M1 de Maths-Info
Théorie de l'information

Alain Chenciner
Université Paris 7 & IMCCE
chenciner@imcce.fr

November 29, 2007

Abstract

A l'origine, il y a l'article "A Mathematical Theory of Communication" [Sh] publié par Claude Shannon en 1948 et l'exposé [Kh1] qu'en a donné Alexandre Khinchine à partir des résultats de B. McMillan et A. Feinstein. Aujourd'hui, la meilleure référence est de loin le beau livre [CT] "Elements of Information theory" de T.C. Cover et J.A. Thomas. Le cours débute par une introduction élémentaire à la théorie des probabilités incluant cependant cette forme forte de la loi des grands nombres qu'est le théorème ergodique de Birkhoff ; les systèmes dynamiques ergodiques sont en effet le cadre général de validité des théorèmes de Shannon (voir [B2]). D'une foule de références se détachent "Probability and measure" de Patrick Billingsley [B1] et "Probability theory, an introductory course" [Si] de Yakov Sinai. Centré sur l'exposé dans leur cadre le plus simple des deux grands théorèmes de Shannon concernant l'entropie : l'équirépartition asymptotique et la possibilité de transmission par un canal "bruyant", le cours se clôt par une brêve incursion dans la théorie des séries de Fourier avec l'exposé d'un troisième théorème de Shannon montrant l'existence d'un codage fini pour les signaux dont la largeur de bande est finie.

1 Les sources discrètes comme processus stochastiques

Les deux citations qui suivent indiquent clairement le cadre mathématique dans lequel nous nous placerons. La première est de Shannon [Sh] en 1948 :

We can think of a discrete source as generating the message, symbol by symbol. It will choose successive symbols according to certain probabilities depending in general on preceding choices as well as the particular symbol in question. A physical system, or a mathematical model of a system which produces such a sequence of symbols governed by a set of probabilities, is known as a stochastic process. We may consider a discrete source, therefore, to be represented by a stochastic process. Conversely, any stochastic process which produces a discrete

sequence of symbols chosen from a finite set may be considered a discrete source. This will include such cases as :

1. Natural written languages such as English, German, Chinese.

2. Continuous information sources that have been rendered discrete by some quantizing process.....

3. Mathematical cases where we merely define abstractly a stochastic process which generates a sequence of symbols.....

La deuxième est de Wiener [W], également en 1948 :

The message is a discrete or continuous sequence of measurable events distributed in time – precisely what is called a time-series by the statisticians.
.....

In doing this, we have made of communication engineering design a statistical science, a branch of statistical mechanics.....
.....

In the case of communication engineering, however, the significance of the statistical element is immediately apparent. The transmission of information is impossible save as a transmission of alternatives. If only one contingency is to be transmitted, then it may be sent most efficiently and with the least trouble by sending no message at all. The telegraph and the telephone can perform their function only if the messages they transmit are continually varied in a manner not completely determined by their past, and can only be designed effectively if the variations of these messages conforms to some sort of statistical regularity.

To cover this aspect of communication engineering, we had to develop a statistical theory of the amount of information, in which the unit amount of information was that transmitted as a single decision between equally probable alternatives. This idea occured at about the same time to several writers, among them the statistician R. A. Fisher, Dr. Shannon of the Bell Telephone Laboratories, and the author. Fisher's motive in studying this subject is to be found in classical statistical theory ; that of Shannon in the problem of coding information ; and that of the author in the problem of noise and message in electrical filters. Let it be remarked parenthetically that some of my speculations in this direction attach themselves to the earlier work of Kolmogoroff in Russia, although a considerable part of my work was done before my attention was called to the work of the Russian school.

Nous commencerons par étudier le cas d'une source "sans mémoire" dans laquelle chaque symbole émis est indépendant des précédents. Dans un premier temps, nous supposerons l'alphabet des symboles réduit à deux éléments, ce qui assimile la source à un jeu de pile ou face, éventuellement biaisé et donne l'occasion d'introduire les notions élémentaires de la théorie des probabilités ainsi que l'interprétation dynamique d'un tel jeu en termes des *décalages de Bernoulli*. Nous passerons ensuite aux *sources markoviennes* dans lesquelles la mémoire est limitée au symbole qui précède immédiatement le symbole émis. Nous introduirons enfin la notion de *source ergodique* qui est le cadre général auquel s'appliquent les théorèmes de Shannon.

Contents

1.1 Du jeu de pile ou face aux "décalages de Bernoulli"

1.1.1 De l'ensemble des événements élémentaires à la tribu des événements

On note $A = \{P, F\}$ ou $A = \{0, 1\}$ l'*alphabet*, ici un ensemble ayant deux éléments "pile" et "face".

On associe à A un ensemble Ω d'*événements élémentaires* :

jeu à 1 tirage : $\Omega = A$, ensemble des mots à 1 lettre ;
jeu à 2 tirages : $\Omega = A^2 = \big\{f : \{1, 2\} \to A\big\}$, ensemble des mots à 2 lettres ;
...
jeu à une infinité de tirages : $\Omega = A^{\mathbb{N}^*} = \big\{f : \mathbb{N}^* \to A\big\}$, ensemble des suites infinies de lettres ; on peut aussi considérer $\Omega = A^{\mathbb{Z}} = \big\{f : \mathbb{Z} \to A\big\}$, ensemble des suites doublement infinies de lettres (tirages dans le passé aussi bien que dans le futur).

Le cas fini : Commençons par les cas où $\Omega = A^n$ est un ensemble fini : on appellera *événement* un sous-ensemble C quelconque de Ω, c'est-à-dire une réunion d'événements élémentaires : il faut interpréter la réalisation de C comme *"il se passe ceci ou cela"*, les événements élémentaires concernés appartenant à C.

Supposons que la *probabilité* de tirer "pile" (que l'on représentera par 0) soit p et que celle de tirer "face" (que l'on représentera par 1) soit q : si $p \neq q$, cela signifie que la pièce est biaisée.

On suppose que les tirages sont *indépendants* les uns des autres. Nous reviendrons sur cette notion qui fonde la *théorie des probabilités* comme théorie distincte de la *théorie de la mesure*. Cela signifie que la probabilité de l'événement

élémentaire $\omega = a_1 a_2 \cdots a_n$ est $P(\omega) = p^{n_0} q^{n_1} = p^{n_0} q^{n-n_0}$, où n_0 est le nombre de 0 (=pile) parmi les a_i et n_1 le nombre de 1 (=face). Il y a exactement $C_n^{n_0} = C_n^{n_1}$ événements élémentaires ayant cette probabilité ; on a donc bien

$$\sum_{\omega \in \Omega} P(\omega) = \sum_{n_0=0}^{n} C_n^{n_0} p^{n_0} q^{n-n_0} = (p+q)^n = 1.$$

De plus, tout événement C a une probabilité qui est la somme des probabilités des événements élémentaires qui le composent.

L'ensemble $\mathcal{P}$ des événements, qui n'est autre que l'ensemble des parties de Ω, forme une *algèbre* au sens suivant :

Définition 1 *Soit Ω un ensemble quelconque. Une collection $\mathcal{G}$ de parties de Ω est appelée une algèbre si elle contient Ω lui-même et si elle est fermée pour les opérations de passage au complémentaire et de réunion finie.*

Il suit de la définition qu'une algèbre est également fermée pour les intersections finies. Dans notre cas $\mathcal{G}$ a un nombre fini d'éléments, puisqu'il en est ainsi de Ω, et le théorème ci-dessous affirme que sa donnée équivaut à celle d'une *partition* de Ω :

Théorème 1 *Si l'algèbre $\mathcal{G}$ est finie, il existe des sous-ensembles $A_1, \ldots, A_r$ deux à deux disjoints de Ω, les "atomes", dont la réunion est Ω, tels que $\mathcal{G}$ coïncide avec l'ensemble des union d'atomes. On dit que les A_i "engendrent" $\mathcal{G}$. Réciproquement, toute partition $\mathcal{P}$ "engendre" l'algèbre finie des unions de ses atomes.*

Indication de démonstration (voir [Si] page 4) : considérer les intersections $G_{\pm 1} \cap G_{\pm 2} \cap \ldots \cap G_{\pm r}$, où les G_i, $i = 1, \ldots, r$ sont une énumération des éléments de $\mathcal{G}$ et $G_{-i} = G_i^c$ est le complémentaire de G_i.
Dans le cas de l'algèbre $\mathcal{P}$ de toutes les parties de Ω fini, les A_i sont les singletons $\{\omega\}$ de Ω, c'est-à-dire les événements élémentaires.

Pourquoi introduire cette notion d'algèbre ? Fondamentale dans le cas où le cardinal de Ω est infini, cette notion se rencontre déjà naturellement dans le cas fini : supposons par exemple n'être capables que de compter le nombre $f(\omega) = \sum_{i=1}^{n} a_i$ de "face" dans $\omega = a_1 a_2 \ldots a_n \in A^n$ (on a représenté les "pile" par 0 et les "face" par 1). L'application $f : \Omega \to \{0, 1, 2, \ldots, n\}$ est l'exemple d'une *variable aléatoire* à valeurs finies (*simple random variable* chez Billingsley [B1] ; pour la définition générale d'une variable aléatoire, voir 1.1.3). Ses niveaux $f^{-1}(k)$ forment une partition de Ω ; ils engendrent donc une algèbre (théorème 1), différente de $\mathcal{P}$ et seule la probabilité des éléments de cette algèbre fait sens pour nous (des événements "plus fins" sont indiscernables par nos moyens d'observation). Si, au lieu de f, on considère la variable aléatoire à deux valeurs qui associe à $\omega \in \Omega$ la parité $g(\omega)$ du nombre de "face", c'est l'algèbre encore plus grossière engendrée par la partition en deux parties "mots pairs" et "mots impairs" qui seule fait sens.

Exercice 1 *Calculer les probabilités des A_i dans les deux cas évoqués ci-dessus.*

Le cas infini : Dans le cas où Ω est un ensemble infini, la définition d'événement est plus subtile. La formalisation qu'a donnée Kolmogorov du calcul des probabilités dans les années 30 (voir [Ko]) est ancrée dans la théorie de la mesure et fait appel à la notion de *σ-algèbre* ou *tribu* due à Emile Borel, celle-la même qui fonde la notion de mesure au sens de Lebesgue. Les événements seront uniquement les sous-ensembles de Ω que l'on peut "mesurer", c'est-à-dire ceux auxquels on peut attacher une probabilité.

Définition 2 *Soit Ω un ensemble quelconque. Une algèbre $\mathcal{F}$ de parties de Ω est appelée une σ-algèbre (ou une "tribu") si elle est fermée pour les opérations de réunion dénombrable. La paire $(\Omega, \mathcal{F})$ est appelée un "espace mesurable".*

Une σ-algèbre est donc fermée pour le passage au complémentaire, les unions et les intersections dénombrables.

Lemme 2 *Une famille $\mathcal{G}$ de parties de Ω (par exemple une algèbre) étant donnée, il existe une plus petite σ-algèbre $\mathcal{F}$ qui la contient, définie comme l'intersection de toutes les σ-algèbres qui la contiennent. $\mathcal{F} = \sigma(\mathcal{G})$ est la σ-algèbre engendrée par $\mathcal{G}$.*

Définition 3 *Une (mesure de) probabilité sur l'espace mesurable $(\Omega, \mathcal{F})$ est une fonction $P : \mathcal{F} \to [0, 1]$ telle que $P(\Omega) = 1$ et $P(\cup C_i) = \sum P(C_i)$ quelle que soit la famille au plus dénombrable d'éléments de $\mathcal{F}$ deux à deux disjoints (en tant que parties de Ω). On dit alors que P possède la propriété d'additivité dénombrable. Si $C \in \mathcal{F}, P(C)$ est la "probabilité" de l'événement C et le triplet $(\Omega, \mathcal{F}, P)$ est appelé "espace de probabilité".*

La restriction à des familles dénombrables est justifiée par le fait que les éléments non nuls d'une famille sommable de nombres réels forment un ensemble au plus dénombrable. Kolmogorov avait choisi comme axiome une propriété équivalente à l'additivité dénombrable, l'*axiome de continuité*, qui est une forme de convergence monotone : si une suite décroissante $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ d'éléments de $\mathcal{F}$ est d'intersection vide, on a $\lim_{n\to\infty} P(A_n) = 0$. Par ailleurs, il note $\sum C_i$ une union disjointe et il m'arrivera de le suivre.

