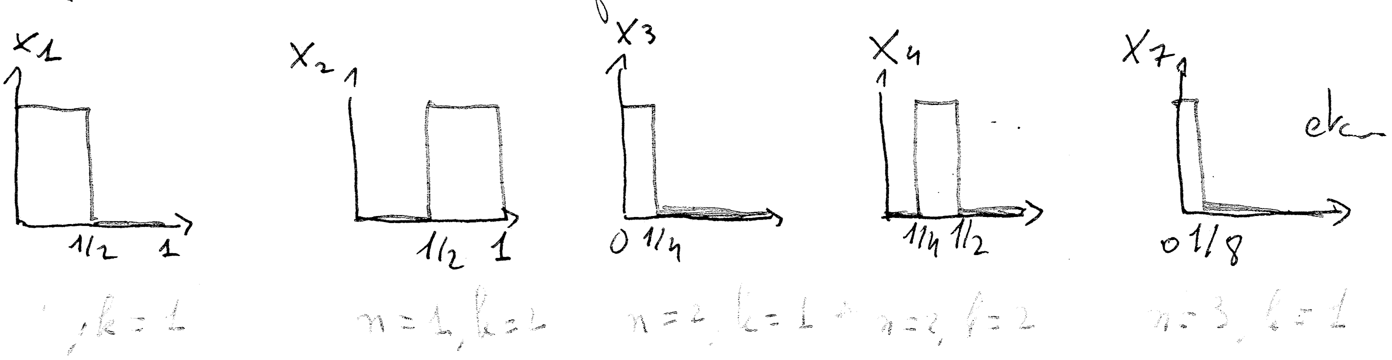
On a vu dans l'exemple de la section 1 que $P[\,|Y_N - m| > \varepsilon\,] \sim \exp(-N\, I_p(\varepsilon))$. On a donc bien convergence en probabilité. Mais on pourrait par exemple aussi avoir $P[\,|Y_N - X^*| > \varepsilon\,] = \frac{C}{\sqrt{N}}$. Dans ce cas on a convergence en loi, mais pas nécessairement convergence presque-sûre (car la série $C \sum_{N=1}^{+\infty} \frac{1}{\sqrt{N}}$ diverge).

De manière générale Convergence presque sûre $\Rightarrow$ Convergence en loi mais la réciproque est fausse

Ex 2

Soit $\{X_n\}$ une suite de v.a telle que $X_{2(2^n-1)+k} = \begin{cases} 1 & \text{si } \omega \in \left[\frac{k-1}{2^n} ; \frac{k}{2^n}\right] \\ 0 & \text{autrement} \end{cases}$

avec $k = 1 \ldots 2^n$. ω va en fonction



$n = 1, k = 1$ $\quad$ $n = 1, k = 2$ $\quad$ $n = 2, k = 1$ $\quad$ $n = 2, k = 2$ $\quad$ $n = 3, k = 1$

La convergence en loi implique en particulier que tous les moments (lorsqu'ils existent) convergent vers ceux de X^*

$$\boxed{E[X_n^k] \longrightarrow E[X^{*k}] \quad , \forall k \geq 0} \qquad (2.9)$$

Exemples

1) Soit X^* v.a uniforme sur $[-1, 1]$ et $X_n = (-1)^n X^*$.

On a $\quad F_{X_n}(x) = P[X_n \leq x] = P[(-1)^n X^* \leq x]$

$$= \begin{cases} P[X^* \leq x] & \text{si } n \text{ pair} \\ P[-X^* \leq x] & \text{si } n \text{ impair} \end{cases} = \begin{cases} F_{X^*}(x) \\ 1 - P[X^* < -x] = 1 - F_{X^*}(-x) = F_{X^*}(x) \end{cases}$$

loi symétrique/0

ici donc $F_{X_n}[x] = F_{X^*}[x]$ et la "convergence est triviale. Notez par contre qu'on n'a pas convergence p.s (la suite $(-1)^n X^*$ ne converge pas)

2)

Un exemple important de convergence faible est le théorème central limite ...

> Soit $\{X_i\}$ une suite de v.a.i.i.d. telles que:
>
> (i) $E[X_i] = m < +\infty$
>
> (ii) $\mathrm{Var}(X_i) = \sigma^2 < +\infty$

Soit $\quad Z_n = \dfrac{\sum_{i=1}^n X_i - nm}{\sigma\sqrt{n}} = \dfrac{n\left[\frac{1}{n}\sum X_i - m\right]}{\sigma\sqrt{n}} = \dfrac{\sqrt{n}}{\sigma}\left[\frac{1}{n}\sum X_i - m\right.$

alors $Z_n \xrightarrow[w]{} \mathcal{N}(0,1)$ i.e.

$$F_{Z_n}(z) \longrightarrow \int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx = F_Z(z)$$