Exercice 2 *Montrer que l'axiome d'additivité dénombrable est équivalent à l'axiome de continuité ainsi qu'à chacune des propriétés suivantes :*
1) Pour toute suite croissante A_i d'éléments de $\mathcal{F}$, $P(\cup A_i) = \lim P(A_i)$.
2) Pour toute suite décroissante A_i d'éléments de $\mathcal{F}$, $P(\cap A_i) = \lim P(A_i)$.

L'exemple paradigmatique est l'intervalle $[0, 1]$ muni de la tribu des *boréliens* et de la mesure de Lebesgue. Nous y reviendrons et nous verrons qu'un le jeu de pile ou face non biaisé avec une infinité de tirages lui est intimement lié.

1.1.2 De l'algèbre à la σ-algèbre : prolongement des mesures

Théorème 3 *Une mesure de probabilité P sur une algèbre $\mathcal{G}$ de parties d'un ensemble Ω se prolonge de façon unique à la σ-algèbre $\mathcal{F}$ engendrée par $\mathcal{G}$.*

Un survol supersonique de la démonstration (voir [B1] section 3) : on commence par définir une *mesure extérieure* P^* qui, à chaque partie A de Ω associe

$$P^*(A) = \inf \sum_n P(A_n),$$

où le inf est pris sur l'ensemble des suite $A_1, A_2, \ldots$ d'éléments de $\mathcal{G}$ telles que $A \subset \cup_n A_n$. On obtient une *mesure intérieure* P_* en passant au complémentaire :

$$P_*(A) = 1 - P^*(A^c),$$

où $A^c = \Omega \setminus A$ est le complémentaire de A. On montre alors que si $A \in \mathcal{F}$, on a pour toute partie E de Ω l'égalité

$$P^*(A \cap E) + P^*(A^c \cap E) = P^*(E).$$

On en déduit ensuite que P^* est une mesure de probabilité sur $\mathcal{F}$ et on vérifie facilement que $P^*(A) = P(A)$ si $A \in \mathcal{G}$. L'unicité découle (avec un peu de travail) de la minimalité de $\mathcal{F}$.

ATTENTION (voir [B1] section 2) : une induction transfinie est nécessaire pour passer d'une algèbre à la σ-algèbre engendrée par les opérations de complémentaire et d'union dénombrable. En particulier, ou bien une tribu est finie, ou bien elle a au moins la puissance du continu. Ceci suggère la difficulté qu'il y peut y avoir à expliciter certains boréliens ; la théorie de l'information nous en fournira cependant des exemples très naturels dans $\{0,1\}^{\mathbb{N}^*}$ ou $\{0,1\}^{\mathbb{Z}}$.

La mesure de Lebesgue sur l'intervalle $[0,1]$ **:** Il est commode de définir la *σ-algèbre $\mathcal{B}$ des boréliens* comme celle engendrée par l'algèbre $\mathcal{I}$ des unions finies d'intervalles $[x_i, y_i[\subset [0,1]$ disjoints. La mesure de Lebesgue λ est l'unique extension de la mesure des intervalles : $\lambda([x,y[) = |y-x|$. La mesure d'un borélien B se définit donc de la manière suivante : c'est l'inf de la somme $\sum_n |y_n - x_n|$ des mesures (longueurs) d'une suite d'intervalles $[x_n, y_n[$ dont la réunion contient B. On peut étendre P à une tribu plus grande que la tribu borélienne mais, si l'on accepte l'axiome du choix et l'hypothèse du continu, on peut montrer qu'il n'existe pas de mesure de probabilité sur la tribu de toutes les parties de $[0,1]$ qui ait la propriété que chaque singleton $\{x\}$ vérifie $P(\{x\}) = 0$. Une extension cependant s'impose, celle de P à tous les ensembles "négligeables", c'est-à-dire aux parties de $[0,1]$ qui peuvent être recouvertes par l'union d'une suite d'intervalles dont la somme des longueurs est arbitrairement petite (on leur attribue donc une mesure nulle, qu'ils appartiennent ou non à la tribu). Les parties dénombrables sont négligeables (exercice) mais il y en a bien d'autres.

Donnons maintenant l'exemple d'un borélien "non trivial" de $[0,1]$, en fait tout simplement un fermé de $[0,1]$. C'est un exemple à propos duquel Emile Borel se plaisait à dire son étonnement : soit $r_1, r_2, \ldots, r_n, \ldots$ une énumération des rationnels de $[0,1]$. Soit I_n un intervalle ouvert de longueur $\frac{\epsilon}{2^n}$ centré sur r_n et $A = \cup_n (I_n \cap [0,1])$. L'ensemble A est un ouvert dense de $[0,1]$ et sa mesure est

majorée par $\sum_{n=1}^{\infty} \frac{\epsilon}{2^n} = \epsilon$. Le complémentaire B de A est un borélien totalement discontinu, précisément un *ensemble de Cantor "épais"* dont la mesure est supérieure à $1 - \epsilon$. Suivant Billingsley [B1], nous donnerons dans ce qui suit un exemple plus complexe, très lié au jeu de pile ou face, l'ensemble des *nombres normaux* de Borel.

1.1.3 Le jeu de pile ou face comme processus stochastique stationnaire

Considérons l'ensemble $\Omega = A^{\mathbb{N}^*} = \{0,1\}^{\mathbb{N}^*}$ des suites infinies

$$\omega = a_1 a_2 \dots$$

formées de 0 et de 1, assimilées chacune à une suite infinie de tirages indépendants d'un jeu de pile ou face. Une telle suite est la réalisation d'un *processus stochastique stationnaire* sans mémoire (la définition précise sera donnée plus loin). La stationnarité du processus signifie que la probabilité p que $a_i = 0$ et la probabilité $q = 1 - p$ que $a_i = 1$ sont indépendantes du "temps" i du tirage ; l'indépendance (ou absence de mémoire) des tirages signifie que la probabilité d'un *cylindre*

$$A_{i_1 i_2 \dots i_k}^{j_1 j_2 \dots j_k} = \big\{\omega \in \Omega; a_{i_1} = j_1, a_{i_2} = j_2, \dots, a_{i_k} = j_k\big\}, i_1, \dots \in \mathbb{N}^*,\ j_1, \dots \in \{0,1\},$$

est

$$P(A_{i_1 i_2 \dots i_k}^{j_1 j_2 \dots j_k}) = P(A_{i_1}^{j_1}) P(A_{i_2}^{j_2}) \dots P(A_{i_k}^{j_k}),$$

c'est-à-dire $p^{k_0} q^{k_1}$ si la suite $j_1 j_2 \dots j_k$ comporte k_0 termes égaux à 0 et $k_1 = k - k_0$ termes égaux à 1.

Exercice 3 *1) Une intersersection finie de cylindres est encore un cylindre ;*
2) le complémentaire d'un cylindre est une union disjointe d'un nombre fini de cylindres ;
3) une union finie de cylindres peut aussi s'écrire comme une union finie de cylindres disjoints ;
4) déduire de 1),2),3) que les unions finies de cylindres disjoints forment une algèbre $\mathcal{G}$ de parties de Ω (à comparer à l'algèbre des unions finies d'intervalles $[a_i, b_i[$ disjoints de $[01]$).

Il est naturel de définir la tribu $\mathcal{F}$ comme celle engendrée par l'algèbre $\mathcal{G}$ des unions finies de cylindres. On dit que la mesure de probabilité $P = P_{p,q}$ dont on a donné ci-dessus la valeur sur les cylindres est le *produit* d'une infinité de copies de la mesure (p, q) sur $\{0, 1\}$.

Exhiber des éléments non triviaux de $\mathcal{F}$ n'est pas si facile, si l'on excepte les réunions dénombrables de cylindres disjoints. En fait, nous allons montrer que le problème est le même que celui d'exhiber un *borélien* non trivial de l'intervalle $[0, 1] \subset \mathbb{R}$. La tribu $\mathcal{F}$ est d'ailleurs la *tribu borélienne* pour la topologie sur Ω engendrée par les cylindres (qui n'est autre que la topologie *produit infini*, voir l'exercice suivant).

Il reste à définir la probabilité (la mesure) d'un élément quelconque de $\mathcal{F}$ comme l'unique extension de la probabilité définie ci-dessus pour les cylindres. Ici encore, ce problème est le même que celui de la théorie de l'intégration où l'on doit généraliser à des boréliens quelconques la mesure des segments.

Exercice 4 (L'espace topologique $\{0,1\}^{\mathbb{N}^*}$ comme ensemble de Cantor) *On munit $\{0,1\}^{\mathbb{N}^*}$ de la topologie produit : c'est celle dont une base d'ouverts est formée par les cylindres. Autrement dit, un ouvert est une union quelconque de cylindres. Une autre façon de la définir est d'introduire la distance $d(a_1a_2\dots, b_1b_2\dots) = \sum_{k=1}^{\infty} \frac{|a_k - b_k|}{2^k}$. Montrer que l'application*

$$f_3 : \{0,1\}^{\mathbb{N}^*} \to [0,1], \quad f_3(a_1a_2\dots a_n\dots) = \sum_{k=1}^{\infty} \frac{2a_k}{3^k}$$

est un homéomorphisme de $\{0,1\}^{\mathbb{N}^*}$ sur l'ensemble triadique de Cantor K. Montrer que K est de mesure de Lebesgue nulle.

De $\{0,1\}^{\mathbb{N}^*}$ à l'intervalle $[0,1]$: Considérons maintenant l'application

$$f_2 : A^{\mathbb{N}^*} \to [0,1], \quad f_2(a_1a_2\dots a_n\dots) = \sum_{k=1}^{\infty} \frac{a_k}{2^k}.$$

Tout élément de $[0,1]$ ayant un *développement dyadique*, cette application est une surjection. Elle n'est pas injective : l'image réciproque de $\frac{1}{2}$ est formée de $1000\dots$ et $0111\dots$ et le même phénomène de non unicité du développement dyadique se produit sur l'ensemble dénombrable dense formé des *nombre dyadiques*, de la forme $\frac{m}{2^k}$ où m et k sont des entiers. Mais, dans le cas équiprobable ($p = q = 1/2$), f_2 est aussi bonne qu'une bijection *du point de vue de la mesure.* Préciser cette affirmation demande qu'on introduise quelques définitions :

Remarque : cylindres dans le cas d'un alphabet quelconque. Lorsque A n'est pas réduit à deux éléments, il faut entendre par *cylindre* une partie de $\Omega = A^{\mathbb{N}^*}$ (ou $A^{\mathbb{Z}}$) définie en astreignant un nombre fini $i_1, i_2, \dots, i_k$ de termes de la suite à appartenir chacun à une certaine partie *non vide* mesurable de A (i.e. une partie non vide quelconque si A est fini). Autrement dit, les indices $j_1, j_2, \dots, j_k$ doivent être entendus comme représentant chacun une partie non vide de A. Cette définition coïncide bien entendu avec celle donnée auparavant dans le cas où $A = \{0,1\}$. Plus conceptuellement (voir [B1]), l'entier k étant donné, on associe à toute partie H de A^k le *cylindre*

$$A_H = \{\omega = a_1 \cdots a_n \cdots \in A^{\mathbb{N}^*},\ (a_1 \cdots a_k) \in H\}$$

(s'il s'agit de $\mathrm{A}^{\mathbb{Z}}$, on part d'un sous-ensemble H de $A^{k_1} \times A^{k_2}$ et l'on demande que $\big((a_{1-k_1}, \cdots a_0), (a_1, \cdots, a_{k_2})\big)$ appartienne à H). Si l'on remplace les sous-ensembles H par des ouverts, on obtient une base d'ouverts de la *topologie produit infini.*

Exercice 5 *Montrer qu'avec cette définition un peu plus générale, un cylindre est ce que nous avions précédemment appelé une union finie de cylindres et que les cylindres forment donc une algèbre. Remarquer également que cette définition rend plus agréables les démonstrations, par exemple celle que $P_{p,q}$ et bien une loi de probabilité sur l'algèbre des cylindres.*

1.1.4 Mesurabilité et préservation de la mesure

Les applications *mesurables* jouent le rôle des flèches dans la *catégorie* dont les objets sont les espaces mesurables (ce ne sont pas des gros mots, penser en termes de catégories est souvent utile) :

Définition 4 *Une application $f : (X, \mathcal{A}) \to (Y, \mathcal{B})$ d'un espace mesurable dans un autre est dite* mesurable *si l'image réciproque $f^{-1}(B)$ d'un élément quelconque de $\mathcal{B}$ est un élément de $\mathcal{A}$.*

Remarquons que *la mesurabilité est une propriété d'autant plus astreignante que la tribu $\mathcal{A}$ de l'espace de gauche est petite.* Si par exemple elle est définie par une partition finie et si l'espace de droite est $\mathbb{R}$ avec sa tribu borélienne, les seules applications mesurables sont celles qui sont constantes sur chacun des morceaux de la partition.

Définition 5 *On dit qu'une application mesurable $f : (X, \mathcal{A}, \mu) \to (Y, \mathcal{B}, \nu)$ d'un espace de probabilité dans un autre préserve la mesure si quel que soit $B \in \mathcal{B}$, on a $\mu(f^{-1}(B)) = \nu(B)$.*

Dorénavant, lorsqu'on parlera d'une application préservant la mesure, on sous-entendra qu'elle est mesurable. Dans la pratique, on utilise toujours le lemme suivant :

Lemme 4 *Si la tribu $\mathcal{B}$ est engendrée par l'algèbre $\mathcal{G}$, il suffit pour que l'application $f : (X, \mathcal{A}) \to (Y, \mathcal{B})$ soit mesurable que l'image réciproque $f^{-1}(G)$ d'un élément quelconque G de $\mathcal{G}$ appartienne à $\mathcal{A}$. De même, il suffit pour que f préserve la mesure que $\mu(f^{-1}(B)) = \nu(B)$ soit vérifié pour tout élément G de l'algèbre $\mathcal{G}$.*

Comme indication de preuve, je laisse le lecteur réfléchir sur la cryptique identité

$$f^{-1}(\sigma(\mathcal{G})) = \sigma(f^{-1}(\mathcal{G})).$$

Définition 6 *Une application $f : (X, \mathcal{A}, \mu) \to (Y, \mathcal{B}, \nu)$ d'un espace de probabilité dans un autre est un isomorphisme d'espaces de probabilité si*
1) c'est une bijection modulo des ensembles de mesure nulle (i.e. il existe $A \subset X$ μ-négligeable et $B \subset Y$ ν-négligeable tels que f définisse une bijection de $X \setminus A$ sur $Y \setminus B$) (on parle de bijection "mod 0")
2) f et f^{-1} sont mesurables et préservent la mesure

Nous pouvons maintenant préciser l'affirmation ci-dessus (formulée la première fois par Steinhaus dans [St]) :

Proposition 5 *L'application*

$$f_2\big(\{0,1\}^{N^*}, \mathcal{F}, P_{\frac{1}{2},\frac{1}{2}}\big) \to \big([0,1], \mathcal{B}, \lambda\big)$$

est un isomorphisme d'espaces de probabilité. L'espace des suite infinies de 0 et de 1 muni de sa tribu borélienne et de la mesure correspondant à des suites de tirages indépendants et non biaisés est donc, au point de vue de la théorie de la mesure, équivalent à l'intervalle $[0,1]$ *muni de sa tribu borélienne et de la mesure de Lebesgue.*

Esquisse de démonstration : du lemme 4, on déduit qu'il suffit de vérifier la mesurabilité et la préservation de la mesure par f_2 sur des générateurs de la tribu borélienne i.e. sur les intervalles et même sur les intervalles de la forme $]\frac{p}{2^k}, \frac{p+1}{2^k}]$. Or, si $x = \sum_{i=1}^{k} \frac{a_i}{2^i}$ et $y = x + \frac{1}{2^k}$, on vérifie immédiatement que $f_2^{-1}[x,y] = A_{12\ldots k}^{a_1 a_2 \ldots a_k}$ et donc que $P_{\frac{1}{2},\frac{1}{2}}\big(f_2^{-1}[x,y]\big) = \frac{1}{2^k} = |y-x| = \lambda([x,y])$.
D'autre part, la non-injectivité de f_2 a lieu sur un *ensemble négligeable* (i.e. de mesure nulle) : soit $\mathcal{D}$ le sous-ensemble de $\{0,1\}^{N^*}$ formé des suites qui à partir d'un certain rang n'ont plus que des 1 ; $\mathcal{D}$ est inclus dans la réunion de cylindres dont la somme des probabilités peut être choisie arbitrairement petite (exercice) et est donc de mesure nulle. Son complémentaire $\{0,1\}^{N^*} \setminus \mathcal{D}$ est en bijection avec l'intervalle $[0,1[$ obtenu en enlevant un seul point (également négligeable) à $[0,1]$.

Définition 7 *Etant données une application mesurable* $f : (X, \mathcal{A}) \to (Y, \mathcal{B})$ *et une mesure (de probabilité)* μ *sur* $(X, \mathcal{A})$*, on appelle image directe de* μ *et on note* $f_*\mu$ *la mesure (de probabilité) sur* $(Y, \mathcal{B})$ *définie par* $f_*\mu(B) = \mu\big(f^{-1}(B)\big)$.

La proposition précédente s'écrit donc $(f_2)_* P_{\frac{1}{2},\frac{1}{2}} = \lambda$ et $(f_2^{-1})_* \lambda = P_{\frac{1}{2},\frac{1}{2}}$.

Un borélien de $[0,1]$: les nombres normaux de Borel. Toute propriété de la mesure de Lebesgue sur $[0,1]$ se traduit donc en une propriété de la probabilité P sur $\Omega = \{0,1\}^{\mathbb{N}^*}$ obtenue lorsque 0 et 1 sont équiprobables et réciproquement. L'exemple des *nombres normaux* de Borel, choisi par Billingsley comme introduction à [B1] est particulièrement pertinent pour notre sujet : c'est l'ensemble $\mathcal{N}$ des $\omega \in [0,1]$ dont le développement dyadique $a_1 a_2 \ldots a_n \ldots$ vérifie : $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} a_i = \frac{1}{2}$: autrement dit, la fréquence asymptotique du nombre de 1 est égale à celle du nombre de 0. L'ensemble $\mathcal{N}$ est un borélien de $[0,1]$ car $\omega \in \mathcal{N}$ équivaut à

$$\forall k \in \mathbb{N}, \exists m \in \mathbb{N}, \forall n \geq m, \ \left| \frac{1}{n} \sum_{i=1}^{n} a_i - \frac{1}{2} \right| < \frac{1}{k},$$

ce qui s'écrit

$$\mathcal{N} = \cap_{k=1}^{\infty} \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} \left[\omega; \left| \frac{1}{n} \sum_{i=1}^{n} a_i - \frac{1}{2} \right| < \frac{1}{k} \right].$$

Une conséquence du théorème ergodique de Birkhoff que l'on énoncera plus loin est que les nombres normaux forment un ensemble de mesure 1.

Exercice 6 (suite de l'exercice 4) *Déduire de ce qui précède que* $f_2 \circ f_3^{-1}$ *est une application continue surjective* δ *de* K *sur* $[0,1]$. *Montrer que cette application prend la même valeurs aux extrémités d'un intervalle de* $[0,1] \setminus K$. *Tracer le graphe de l'unique application continue de* $[0,1]$ *dans lui-même obtenue en prolongeant* δ *par une constante dans chaque intervalle de* $[0,1] \setminus K$. *Vérifier que ce graphe mérite l'appellation d'"escalier du diable" que lui donnent les dynamiciens : c'est un bel exemple de fonction à variation bornée mais non absolument continue (i.e. non égale à l'intégrale de sa dérivée, qui existe et est nulle presque partout pour la mesure de Lebesgue).*

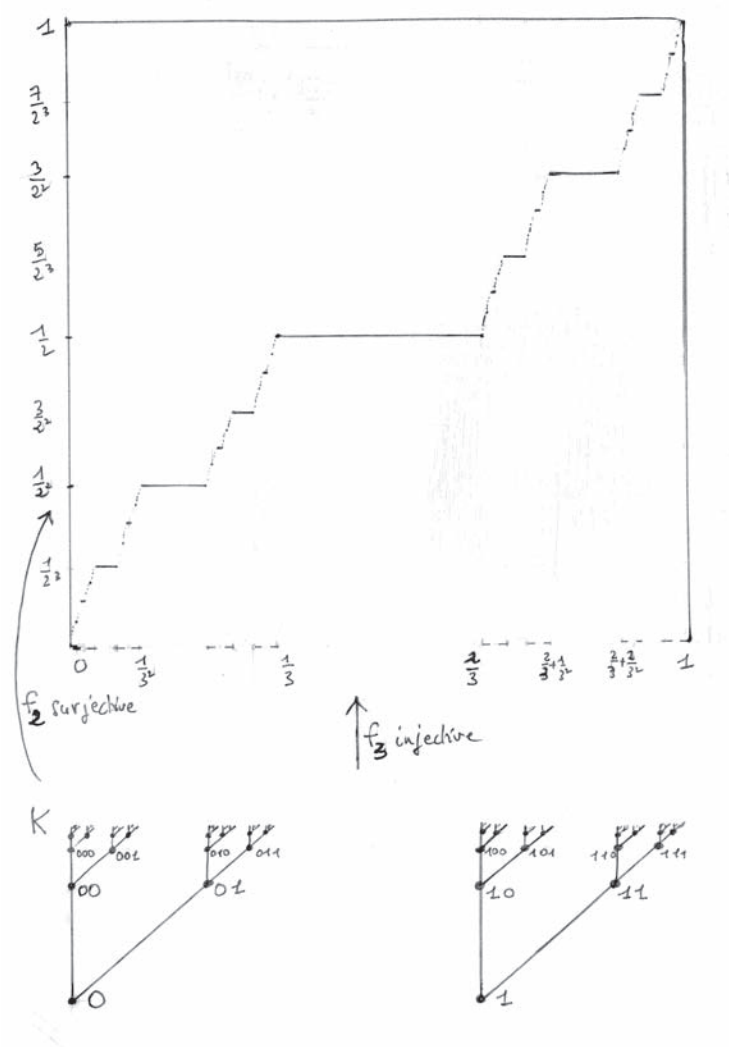


Figure 1 : l'escalier du diable.

1.1.5 Variables aléatoires

Définition 8 *Soit $(\Omega, \mathcal{F}, \mu)$ un espace de probabilité. Une application mesurable $\xi : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ à valeurs dans la droite réelle munie de sa tribu borélienne (ou plus généralement à valeurs dans un espace topologique X muni de sa tribu borélienne) est appelée une* variable aléatoire.

A une variable aléatoire ξ est associée la mesure de probabilité image $\xi_*\mu$ sur $(\mathbb{R}, \mathcal{B})$ (ou sur $(X, \mathcal{B})$). Notons que cette mesure image est souvent la seule à laquelle nous ayons vraiment accès.

Le cas le plus simple, celui où ξ ne prend qu'un nombre fini de valeurs, sera particulièrement important pour nous : soit $A = \{A_1, \dots, A_r\}$ l'ensemble de ces valeurs et $\xi_*\mu = p = (p_1, \dots, p_r)$ la loi de probabilité image : la probabilité d'un intervalle $I \subset \mathbb{R}$ (i.e. la probabilité que la valeur de la variable aléatoire ξ appartienne à I) est la somme $\sum_{A_i \in I} p_i$. La donnée de la variable aléatoire ξ équivaut à celle de l'espace de probabilité fini (A, p) ou encore à celle de la partition finie $\mathcal{P} : \Omega = P_1 + \cdots + P_r$ de l'espace source $(\Omega, \mathcal{F}, \mu)$ dont les atomes sont les images réciproques $P_i = \xi^{-1}(A_i)$ (rappelons que, par définition, $\mu(P_i) = p_i$). Les trois notions suivantes sont donc intimement liées :

- partition finie d'un espace de probabilité,
- variable aléatoire à valeurs finies,
- espace de probabilité fini

Remarque sur les notations. En théorie de la mesure on note en général $p(I)$ ce qu'en théorie des probabilités on note plutôt $\mu\{\xi \in I\}$ ou $Pr\{\xi \in I\}$.

Définition 9 *Deux variables aléatoires ayant les mêmes lois image sont dites* identiquement distribuées.

L'exemple typique de variables aléatoires à valeurs finies identiquement distribuées est donné par les

$$\xi_i : (\{0,1\}^{\mathbb{N}^*}, \mathcal{B}, P_{p,q}) \to \mathbb{R}, \quad \xi_i(a_1 a_2 \dots) = a_i.$$

Ceci vient de ce que la mesure $P_{p,q}$ ne fait pas intervenir l'indice i du tirage.

On laisse au lecteur le soin de généraliser ceci du fini au dénombrable (variables alétoires à valeurs discrètes).