Remarques

1) Ce théorème se généralise aux cas suivants

i) Les v.a. ne sont pas i.d. $E[X_i]=m_i$, $Var(X_i)=\sigma_i^2<\infty$

Il faut alors de plus qu'elles satisfassent à la condition de Lindenberg

$$\lim_{n\to+\infty} \frac{1}{B_n^2}\sum_{k=1}^{n}\int_{|x-m_k|>\varepsilon B_n}(x-m_k)^2\,\rho_{X_k}(x)\,dx = 0 \qquad \forall \varepsilon>0$$

avec $B_n^2 = \sum_{k=1}^{n}\sigma_k^2$. Dans ce cas, la variable renormalisée est $Z_n = B_n^{-1}\sum_{k=1}^{n}(X_k - m_k)$ et $Z_n \xrightarrow[w]{} \mathcal{N}(0,1)$

ii) Les v.a ne sont pas indépendantes mais leur corrélation [décroît] exponentiellement vite. Cette propriété apparait en sous le nom de mélange ou chaos moléculaire.

2) Si la variance n'est pas finie (lois de Levy par exemple) on montre qu'on a l'équivalent d'un TCL. Il y a convergence en loi ~~mais~~ pas vers une gaussienne mais vers une autre loi (loi stable) et l'exposant de renormalisation n'est pas 1/2 mais un autre, dépendant de la loi

Interprétation. On a déjà discuté de l'interprétation de ce théorème dans l'exemple de la section 1. Dans le cadre des hypothèses du TCL, la moyenne empirique $\frac{1}{n}\sum X_i \xrightarrow[p.s]{} m$ (loi des grands nombres). Le théorème C.L nous dit qu'en calculant la différence $\frac{1}{n}\sum X_i - m$ et en la renormalisant correctement (on la multiplie par $\frac{\sqrt{n}}{\sigma}$)

la variable obtenue, Z_n, tend vers une gaussienne centrée réduite. Pour n fini (mais grand) cela signifie que

$\frac{1}{n}\sum X_i - m = \frac{\sigma}{\sqrt{n}} Z_n$ se comporte comme une v.a. gaussienne centrée d'écart type $\frac{\sigma}{\sqrt{n}}$. Le TCL nous apprend donc que les fluctuations autour de la moyenne, sont, pour n assez grand, approximativement gaussiennes, avec une variance qui décroît en $1/\sqrt{n}$.

Pour de nombreuses raisons, il ne faut pas confondre la loi des grands nombres et le TCL. Tout d'abord, le type de convergence est différent : convergence p.s. pour la loi forte des grands nombres et convergence faible pour le TCL. On prenait ... propre aux mathématiciens et la dédaigner superbement. Et, du haut de notre superbe, interpréter le TCL en disant par exemple :

"la v.a. Z_n converge vers une v.a. Z qui est gaussienne $\mathcal{N}(0,1)$"

C'est absolument faux. Prenons par exemple $m = 0$, $\sigma = 1$, alors $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$. Cette suite n'est pas (avec proba 1) une suite de Cauchy et elle ne converge donc pas. On peut s'amuser à tracer sur un ordinateur le comportement de Z_n avec n. On voit que la variable ne converge pas, même si la distance entre Z_n et Z_{n+1} tend vers zéro quand $n \to +\infty$. On voit donc ici toute la subtilité qui se cache sous les différentes notions de convergence

Ex : Vérifier les assertions ci-dessus i.e. Z_n n'est pas une suite de Cauchy et simuler Z_n sur un ordinateur.

Démonstration du TCL.

On passe par les fonctions caractéristiques.

$$E[e^{itZ_n}] = E[e^{it(\Sigma X_i - nm)/\sigma\sqrt{n}}] = e^{-it\frac{nm}{\sigma\sqrt{n}}} E[e^{\frac{it}{\sigma\sqrt{n}}\Sigma X_i}]$$

Les X_i sont indépendantes donc

$$\varphi_{Z_n}(t) = e^{-itnm/\sigma\sqrt{n}} \varphi_X(t/\sigma\sqrt{n})^n$$

Pour t fixé, lorsque $n \to +\infty$ $\frac{t}{\sqrt{n}} \to 0$. On fait alors un DL de φ en 0.

$$\varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right) = \varphi_X(0) + \frac{t}{\sigma\sqrt{n}}\varphi_X'(0) + \frac{t^2}{n\sigma^2}\varphi_X''(0) + O\left(\left(\frac{t}{\sqrt{n}\sigma}\right)^3\right)$$

$$= 1 + \frac{itm}{\sigma\sqrt{n}} - \frac{t^2}{2\sigma^2 n}\mathrm{Var}(X) + \cdots = 1 + i\frac{tm}{\sigma\sqrt{n}} - \frac{t^2}{2n} + \cdots$$