1.1.6 Indépendance d'expériences, d'événements, de partitions, de tribus, de variables aléatoires ; probabilités conditionnelles

Le jeu de pile ou face décrit dans le paragraphe précédant est l'exemple typique d'une suite d'*expériences indépendantes*. Suivant Kolmogorov [Ko], commençons par définir cette notion. Une expérience sera assimilée à la donnée d'une partition mesurable (que l'on peut supposer finie) de l'ensemble Ω des événements élémentaires. Une telle partition correspond à l'ensemble des questions posées lors de l'expérience : par exemple, si Ω est l'intervalle $[0,1]$, la partition $[0,1] = [0, \frac{1}{10}[+[\frac{1}{10}, \frac{1}{2}]+]\frac{1}{2}, 1]$ correspond aux questions : "auquel de

ces trois intervalles le nombre donné par l'expérience appartient-il ?" (j'utilise la notation + de Kolmogorov pour les unions disjointes.)

Définition 10 *Soient $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \dots, \mathcal{E}^{(n)}$ des expériences correspondant aux partitions $\Omega = A_1^{(i)} + A_2^{(i)} + \dots + A_{r_i}^{(i)}$, $i = 1, \dots, n$. Ces expériences sont dites indépendantes si quels que soient les $k_i \in \{1, r_i\}$, $i = 1, \dots n$,*

$$p_{k_1 k_2 \dots k_n} := P\left(A_{k_1}^{(1)} \cap A_{k_2}^{(2)} \cap \dots \cap A_{k_n}^{(n)}\right) = P(A_{k_1}^{(1)}) P(A_{k_2}^{(2)}) \dots P(A_{k_n}^{(n)}).$$

Autrement dit, la probabilité pour que la première expérience donne le résultat $A_{k_1}^{(1)}$, la deuxième $A_{k_2}^{(2)}$...et la dernière $A_{k_n}^{(n)}$ est le produit des probabilités de chacun de ces événements. On dira également que les partitions correspondantes sont indépendantes. Enfin, des variables aléatoires à valeurs finies $\xi_1, \cdots, \xi_n : \Omega \to \mathbb{R}$ sont dites indépendantes si les partitions qu'elles définissent le sont.

Exercice 7 *Donner un exemple de trois partitions qui soient deux à deux indépendantes mais non indépendantes dans leur ensemble.*

Définition 11 *Les événements $A_1, A_2, \dots, A_n \subset \Omega$ sont dits indépendants si les partitions $\Omega = A_i + A_i^c$, $i = 1, 2, \dots n$, sont indépendantes.*

Exercice 8 *Montrer que les événements $A_1, A_2, \dots, A_n$ sont indépendants si et seulement si, pour tout $r \leq n$ et toute sous-famille $A_{i_1}, A_{i_2}, A_{i_r}$, on a l'égalité $P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_r}) = P(A_{i_1}) P(A_{i_2}) \dots P(A_{i_r})$.*

Enfin, on définit de même la notion de *tribus indépendantes* par la condition que la probabilité de l'intersection d'un nombre fini d'événements appartenant chacun à une tribu différente soit le produit des probabilités.

Une réécriture formelle mais fondamentale de la notion d'indépendance est celle en termes de probabilité conditionnelle :

Définition 12 *Soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité. Etant donnés A et B dans $\mathcal{F}$ avec $P(B) \neq 0$, la probabilité conditionnelle de A par rapport à B est*

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

Cette formule définit une nouvelle loi de probabilité sur Ω, qui est concentrée sur B. On peut également la considérer comme une loi de probabilité sur B, muni de la tribu $B \cap \mathcal{F}$.

Exercice 9 *L'indépendance de deux événements A et B équivaut, lorsque $P(B) \neq 0$, à l'égalité $P(A|B) = P(A)$: autrement dit, la réalisation de B n'a aucune influence sur celle de A.*

On définit plus généralement la probabilité conditionnelle de A par rapport à une partition, finie (ou dénombrable), ainsi que l'indépendance par rapport à une tribu (pour une présentation détaillée et de nombreux exemples, voir la section 33 de [B1]) :

Définition 13 *La probabilité conditionnelle de $A \subset \Omega$ par rapport à la partition $\Omega = B_1 + B_2 + \ldots + B_n$ est la variable aléatoire $f : \Omega \to \mathbb{R}$ qui prend la valeur $P(A|B_i)$ en $\omega \in B_i$.*

La fonction f est donc constante sur chacun de B_i et elle vérifie :

$$P(A \cap B_i) = \int_{B_i} f(\omega) dP(\omega).$$

En particulier, f est mesurable pour la tribu $\mathcal{T}$ engendrée par la partition et pour tout élément B de cette tribu (c'est-à-dire pour toute union de certains des B_i), on a de même $P(A \cap B) = \int_B f(\omega) dP(\omega)$. On notera dorénavant $P(A||\mathcal{T})$ la variable aléatoire f et on l'appellera *probabilité conditionnelle de A par rapport à la tribu $\mathcal{T}$.*

Plus généralement, étant donnée une tribu $\mathcal{T}$ quelconque, le théorème de Radon-Nikodym permet de définir une *probabilité conditionnelle de A par rapport à $\mathcal{T}$* : c'est une (classe modulo 0 de) fonction $P(A||\mathcal{T}) : \Omega \to \mathbb{R}$ qui est $\mathcal{T}$-mesurable et telle que, pour tout élément B de $\mathcal{T}$, on ait

$$P(A \cap B) = \int_B P(A||\mathcal{T}) dP.$$

1.1.7 Espérance, variance, loi faible des grands nombres

Pour plus de détails sur cette partie, consulter [Si]

Définition 14 *On appelle* espérance *d'une variable aléatoire $\xi : (\Omega, \mathcal{F}, \mu) \to \mathbb{R}$ sa moyenne*

$$E(\xi) = \int_\Omega \xi d\mu.$$

L'écart d'une variable aléatoire à son espérance est mesuré par sa variance :

Définition 15 *On appelle* variance *d'une variable aléatoire le nombre réel positif défini par*

$$Var\xi = \sigma_\xi^2 = E(\xi - E\xi)^2 = E(\xi^2) - (E\xi)^2.$$

Si ξ ne prend qu'un nombre fini de valeurs $A_1, \cdots, A_r$, notons $\Omega = C_1 + \cdots + C_r$ la partition de Ω définie par les $C_i = \xi^{-1}(A_i)$. Si $\mu(C_i) = p_i$, l'espérance et la variance de ξ deviennent respectivement

$$E\xi = \sum_{i=1}^r p_i A_i, \quad Var\xi = \sum_{i=1}^r p_i (A_i - E\xi)^2 = \sum_{i=1}^r p_i A_i^2 - \big(\sum_{i=1}^r p_i A_i\big)^2.$$

Dans la suite du paragraphe, nous nous placerons dans ce cas, laissant au lecteur le soin de généraliser ceci aux cas où ξ est à valeurs discrètes.

Le lemme qui suit, bien qu'élémentaire, implique, dans le cas de tirages indépendants, un résultat fondamental : la "loi faible des grands nombres" :

Lemme 6 (inégalité de Tschebishev) *Si ξ est une une variable aléatoire (i.e.≥ 0) à valeurs finies, et si $\alpha > 0 \in \mathbb{R}$, on a*

$$\mu\{\omega \in \Omega, \xi(\omega) \geq \alpha)\} \leq \frac{E\xi}{\alpha}.$$

Démonstration.

$$\mu\{\omega \in \Omega, \xi(\omega) \geq \alpha\} = \sum_{i, A_i \geq \alpha} p_i \leq \sum_{i, A_i \geq \alpha} p_i \frac{A_i}{\alpha} \leq \sum_i p_i \frac{A_i}{\alpha} = \frac{E\xi}{\alpha}.$$

Remarque sur la notation : les probabilistes ne notent en général pas la variable ω. L'inégalité ci-dessus est alors écrite simplement $Pr\{\xi \geq \alpha\} \leq \frac{E\xi}{\alpha}$.

Appliqué à la variable aléatoire positive $(\xi - E\xi)^2$, le lemme de Tschebishev devient

$$Pr\{|\xi - E\xi| \geq t\} \leq \frac{Var\xi}{t^2}.$$

Exercices (espérance et variance d'une somme de variables aléatoires).
1) Montrer que si $\xi_i, \xi_j : \omega \to \mathbb{R}$ sont des variables aléatoires à valeurs finies *indépendantes*, on a $E(\xi_i \xi_j) = E(\xi_i)E(\xi_j)$.
2) Montrer que si $\xi_1, \cdots, \xi_n : \Omega \to \mathbb{R}$ sont des variables aléatoires à valeurs finies *deux à deux indépendantes*, on a $Var(\xi_1 + \cdots + \xi_n) = Var\xi_1 + \cdots + Var\xi_n$.

Nous sommes maintenant en mesure de démontrer dans le cas le plus simple la loi faible des grands nombres, à la base de l'interprétation statistique de la notion même de probabilité. Nous considérons des variables aléatoires à valeurs finies indépendantes et identiquement distribuées ; ici encore, l'exemple le plus simple est donné par les

$$\xi_i : (\{0,1\}^{\mathbb{N}^*}, \mathcal{B}, P_{p,q}) \to \mathbb{R}, \quad \xi_i(a_1 a_2 \dots) = a_i.$$

Comme pour l'indépendance, le fait qu'elles soient identiquement distribuées vient de ce que la mesure $P_{p,q}$ ne fait pas intervenir l'indice i du tirage.

Théorème 7 (Loi faible des grands nombres dans le cas indépendant) *Si $\xi_1, \cdots, \xi_n : (\Omega, \mathcal{F}, \mu) \to \mathbb{R}$ sont des variables aléatoires* indépendantes *et* identiquement distribuées*, d'espérance m et de variance σ^2, on a*

$$\forall \epsilon > 0, \ Pr\left\{\left|\frac{\xi_1 + \cdots + \xi_n}{n} - m\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

En particulier, ϵ étant fixé, cette probabilité tend vers 0 lorsque n tend vers $+\infty$.

Démonstration. On applique l'inégalité de Tschebishev à la variable aléatoire $\left(\frac{s_n}{n} - E(\frac{s_n}{n})\right)^2$ et on utilise le résultat des exercices ci-dessus.

Appliquons ce théorème aux variables aléatoires

$$\xi : \{A_1, \cdots, A_r\}^{\mathbb{N}^*}, B, P_{p_1,\dots p_r}) \to \mathbb{R}, \quad \xi(a_1 a_2 \cdots) = \log \frac{1}{p(a_i)},$$

où le log est pris en base r et $p(a_i) = p_k$ si $a_i = A_k$. Notant $p(a_1 \cdots a_n) = p(a_1) \cdots p(a_n)$ la probabilité du cylindre $A_{1\cdots n}^{a_1 \cdots a_n}$ (qui est aussi la probabilité de $a_1 \cdots a_n \in A^n$), on obtient le

Théorème 8 Le premier théorème de Shannon dans le cas indépendant)

$$\forall \epsilon > 0,\ Pr\left\{ \left| \frac{1}{n} \log \frac{1}{p(a_1 \cdots a_n)} - \sum_{i=1}^{r} p_i \log \frac{1}{p_i} \right| \geq \epsilon \right\} \leq \frac{\sigma^2}{n\epsilon^2},$$

où $\sigma^2 = \sum_{i,j=1}^{r} p_i p_j^2 (\log \frac{p_j}{p_i})^2$ *et la probabilité est calculée avec la mesure* $P_{p_1,\cdots,p_r}$ *sur* $\{A_1, \cdots, A_r\}^{\mathbb{N}^*}$ *(ou, c'est équivalent, sur* $\{A_1, \cdots, A_r\}^n$*).*

la quantité $h = \sum_{i=1}^{r} p_i \log \frac{1}{p_i}$ est l'entropie de Shannon. Elle sera étudiée dans le chapitre 2. L'interprétation de ce théorème est que, si n est assez grand, on a une probabilité très grande de ne rencontrer que des suites (des messages) $a_1 \cdots a_n$ dont la probabilité est très proche de r^{-nh}. Il n'y a donc qu'environ r^{nh} de ces messages très probables parmi les r^n messages possibles de longueur n. Si $h = 1/2$, cela représente 100 messages parmi 10000 ! La figure qui suit, dans laquelle la taille des éléments est proportionnelle à leur probabilité, illustre ce fait.

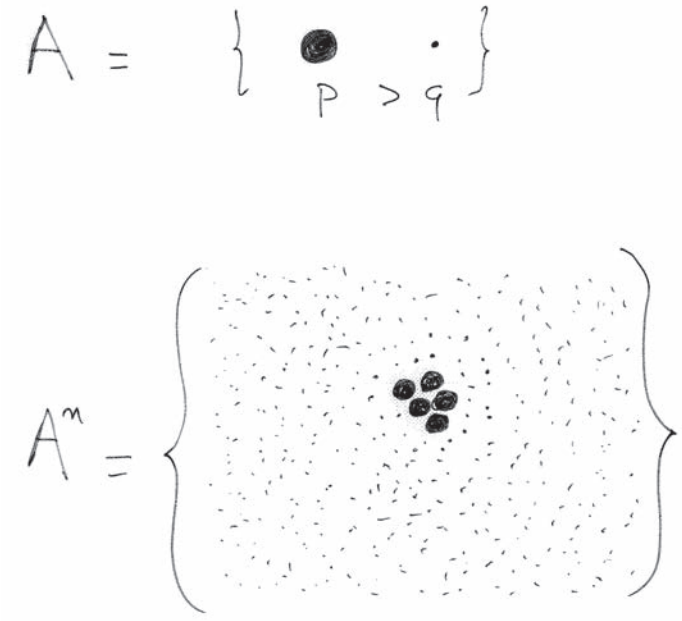


Figure 2 : le premier théorème de Shannon.

1.1.8 Le processus comme dynamique : les décalages de Bernoulli

Les propriétés stochastiques d'une suite infinie de tirages indépendants (avec probabilités respectivement p, q de $0, 1$), sont élégamment reflétées dans les propriétés dynamiques d'un unique objet, le *décalage de Bernoulli* (ou *Bernoulli shift*), qui est l'application

$$T : (\{0,1\}^{\mathbb{N}^*}, \mathcal{F}, P_{p,q}) \to (\{0,1\}^{\mathbb{N}^*}, \mathcal{F}, P_{p,q}), \quad T(a_1 a_2 a_3 \dots) = (a_2 a_3 a_4 \dots).$$

"Oubliant" a_1, cette application est surjective mais non injective : l'image réciproque de n'importe quel élément est formée de deux éléments. Sa propriété fondamentale est d'être une transformation *mesurable* qui *préserve les mesures* de probabilité $P = P_{p,q}$ sur la tribu borélienne $\mathcal{F}$ de $\{0,1\}^{\mathbb{N}^*}$ que l'on a associées à un couple (p, q) de probabilités de 0 et 1 : en effet, l'image réciproque $T^{-1}(A)$ du cylindre $A = A^{j_1 j_2 \dots j_k}_{i_1 i_2 \dots i_k}$ est le cylindre $A^{j_1 j_2 \dots j_k}_{i'_1 i'_2 \dots i'_k}$, où $i'_n = i_n + 1$ et a donc la même probabilité $p^{k_0} q^{k_1}$ que A puisque le processus est stationnaire. On applique le lemme 4 pour conclure.

Orbites et dynamique. Une *orbite* $\{\omega, T\omega, T^2\omega, \dots, T^n\omega, \dots\}$ de T est une description dynamique de la suite de tirages ω et le langage et les méthodes de la *théorie des systèmes dynamiques* – qui traite une telle orbite comme la version discrète d'une courbe intégrale d'une équation différentielle – s'avèrent remarquablement pertinentes et efficaces dans la description des processus stationnaires de ce type.

Exercice 10 *Lorsque $p = q = \frac{1}{2}$, la traduction de T dans le monde de l'intervalle $[0, 1]$ est l'application $x \mapsto 2x(mod.1) = 2x - [2x]$ dont on vérifiera directement qu'elle préserve la mesure de Lebesgue ($[x]$ désigne la partie entière).*

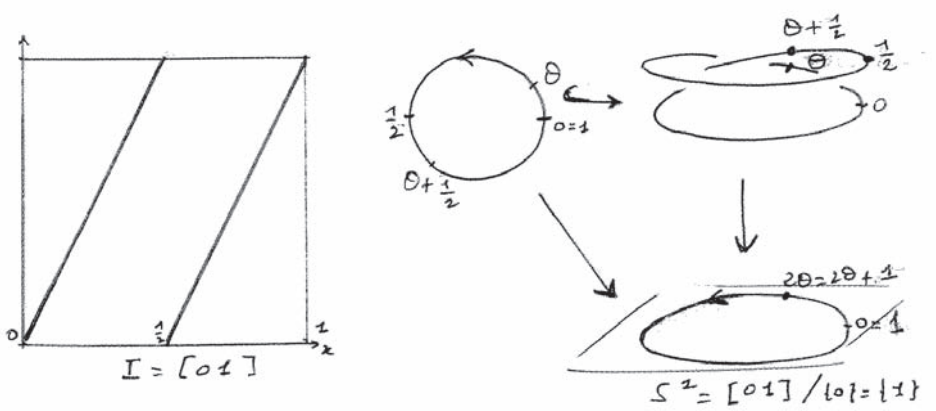


Figure 3 : L'application $x \mapsto 2x$ sur l'intervalle et sur le cercle.

Des suites simplement infinies aux suites doublement infinies :
Il est souvent plus agréable de travailler avec des transformations inversibles et, dans notre cas, c'est exactement ce qu'accomplit la considération d'une double infinité de tirages, à la fois dans le passé et dans le futur. On définit maintenant une bijection bimesurable $T : \{0,1\}^{\mathbb{Z}} \to \{0,1\}^{\mathbb{Z}}$ préservant toutes les mesures de probabilité $P = P_{p,q}$ par

$$T(\dots a_{-2}a_{-1}a_0a_1a_2\dots) = (\dots b_{-2}b_{-1}b_0b_1b_2\dots), \quad \text{où} \quad b_i = a_{i+1}.$$

Exercice 11 *On suppose que $p = q = \frac{1}{2}$. Montrer que si $g_2 : \{0,1\}^{\mathbb{Z}} \to [0,1]^2$ est l'application définie par*

$$g_2(\dots a_{-2}a_{-1}a_0a_1a_2\dots) = \left(\sum_{k=0}^{-\infty} \frac{a_k}{2^{1-k}}, \sum_{k=1}^{\infty} \frac{a_k}{2^k}\right),$$

l'image directe de P par g_2 est la mesure de Lebesgue sur $[0,1]^2$ et g_2 conjugue (mesurablement) $T : \{0,1\}^{\mathbb{Z}} \to \{0,1\}^{\mathbb{Z}}$ à la transformation $\tau : [0,1]^2 \to [0,1]^2$ définie par

$$\tau(x,y) = \big(\frac{1}{2}(x + [2y]), 2y - [2y]\big),$$

où $[2y]$ désigne le plus grand entier $\leq 2y$ (bien entendu, $2y - [2y]$ n'est autre que $2y (mod 1)$). On expliquera pourquoi τ est appelée "transformation du boulanger" par les dynamiciens.

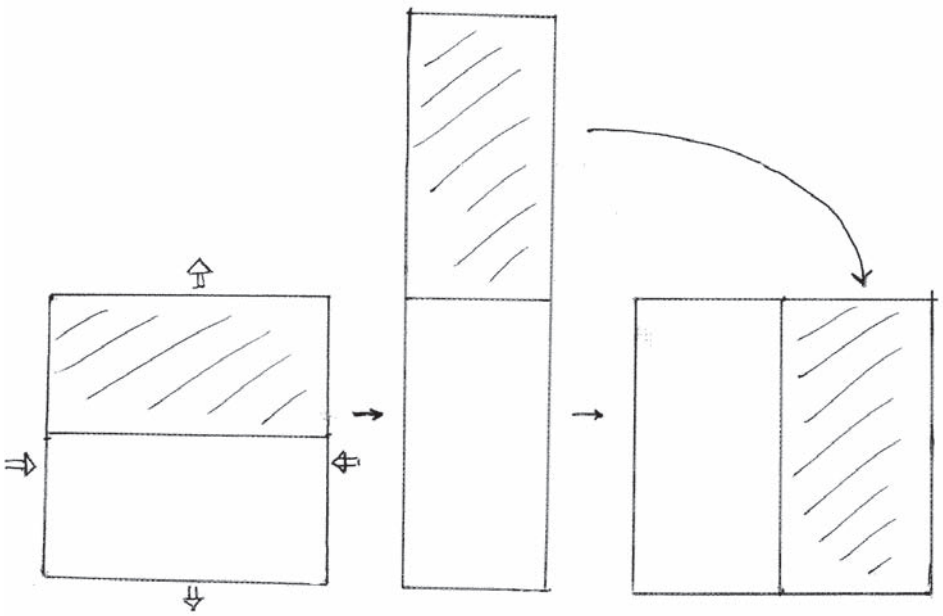


Figure 4 : La transformation du boulanger.

Processus stochastiques stationnaires et décalages : dictionnaire
Nous avons jusqu'ici supposé que les tirages étaient indépendants et cela s'est traduit par le choix de la mesure de probabilité $P = P_{p,q}$ sur $\{0,1\}^{\mathbb{Z}}$. Dans la théorie de l'information, on envoie des messages dont la structure n'est pas en général aussi simple : dans une langue donnée, une lettre n'est pas suivie de façon équiprobable de n'importe quelle lettre et ceci conduit à la considération de processus plus complexes, par exemple les *chaînes de Markov* que nous étudierons plus loin.

Définition 16 (source discrète) *On appelle source dicrète (stationnaire) la donnée d'une mesure de probabilité T-invariante sur la tribu borélienne de $\{0,1\}^{\mathbb{Z}}$ (resp. dans $A^{\mathbb{Z}}$ où $A = \{A_1, A_2, \dots, A_r\}$ est un alphabet fini).*

Une telle donnée équivaut à la donnée d'un processus stochastique stationnaire à valeurs dans $\{0,1\}$ (resp. dans A) au sens de la définition suivante :

Définition 17 *Un processus stochastique à temps discret sur l'espace de probabilité $(\Omega, \mathcal{F}, \mu)$ à valeurs dans l'espace topologique X est une suite $\{\xi_n\}_{n\in\mathbb{Z}}$ de variables aléatoires $\xi_n : \Omega \to X$.*

On associe à un tel processus une loi de probabilité ν sur $X^{\mathbb{Z}}$ en posant

$$\nu(A_{i_1 i_2 \dots i_k}^{j_1 j_2 \cdots j_k}) = \mu\big(\tilde{\xi}^{-1}(A_{i_1 i_2 \dots i_k}^{j_1 j_2 \cdots j_k})\big),$$

où $\tilde{\xi}(\omega) = \dots \xi_{-2}(\omega)\xi_{-1}(\omega)\xi_0(\omega)\xi_1(\omega)\xi_2(\omega)\dots$, autrement dit en définissant cette loi comme l'image directe $\nu = \tilde{\xi}_*\mu$ de μ par l'application $\tilde{\xi}$. Le processus est dit *stationnaire* si ν est invariante par le décalage T. Si les variables aléatoires ξ_i sont indépendantes, également distribuées et à valeurs dans l'alphabet fini $X = (A_1, \cdots, A_r)$ avec pour loi image $(p_1, \cdots, p_r)$, la mesure $\nu = \tilde{\xi}$ sur $X^{\mathbb{Z}}$ n'est autre que $P_{p_1,\cdots,p_r}$.

1.2 Ergodicité et théorème de Birkhoff

L'indépendace des tirages d'un jeu de pile ou face (doublement) infini peut se traduire en un "oubli des conditions initiales" : chaque tirage ignore le résultat de tous les tirages précédents ; à cet oubli correspond une propriété très forte du décalage de Bernoulli T, appelée *ergodicité* :

Définition 18 *Soit $f : (X, \mathcal{A}, P) \to (X, \mathcal{A}, P)$ une application préservant la mesure de probabilité P. On dit que f est ergodique si tout ensemble $A \in \mathcal{A}$ invariant par f vérifie $P(A) = 0$ ou $P(A) = 1$. On dit aussi, f étant sous-entendue, que la mesure P est ergodique.*

Dans cette définition,; il faut entendre le mot *invariant* au sens de la mesure, i.e. modulo des ensembles de mesure nulle. Plus précisément :

Définition 19 *Soit $f : (X, \mathcal{A}, P) \to (X, \mathcal{A}, P)$ une application préservant la mesure de probabilité P. On dit que $A \subset X$ est invariant si $P\big(f^{-1}(A)\Delta A\big) = 0$, où $U\Delta V = (U \cap V^c) \cup (V \cap U^c)$ est la différence symétrique de U et V.*

Cette définition suppose bien entendu que A appartienne à la tribu $\mathcal{A}$ (à laquelle on a ajouté les ensembles de mesure nulle). On peut montrer qu'une partie A invariante en ce sens est presque partout égale à une partie B invariante au sens strict, i.e. $P(A\Delta B) = 0$ et $f^{-1}(B) = B$.