Donc $\varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right)^n = \exp\left[n \log \varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right] \simeq \exp\left[n\log\left(1 + \frac{itm}{\sigma\sqrt{n}} - \frac{t^2}{2n}\right)\right]$

$$\sim \exp\left[n\left(\frac{itm}{\sigma\sqrt{n}} - \frac{t^2}{2n}\right)\right] = \exp i\frac{tnm}{\sigma\sqrt{n}} \exp -t^2/2$$

Donc $\varphi_{Z_n}(t) \sim \exp -t^2/2$ CQFD

d) Convergence L^P

On dit que $X_n \xrightarrow{L^P} X$ soi $E(|X_n - X|^P) \to 0$

avec $X_n, X \in L^P$ i.e les fonctions telles que $E(|X_n|^P)$ existe $< \infty$.

~~Cette~~ convergence L^P inclut :

→ La convergence en moyenne $E(|X_n - X|) \to 0$

→ La convergence en moyenne quadratique $E(|X_n - X|^2) \to 0$

Cela signifie dans le second cas que l'écart moyen entre X_n et X tend vers zero. C'est un type très fort de convergence.

e) Relations entre les différents type de convergence

~~Cas~~

$$\begin{array}{ccc} L^P & & P.s \\ \Downarrow & & \Downarrow \\ \text{moyenne} & \Longrightarrow & \text{Proba} \\ & & \Downarrow \\ & & \text{faible} \end{array}$$

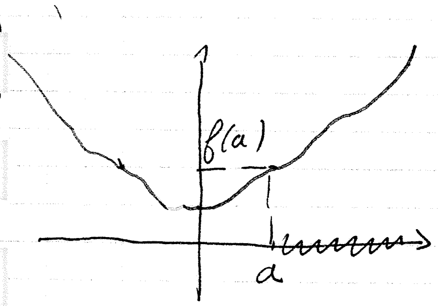
3) Inégalités fondamentales

Pour démontrer la convergence de variables aléatoires ~~des~~ les inégalités suivantes sont utiles.

a) Inégalité de Tchebychev

Soit f une fonction positive, paire, croissante sur $[0,+\infty[$ alors $\forall a \geq 0$.

$$\frac{E(f(X)) - f(a)}{\text{ess sup } f(X)} \leq P[\,|X| \geq a\,] \leq \frac{E[f(X)]}{f(a)}$$



Le sup essentiel est défini par $\inf\{ C \,;\, C \geq 0 \text{ et } P[\{ f(X) > C \} = 0] \}$ c'est donc le plus petit nombre C t.q. la proba que $f(x) > C$ est nulle.

* On suppose que P a une densité (pour simplifier).

$$P[\,|X| \geq a\,] = P[\,X \in]-\infty,-a] \cup [a,+\infty[\,] = P[X \in]-\infty,-a]] + P[X \in [a,+\infty[]$$

$$= \frac{1}{f(a)} \left(f(a)\, P[X \in]-\infty,-a]] + f(a)\, P[X \in [a,+\infty[] \right)$$

$$= \frac{1}{f(a)} \left[\int_{-\infty}^{-a} f(a) \rho_X(x)\,dx + \int_{a}^{+\infty} f(a) \rho_X(x)\,dx \right] \quad \text{or } f(x) \geq f(a) \text{ si } |x| \geq a$$

$$\Rightarrow P[\,|X| \geq a\,] \leq \frac{1}{f(a)} \left[\int_{-\infty}^{-a} f(x) \rho_X(x)\,dx + \int_{a}^{+\infty} f(x) \rho_X(x)\,dx \right] \quad \text{or } f(x) \geq 0$$

$$P[\,|X| \geq a\,] \leq \frac{1}{f(a)} \int_{-\infty}^{+\infty} f(x) \rho_X(x)\,dx = \frac{E[f(X)]}{f(a)}$$

Un corollaire important est l'inégalité de Bienaymé-Tchebichev. On pose $f(x) = x^2$ alors

$$P\left[|X - \bar{X}| \geq a\right] \leq \frac{Var(X)}{a^2}$$

Exemple d'application : loi faible des grands nombres.

$$P\left[\left|\frac{1}{N}\sum X_i - m\right| \geq \varepsilon\right] \leq \frac{Var\left(\frac{1}{N}\sum X_i\right)}{\varepsilon^2} = \frac{N V_X}{N^2\varepsilon^2} = \frac{V_X}{N\varepsilon^2} \to 0$$

b) Inégalité de Jensen.

Si $f(x)$ est une fonction convexe continue, convexe sur $]a,b[$ avec $X \in [a,b]$ (a, b peuvent être ∞) alors :

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

c) Inégalité de Holder

$$E[|XY|] \leq E(|X|^p)^{\frac{1}{p}} E(|Y|^q)^{1/q}$$