Exemple Munissons le cercle unité R^2 de sa mesure des angles. Une rotation d'angle α est ergodique si et seulement si $\frac{\alpha}{2\pi}$ est irrationnel. Une très jolie démonstration directe basée sur la densité des orbites d'une telle rotation se trouve dans Billingsley. Une démonstration beaucoup plus simple mais plus mystérieuse est basée sur les séries de Fourier (voir [AA]).

Tester directement une propriété sur un élément quelconque d'une tribu $\mathcal{F}$ peut être difficile alors même qu'il est facile de la vérifier sur les éléments d'un ensemble de générateurs (par exemple les éléments d'une algèbre $\mathcal{G}$ telle que $\mathcal{F} = \sigma(\mathcal{G})$). Le lemme suivant permet souvent de se restreindre à cette seule vérification.

Lemme 9 *Soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité et $\mathcal{G}$ une algèbre engendrant $\mathcal{F}$ (i.e. $\mathcal{F} = \sigma(\mathcal{G})$). Quels que soient $A \in \mathcal{F}$ et $\epsilon > 0$, il existe $A_0 \in \mathcal{G}$ tel que $P(A\Delta A_0) < \epsilon$.*

Autrement dit, tout élément de la tribu est approché arbitrairement près par un élément de l'algèbre.

Esquisse de démonstration. Elle découle de la construction même de l'extension de $\mathcal{G}$ à $\sigma(\mathcal{F})$ de la probabilité P : par définition de $P^*(A)$ comme le inf des sommes $\sum_{n=1}^{\infty} P(A_n)$ sur l'ensemble des suites d'éléments de $\mathcal{G}$ dont la réunion contient A (démonstration du théorème 3), il existe, pour tout $\epsilon > 0$, une suite C_n d'éléments de $\mathcal{G}$ telle que $A \subset \cup_{i=n}^{\infty} C_n$ et $\sum_{i=n}^{\infty} P(C_n) < P(A) + \frac{1}{2}\epsilon$. Puisque la suite des $P(C_n)$ converge, il existe un entier N tel que $\sum_{n=N+1}^{\infty} C_n < \frac{1}{2}\epsilon$. On montre alors (exercice) que $A_0 = \sum_{i=1}^{N} C_n$ convient.

Voici enfin quelques propriétés de la différence symétrique qui permettront en particulier d'utiliser le lemme d'approximation :

Exercice 12 *1) Montrer que quels que soient U, V, W sous-ensembles de X, on a $U\Delta V \subset (U\Delta W) \cup (W\Delta V)$ et donc $P(U\Delta V) \leq P(U\Delta W) + P(W\Delta V)$.*

Exercice 13 *Montrer que, si $P(A\Delta A_0) < \epsilon$ et $P(B\Delta B_0) < \epsilon$, on a*

$$P\big((A \cap B)\Delta(A_0 \cap B_0)\big) < 2\epsilon \quad \textit{et} \quad P\big((A\Delta B)\Delta(A_0\Delta B_0)\big) < 2\epsilon.$$

1.2.1 L'ergodicité du décalage de Bernoulli

Pour montrer l'ergodicité des décalages de Bernoulli, nous allons montrer une propriété plus forte, le *mélange* :

Définition 20 *Soit $f : (X, \mathcal{A}, P) \to (X, \mathcal{A}, P)$ une application préservant la mesure de probabilité P. On dit que f est mélangeante (ou mieux que f possède la propriété de mélange) si quels que soient $A, B \in \mathcal{A}$, on a*

$$\lim_{n\to\infty} P(f^{-n}(A) \cap B) = P(A)P(B).$$

Cette propriété entraine immédiatement l'ergodicité car si A est invariant, $A = f^{-n}(A)$ mod 0 et le choix de $B = A$ donne $P(A) = P(A)^2$. Elle est en fait strictement plus forte, l'exemple typique étant une rotation non périodique du cercle muni de sa mesure naturelle (voir n'importe quel livre de théorie ergodique, par exemple [AA, B2, M1]).

Théorème 10 *Le décalage de Bernoulli T sur $\{0,1\}^{\mathbb{N}^*}$ ou sur $\{0,1\}^{\mathbb{Z}}$ possède la propriété de mélange (et est en conséquence ergodique) pour n'importe laquelle des mesures de probabilité produit $P = P_{p,q}$.*

Démonstration. On commence par déduire du lemme 6 et de l'exercice 9 qu'il suffit de vérifier la propriété de définition du mélange sur les éléments de l'algèbre $\mathcal{G}$, c'est-à-dire sur les unions finies de cylindres. En effet, si $A, B \in \mathcal{F}$ et si $A_0, B_0 \in \mathcal{G}$ vérifient $P(A\Delta A_0) < \epsilon$ et $P(B\Delta B_0) < \epsilon$, on a, pour tout n, $P(f^{-n}(A)\Delta f^{-n}(A_0)) = P(A\Delta A_0) < \epsilon$ par préservation de la mesure et donc $P\big((f^{-n}(A) \cap B)\Delta(f^{-n}(A_0) \cap B_0)\big) < 2\epsilon$. On en déduit que $|P(f^{-n}(A) \cap B) - P(f^{-n}(A_0) \cap B_0)| < 2\epsilon$ et donc que les lim sup et lim inf de $P(f^{-n}(A) \cap B)$ diffèrent de celles de $P(f^{-n}(A_0) \cap B_0)$ d'au plus 2ϵ.
Mais étant donnés deux unions finies de cylindres A_0 et B_0, les ensembles d'indices mis en jeu par A et $T^{-n}(B_0)$ sont disjoints dès que n est assez grand, ce qui implique que $P(f^{-n}(A_0) \cap B_0) = P(A_0)P(B_0)$. Je laisse le soin au lecteur de terminer la démonstration.

Corollaire 11 *L'application $x \mapsto 2x(mod.1)$ de $[0,1]$ dans lui-même et l'application du boulanger τ de $[0,1]^2$ dans lui-même sont mélangeantes, et donc ergodiques, pour la mesure de Lebesgue.*

1.2.2 Le théorème ergodique de Birkhoff

La *théorie ergodique* a pour origine les travaux de mécanique statistique de Boltzmann. Sa forme proprement mathématique date des années 1930, avec les thérèmes ergodiques de Von Neumann, Birkhoff et Koopman, qui sont des formes dynamiques très fortes de la loi des grands nombres, mais on peut aussi la faire remonter au *théorème de récurrence de Poincaré* [Po] qui exploite déjà de manière fine les contraintes qu'impose à un système dynamique la préservation d'une mesure de masse totale finie. Dans ce qui suit, on se permettra systématiquement l'abus de notation qui consiste à identifier une fonction et l'élément qu'elle définit dans $L^1(X, \mathcal{F}\mu)$.

Théorème 12 *Soit $(X, \mathcal{F}, \mu)$ un espace de probabilité et $T : (X, \mathcal{F}, \mu) \to (X, \mathcal{F}, \mu)$ une application préservant la mesure. Quelle que soit la fonction $f \in L^1(X, \mathcal{F}, \mu)$, la limite des "sommes de Birkhoff"*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f\big(T^k(x)\big) := f^*(x)$$

existe pour μ-presque tout $x \in X$ et définit une fonction $f^* \in L^1(X, \mathcal{F}, \mu)$ qui vérifie $f^* \circ T = f^*$ (μ-presque partout) et $\int_X f(x)\, d\mu(x) = \int_X f^*(x)\, d\mu(x)$. Si T est inversible, les fonctions f^* et $\bar{f}^*$ définies respectivement par T et T^{-1} coincident presque partout.

Corollaire 13 *Sous les mêmes hypothèses, si de plus T est ergodique, f^* est une constante, égale à l'intégrale $\int_X f d\mu$.*

En mots, cela signifie que si T est ergodique, la *moyenne temporelle*, définie comme la limite des *sommes de Birkhoff* existe presque partout et est égale à l'intégrale, c'est-à-dire à la moyenne spatiale. Ce résultat est particulièrement parlant si l'on choisit pour f la fonction caractéristique $\mathcal{X}_A$ d'une partie mesurable $A \in \mathcal{F}$. Le corollaire affirme alors que la proportion de "temps" passé dans A par l'orbite de x coïncide, pour presque tout x avec la mesure (la probabilité) de A ($n \in \mathbb{N}^*$ ou $n \in \mathbb{Z}$ doit être interprétée comme un temps discret, l'unité correspondant à une itération de la transformation T).

1.2.3 Loi forte et loi faible des grands nombres

Appliqué au décalage de Bernoulli, que nous dit le corollaire précédent ? Ceci que la structure statistique de presque toutes les suites est la même, c'est-à-dire une forme très forte de la *loi des grands nombres*. L'énoncé de cette loi, à la base de l'interprétation "fréquentielle" des probabilités, est le suivant *(comme précédemment nous ne considérons que le cas de variables aléatoires à valeurs finies)* :

Théorème 14 (Loi forte des grands nombres dans le cas indépendant) *Si $\xi_1, \cdots, \xi_n, \cdots : (\Omega, \mathcal{F}, \mu) \to \mathbb{R}$ sont des variables aléatoires indépendantes et identiquement distribuées de moyenne m, on a*

$$Pr\left\{ \lim_{n\to\infty} \frac{1}{n}(\xi_1 + \cdots + \xi_n) = m \right\} = 1.$$

Démonstration. Soient $A_1, \cdots, A_r \subset \mathbb{R}$ les valeurs prises par les variables aléatoires ξ_i et $(p_1, \cdots, p_r)$ leurs probabilités (qui sont bien définies par l'hypothèse d'identique distribution). On applique le théorème de Birkhoff, ou plutôt son corollaire, au décalage

$$T : (\{A_1, \cdots, A_r\}^{\mathbb{N}^*} \mathcal{B}, P_{p_1,\cdots,p_r}) \to : (\{A_1, \cdots, A_r\}^{\mathbb{N}^*} \mathcal{B}, P_{p_1,\cdots,p_r})$$

et à la fonction, évidemment intégrable, $f(a_1 \cdots a_n \cdots) = a_1$. La conclusion suit de ce que, d'une part $\xi_1(\omega) + \cdots + \xi_n(\omega) = \sum_{k=0}^{n-1} T^k \xi_1(\omega) = \sum_{k=0}^{n-1} T^k f(\tilde{\xi}(\omega))$, d'autre part l'intégrale de f sur $\{A_1, \cdots, A_r\}^{\mathbb{N}^*}$ est égale à $\sum_{i=1}^{r} p_i A_i = m$.

Théorème 15 *La loi forte implique la loi faible*

Démonstration. Elle est d'un type classique en théorie des probabilités et est en général énoncée sous la forme *"La convergence avec probabilité 1 implique la convergence en probabilité".* On considère une suite X_n de variables aléatoires (les $\frac{1}{n}\xi_n$ du théorème 22) qui converge avec probabilité 1 vers une variable aléatoire X (la constante égale à l'entropie dans le théorème 22) :

$$\mu\{\omega \in \Omega,\ \lim_{n\to\infty}(X_n(\omega) = X(\omega)\} = 1.$$

Notons L le sous-ensemble de mesure pleine ci-dessus. Son complémentaire peut s'écrire

$$L^c = \cup_\epsilon\{\omega \in \Omega,\ |X_n(\omega) - X(\omega| \geq \epsilon \text{ pour une infinité de } n\}.$$

On peut bien entendu se restreindre à l'ensemble dénombrable des ϵ rationnels, ce qui montre en passant que l'ensemble L est mesurable (i.e. qu'il appartient à la σ-algèbre $\mathcal{F}$ de Ω). Le corollaire découle alors de la

Proposition 16 *Supposons que*

$$\mu\{\omega \in \Omega,\ |X_n(\omega) - X(\omega| \geq \epsilon \textit{ pour une infinité de } n\} = 0\,;$$

Alors

$$\lim_{n\to\infty} \mu\{\omega \in \Omega,\ |X_n - X| \geq \epsilon\} = 0.$$

Si cette dernière propriété est vérifiée pour tout $\epsilon > 0$, on dit que X_n converge en probabilité vers X.

Démonstration. Notons $G_n = \{\omega \in \Omega,\ |X_n(\omega) - X(\omega| \geq \epsilon\}$. L'ensemble des ω qui appartiennent à G_n pour une infinité de valeurs de n s'appelle la lim sup des G_n et peut être défini par la formule :

$$\limsup_n G_n = \cap_{n=1}^\infty \cup_{k=n}^\infty G_k.$$

On déduit de cette écriture que

$$\limsup_n \mu(G_n) \leq \mu(\limsup_n G_n).$$

En effet, $\limsup G_n$ est l'intersection des $U_n = \cup_{k=n}^\infty G_k$ qui forment une famille décroissante d'ensembles (i.e. $U_{n+1} \subset U_n$). Donc $\lim_{n\to\infty} \mu(U_n) = \mu(\limsup G_n)$. Mais $G_n \subset U_n$ donc $\mu(G_n) \leq \mu(U_n)$ et enfin

$$\limsup(\mu(G_n) \leq \limsup \mu(U_n) = \lim \mu(U_n) = \mu(\limsup G_n).$$

L'hypothèse de la proposition est que $\mu(\limsup G_n) = 0$. On en déduit que $\limsup \mu(G_n) \leq 0$ et donc, puisque μ est à valeurs positives, que $\lim \mu(G_n) = 0$.

Un exemple de résultats plus précis Considérons dans $(\{0,1\}^{\mathbb{N}^*}, \mathcal{B}, P_{p,q})$ le cylindre A défini par $a_1 = a_2 = \ldots = a_{1000} = 0$. La somme de Birkhoff $\frac{1}{n}\sum_{k=0}^{n-1} \mathcal{X}_A\big(T^k(x)\big)$, où T est le décalage, représente la fréquence avec laquelle on a $a_{k+1} = a_{k+2} = \ldots = a_{k+1000} = 0$ lorsque k varie de 0 à n. Le théorème affirme que, pour presque toute suite, cette fréquence tend vers une limite, égale à p^{1000}, lorsque n tend vers $+\infty$. Il en est de même avec tout cylindre, c'est-à-dire avec tout motif fini de 0 et de 1 mais ceci n'épuise pas la richesse du théorème puisque la fonction f peut dépendre d'un nombre quelconque de coordonnées.

Exercice 14 *Appliquer le théorème ergodique au même exemple sous le déguisement de l'application $T(x = 2x$ (mod 1) de l'intervalle $[0,1]$ muni de la tribu borélienne et de la mesure de Lebesgue dans lui-même. En déduire une démonstration de l'affirmation qui clôt le paragraphe 1.1.3 (presque tout élément de $[0,1]$ est normal au sens de Borel, i.e. appartient à $\mathcal{N}$).*

On remarquera que, de la définition de l'ergodicité, on pouvait conclure seulement que la mesure de $\mathcal{N}$ valait 0 ou 1.

Exercice 15 (forme forte du théorème de Shannon (cas indépendant)) *Si $\xi_1, \cdots, \xi_n, \cdots : (\Omega, \mathcal{F}, \mu) \to \mathbb{R}$ sont des variables aléatoires indépendantes et identiquement distribuées à valeurs dans $\{A_1, \cdots, A_r\}$ avec loi image $(p_1, \cdots, p_r)$, on a*

$$Pr\left\{\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{p(a_1 \cdots a_n)} = \sum_{i=1}^{r} p_i \log \frac{1}{p_i}\right\} = 1.$$

1.2.4 Démonstration du théorème ergodique de Birkhoff

La démonstration donnée ci-dessous est celle de Katznelson et Weiss *(A simple proof of some ergodic theorems, Israel Journal of Mathematics 42 (1982), pages 291-296).* Inspirée par un travail de Kamae qui utilise les méthodes de l'Analyse Non Standard, elle est beaucoup plus simple que celle de Birkhoff. J'ai choisi de l'exposer en détail car ce théorème fait partie du très petit nombre de résultats signifiants dont on dispose sur les Systèmes Dynamiques généraux. Pour une autre démonstration, très courte, voir [KH].

Notons

$$\overline{f}(x) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f\big(T^i(x)\big),\ \underline{f}(x) = \liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f\big(T^i(x)\big).$$

Il suffit de montrer que pour μ presque tout x, on a l'inégalité

$$\overline{f}(x) \leq \underline{f}(x).$$

Ecrivant $f \in L^1(X, \mu)$ comme différence de fonctions intégrables positives, on peut supposer que f est positive, donc également $\overline{f}$ et $\underline{f}$. Le Théorème découle alors de la

Proposition 17 *Si $f \in L^1(X,\mu)$ est positive ($f \geq 0$), on a*

$$\int_X \overline{f}(x)\, d\mu(x) \leq \int_X f(x)\, d\mu(x) \leq \int_X \underline{f}(x)\, d\mu(x)$$

.

Démonstration de la Proposition. On commence par se débarasser d'éventuels points $x \in X$ en lesquels $\overline{f}(x) = +\infty$ en remplaçant $\overline{f}(x)$ par

$$\overline{f}_M(x) = \min\{\overline{f}(x), M\}$$

et en remarquant qu'il suffit, par le *théorème de convergence monotone* ([R], chapitre 1), de montrer que pour tout M,

$$\int_X \overline{f}_M(x)\, d\mu(x) \leq \int_X f(x)\, d\mu(x).$$

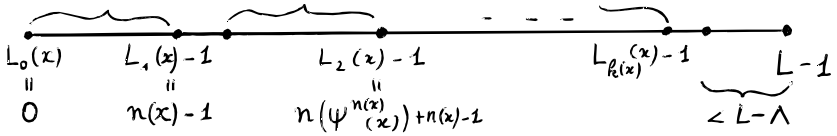
Par définition de la lim sup, on peut alors affirmer que

$$\forall \epsilon > 0, \forall x \in X, \exists n(x),\ \overline{f}_M(x) \leq \frac{1}{n(x)} \sum_{i=0}^{n(x)-1} f\big(T^i(x)\big) + \epsilon.$$

Puisque $\overline{f}_M \circ T = \overline{f}_M$, cette inégalité s'écrit encore

$$\sum_{i=0}^{n(x)-1} \overline{f}_M\big(T^i(x)\big) \leq \sum_{i=0}^{n(x)-1} f\big(T^i(x)\big) + n(x)\epsilon.$$

Si maintenant L est un entier assez grand, on peut, pour chaque élément x de X, décomposer l'ensemble $\{0, 1, \cdots, L-1\}$ en morceaux sur lesquels (à l'exception du dernier) une inégalité du type ci-dessus est vérifiée :



Définissant par récurrence $L_0(x) = 0$, $L_{i+1}(x) = L_i(x) + n\big(T^{L_i(x)}(x)\big)$ et notant $k(x) = \sup\big\{i, L_i(x) \leq L\big\}$, on obtient

$$\sum_{i=0}^{L-1} \overline{f}_M\big(T^i(x)\big) \leq \sum_{i=0}^{L_{k(x)}(x)-1} f\big(T^i(x)\big) + \sum_{i=L_{k(x)}(x)}^{L-1} \overline{f}_M\big(T^i(x)\big) + L_{k(x)}(x)\epsilon,$$

et a fortiori (puisque f est positive)

$$\sum_{i=0}^{L-1} \overline{f}_M\big(T^i(x)\big) \leq \sum_{i=0}^{L-1} f\big(T^i(x)\big) + (L-\Lambda)M + L\epsilon,$$

où $\Lambda = \inf_{x\in X} L_{k(x)}(x)$. Intégrant sur X et divisant par L, on obtient enfin

$$\int_X \overline{f}_M(x)\, d\mu(x) \leq \int_X f(x)\, d\mu(x) + \left(\frac{L-\Lambda}{L} M + \epsilon\right).$$

A priori, on ne peut pas choisir une fonction $x \mapsto n(x)$ bornée, ce qui exclut de contrôler Λ, mais on peut cependant la choisir mesurable. Les sous-ensembles

$$X_N = \{x \in X,\, n(x) \leq N\}$$

sont alors mesurables et leur réunion croissante épuise X. Le nœud de la preuve consiste en la remarque suivante : si l'on définit $\tilde{n}(x)$ et $\tilde{f}(x)$ par

$$\tilde{n}(x) = n(x) \text{ si } x \in X_N, \quad \tilde{n}(x) = 1 \text{ si } x \in X - X_N,$$

$$\tilde{f}(x) = f(x) \text{ si } x \in X_N, \quad \tilde{f}(x) = \sup\big(M, f(x)\big) \text{ si } x \in X - X_N,$$

on a comme précédemment, pour tout $x \in X$,

$$\sum_{i=0}^{\tilde{n}(x)-1} \overline{f}_M\big(T^i(x)\big) \leq \sum_{i=0}^{\tilde{n}(x)-1} \tilde{f}\big(T^i(x)\big) + \tilde{n}(x)\epsilon,$$

et donc

$$\int_X \overline{f}_M(x)\, d\mu(x) \leq \int_X \tilde{f}(x)\, d\mu(x) + \left(\frac{L-\Lambda}{L} M + \epsilon\right),$$

avec maintenant $L - \Lambda < N$. Finalement,

$$\int_X \overline{f}_M(x)\, d\mu(x) \leq \int_X f(x)\, d\mu(x) + M\mu(X - X_N) + \left(\frac{N}{L} M + \epsilon\right).$$

M et ϵ étant donnés, choisissons N de façon que $M\mu(X - X_N) < \epsilon$, puis L de façon que $NM < L\epsilon$. Faisant tendre ϵ vers 0, il vient

$$\int_X \overline{f}_M(x)\, d\mu(x) \leq \int_X f(x)\, d\mu(x).$$

Le lecteur démontrera de même la deuxième inégalité de la Proposition. Le point fondamental a donc été de rendre finie (choix de M) puis uniformément bornée (remplacement par $\tilde{n}$) la fonction n.

Le cas où T est inversible : on note

$$f^{*+}(x) = \lim_{n\to+\infty} \frac{1}{n} \sum_{j=0}^{n-1} f\big(T^j(x)\big),\ f^{*-}(x) = \lim_{n\to+\infty} \frac{1}{n} \sum_{j=0}^{n-1} f\big(T^{-j}(x)\big).$$

Le Théorème de Birkhoff implique que ces deux limites existent et appartiennent à $L^1(X, \mu)$. Supposons que le sous-ensemble

$$Y = \left\{x \in X,\, f^{*+}(x) > f^{*-}(x)\right\},$$

bien défini et invariant par T, à un ensemble de mesure nulle près, soit de mesure positive. Appliquant le Théorème de Birkhoff aux restrictions à Y de f, T ou T^{-1}, μ, on obtient

$$\int_Y f^{*+}(x)\, d\mu(x) = \int_Y f(x)\, d\mu(x) = \int_Y f^{*-}(x)\, d\mu(x),$$

donc

$$\int_Y (f^{*+} - f^{*-})(x)\, d\mu(x) = 0,$$

ce qui est impossible puisque $f^{*+} - f^{*-}$ est positive sur Y. On conclut en itérant le raisonnement après remplacement de $>$ par $<$.

1.2.5 Espérance conditionnelle et théorème de Birkhoff

Définition 21 *L'espérance conditionnelle d'une variable aléatoire $\xi : \Omega \to \mathbb{R}$ par rapport à une partition $\Omega = B_1 + B_2 + \ldots + B_n$ est la variable aléatoire qui prend la valeur constante $\int_{B_i} \xi dP$ sur chacune des parties B_i de la partition.*

Plus généralement, on montre qu'on peut définir comme ci-dessus l'*espérance conditionnelle d'une variable aléatoire par rapport à une tribu* comme une (classe modulo 0 de) variable aléatoire $E(\xi||\mathcal{T})$ qui est $\mathcal{T}$-mesurable et vérifie

$$\int_B \xi dP = \int_B E(\xi||\mathcal{T}) dP$$

pour tout $B \in \mathcal{T}$. Je renvoie au paragraphe 34 de [B1] pour son existence.

Attention, comme dans le cas des probabilités conditionnelles, la condition de $\mathcal{T}$-mesurabilité est ici fondamentale : par exemple, dans le cas d'une tribu $\mathcal{T}$ engendrée par une partition, la mesurabilité signifie que la fonction $E(\xi||\mathcal{T})$ est constante sur chaque morceau de la partition : sa valeur sur le morceau B_i est simplement l'espérance de sa restriction $\int_{B_i} \xi dP$, ce qui rejoint la définition précédente. Les notions de probabilité conditionnelle et d'espérance conditionnelle sont d'ailleurs liées par la relation $P(A||\mathcal{T}) = E(\mathcal{X}_A||\mathcal{T})$, où $\mathcal{X}_A$ désigne la fonction caractéristique (ou indicatrice) de A.

Avec cette dernière définition, on peut réinterpréter le théorème ergodique de Birkhoff comme affirmant la convergence des sommes de Birkhoff vers l'espérance conditionnelle f^* de f par rapport à la tribu des sous-ensembles T-invariants : avant même de démontrer le théorème, on sait que, si elle existe, la limite des sommes de Birkhoff ne peut être que cette espérance conditionnelle f^*.

1.3 Processus à mémoire et décalages de Markov

Après les tirages indépendants, auxquels correspondent les mesures de probabilité $P_{p,q}$ invariantes par le décalage sur $\{0,1\}^{N^*}$ ou sur $\{0,1\}^{\mathbb{Z}}$, viennent les *chaînes de Markov*, dans lesquels chaque tirage (ou émission) d'un symbole de l'alphabet dépend d'un nombre fini des tirages précédents. Il leur correspond

des mesures de probabilités invariantes par le décalage et l'on peut énoncer des conditions nécessaires et suffisantes sur le processus pour que le décalage soit ergodique ou mélangeant. De bonnes références sont [M1, Si].

1.3.1 Chaînes de Markov

Une chaîne de Markov se définit bien à partir de son *graphe* :
– chaque sommet (encore appelé *état*) est caractérisé par l'ensemble des traits dont la mémoire influence le tirage (*"the residue of influence" from preceding letters*, dit Shannon); à chacun est associée une distribution de probabilité sur l'alphabet A (que l'on supposera fini).
– chaque arête joint un sommet à un autre (éventuellement lui-même) et caractérise l'émission d'un élément de A. Elle est munie d'une probabilité caractérisant la probabilité de l'émission de cette lettre à partir de cet état. La *probabilité de transition* entre deux états est la somme des probabilités des arêtes qui les joignent.

Dans le cas d'un processus de Bernoulli (tirages indépendants d'éléments de $A = \{A_1, A_2, \dots, A_r\}$), il y a un seul état et r arêtes joignant à lui-même l'unique sommet correspondant. Ces arêtes sont associées respectivement aux probabilités $p_1, p_2, \dots, p_r$.

Dans le cas d'un processus n'ayant la mémoire que du seul tirage qui précède immédiatement, il y a, dans le cas général, autant d'états que d'éléments dans A. Supposant toujours que $A = \{A_1, A_2, \dots, A_r\}$, l'arête joignant l'état A_i à l'état A_j est munie de la *probabilité conditionnelle* p_{ij}, qui est la *probabilité de tirer (émettre)* A_j *après que l'on ait tiré (émis)* A_i. Les p_{ij} doivent bien entendu être des nombres réels entre 0 et 1 qui satisfont : $\sum_{j=1}^{r} p_{ij} = 1$, puisqu'on est certains de tirer quelque chose après avoir tiré A_i.

Dans le cas d'un processus ayant la mémoire des m tirages précédents, un état est en général une suite quelconque de m éléments de A mais la définition peut être plus grossière (ceci vaut également pour les cas qui précèdent) et un état peut être en correspondance avec un ensemble de telles suites. Les exemples donnés par Shannon éclairent tout ceci.

1.3.2 Mesure sur l'ensemble des suites associée à une chaîne de Markov

Nous ne nous intéresserons dorénavant qu'aux processus ayant seulement la mémoire du tirage précédent. Si l'alphabet possède r lettres, le processus est caractérisé par la $r \times r$ matrice M des p_{ij} (i est l'indice de ligne, j celui de colonne). Une telle matrice, à coefficients dans l'intervalle [01] et telle que la somme $\sum_{j=1}^{r} p_{ij}$ des termes de chaque ligne soit égale à 1 (i.e. telle que $(1, , , \cdots, 1)$ en soit un vecteur propre de valeur propre 1), est appelée une *matrice stochastique*
Nous ne considérerons explicitement que le cas particulièrement simple d'un alphabet à deux lettres, $A = \{0, 1\}$. La matrice M est alors une matrice 2×2 à coefficients dans $[0, 1]$ qui vérifient $p_{00} + p_{01} = 1$ et $p_{10} + p_{11} = 1$.

Cherchons à associer à un tel processus une mesure sur $\Omega = \{0,1\}^{\mathbb{N}^*}$ ou $\{0,1\}^{\mathbb{Z}}$ dont nous noterons $\omega = a_1 a_2 \dots a_n \dots$ ou $\dots a_1 a_2 \dots a_n \dots$ les éléments : il faut tout d'abord choisir un couple (p_0, p_1) de "probabilités initiales", celles d'avoir a_1 égal à 0 ou à 1 ; on peut alors définir la mesure des cylindres "élémentaires" correspondant à la fixation des n premiers éléments de la suite par

$$P_{M,p_0,p_1}(A_{12\dots n}^{j_1 j_2 \dots j_n}) = p_{j_1} p_{j_1 j_2} \dots p_{j_{n-1} j_n}.$$

La mesure d'un cylindre quelconque dans $\{0,1\}^{\mathbb{N}^*}$ se déduit du fait qu'il est une union disjointe de cylindres élémentaires obtenus en fixant à 0 et à 1 de toutes les manières possibles les termes entre 1 et i_n qui n'appartiennent pas à la liste.

Exercice 16 *1) Montrer par récurrence que les probabilités $p_0^{(n)}, p_1^{(n)}$ pour que a_n soit un 0 ou un 1 sont données par*

$$\left(p_0^{(n)} p_1^{(n)}\right) = (p_0\ p_1) M^{n-1}.$$

Interpréter de même les coefficients $p_{ij}^{(n)}$ de M^n comme les probabilités conditionnelles de tirer un j en n coups à partir d'un i.

On déduit de l'exercice précédent qu'une condition nécessaire pour que la mesure soit stationnaire (i.e. invariante par le décalage) est qu'il existe un *vecteur de probabilité*, c'est-à-dire une distribution de probabilité (p_0, p_1) sur $\{0,1\}$ telle que

$$(p_0\, p_1) M = (p_0\, p_1).$$

Il n'est pas difficile de montrer que, réciproquement, à la donnée d'un triplet (M, p_0, p_1), où M est une matrice à coefficients ≥ 0 admettant $(1\ 1)$ comme vecteur propre de valeur propre 1 et un vecteur de probabilité $(p_0\, p_1)$ comme covecteur propre (i.e. vecteur propre de sa transposée) de valeur propre 1, est associée une mesure de probabilité P_{M,p_0,p_1} sur $\{0,1\}^{\mathbb{N}^*}$ ou $\{0,1\}^{\mathbb{Z}}$ invariante par le décalage. Les mesures correspondant à une telle matrice sont appelées *mesures de Markov*.

Nous faisons maintenant sur M une hypothèse plus forte qui assurera que T possède la propriété de mélange pour cette mesure.

Lemme 18 (Perron-Frobenius) *Soit M ne matrice positive 2×2 à coefficients ≥ 0 et telle que le vecteur de coordonnées $(1,1)$ soit invariant. Supposons qu'il existe un entier s tel que la matrice M^s ait tous ses coefficients, notés $p_{ij}^{(s)}$, strictement positifs. Alors il existe une unique distribution de probabilité (p_0, p_1) sur $\{0,1\}$ telle que*
1) $(p_0\, p_1) M = (p_0\, p_1)$,
2) $\lim_{s\to\infty} p_{ij}^{(s)} = p_j,\ j = 0, 1.$

Démonstration. L'application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^2$ définie par ${}^tM$ envoie dans-lui-même le quadrant "positif" $\{(x,y) \in \mathbb{R}^2,\ x \geq 0, y \geq 0\}$ puisque les coefficients p_{ij} de M sont tous ≥ 0. Que $(1\,1)$ soit vecteur propre de M implique

de plus que ${}^tM$ envoie dans elle-même la droite d'équation $x + y = 1$, donc également le segment I obtenu en restreignant cette droite au quadrant positif. Elle a donc au moins un point fixe : en dimension 1, le théorème des valeurs intermédiaires suffit ; en dimension supérieure (alphabet à plus de deux lettres) on invoquerait le théorème de point fixe de Brouwer. Mais en fait la situation est beaucoup plus simple : l'hypothèse sur M^s implique en effet que l'application de I dans lui-même est une contraction, ce qui montre l'unicité du point fixe. La deuxième propriété s'obtient en écrivant que l'image par $({}^tM)^s$ de n'importe quel élément de I, en particulier ses extrémités $(0\,1)$ et $(1\,0)$, converge vers le point fixe $(p_0\,p_1)$. Il reste enfin à montrer que, non seulement M^s mais Melle-même possède ces propriétés, ce que je laisse en exercice au lecteur.

Remarque. L'interprétation de l'hypothèse est que quels que soient i et j dans $A = \{0, 1\}$, on a une probabilité non nulle de tirer un j en un nombre s de coups à partir d'un i ; La conclusion, outre la stationnarité, est que les cylindres $a_k = i$ et $a_{k+N} = j$ sont asymptotiquement indépendants lorsque $N \to \infty$ (la probabilité conditionnelle $p_{ij}^{(s)}$ dépend de moins en moins de i). Cette dernière remarque est à la base de la démonstration du

Théorème 19 *Sous les hypothèses du lemme précédent, la mesure de Markov P_{M,p_0,p_1} sur $\{0,1\}^{\mathbb{N}^*}$ (ou $\{0,1\}^{\mathbb{Z}}$) est invariante par le décalage T et possède la propriété de mélange. Elle est donc ergodique.*

Esquisse de démonstration. Je me contenterai de vérifier la propriété de mélange pour les cylindres élémentaires obtenus en fixant la valeur d'une seule décimale. La démonstration complète s'en déduit en considérant d'abord deux cylindres quelconques puis en concluant comme dans le cas des décalages de Bernoulli. Pour les cylindres élémentaires, on a

$$P_{M,p_0,p_1}(T^{-n}(A_{i_1}^{j_1}) \cap A_{i_2}^{j_2}) = P_{M,p_0,p_1}(A_{i_1+n}^{j_1} \cap A_{i_2}^{j_2}) = p_{j_2} p_{j_2 j_1}^{(n+i_1-i_2)},$$

(on a supposé $n + i_1 > i_2$) qui tend vers $p_{j_2} p_{j_1} = P_{M,p_0,p_1}(A_{i_1}^{j_1}) P_{M,p_0,p_1}(A_{i_2}^{j_2})$ lorsque n tend vers $+\infty$.

Je laisse au lecteur le soin de généraliser tout ceci aux chaînes de Markov associées à un alphabet fini quelconque.

1.4 Sources ergodiques plus générales

Il existe bien d'autres façons de définir sur $A^{\mathbb{N}^*}$ ou $A^{\mathbb{Z}}$ des mesures invariantes par le décalage. Voici un exemple, tiré de [B2], de source discrète ergodique non markovienne :

Soit A un alphabet fini, muni de la mesure de probabilité $(p_1, \ldots p_r)$. On munit $A^{\mathbb{Z}}$ de la mesure μ définie comme suit : pour i pair, le cylindre A_i* a la mesure p_j ; de plus ces cylindres sont indépendants : si i_1 et i_2 sont tous les deux pairs, $\mu(A_{i_1 i_2}^{j_1 j_2}) = p_{j_1} p_{j_2}$. Enfin si i est impair, a_i est égal à a_{i-1} ou a_{i+1} avec probabilité $1/2$ dans chaque cas. Autrement dit, $\mu(A_{i\,i+1}^{j_1 j_2}) = p_{j_2}/2$ et

$\mu(A_{i-1i}^{j_1 j_2}) = p_{j_1}/2$. Montrer que μ est invariante par T et qu'elle vérifie pour tout couple de cylindres C, D et donc pour tous boréliens C, D :

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n-1} \mu(C \cap T^{-k}(D)) = \mu(C)\mu(D).$$

En déduire que μ est ergodique. Montrer par ailleurs que le comportement de $\mu(C \cap T^{-k}(D))$ implique que μ n'est pas mélangeante.

Le lecteur trouvera par exemple une discussion des *sources gaussiennes* dans [M1].

2 L'entropie d'une source

2.1 L'entropie d'un espace de probabilité fini

2.1.1 De la définition de Hartley à celle de Shannon

Nous avons déjà rencontré la définition $h = \sum_{i=1}^{r} p_i \log \frac{1}{p_i}$ de l'entropie de Shannon associée à un alphabet fini $A = \{A_1, \cdots, A_r\}$ muni des probabilités $\{p_1, \cdots, p_r\}$, ou plus exactement à l'espace de probabilité $(A^{\mathbb{N}^*}, P_{p_1,\cdots,p_r})$ (ou $(A^{\mathbb{Z}}, P_{p_1,\cdots,p_r})$) des suites de tirages *indépendants* de lettres de cet alphabet.

Cette définition est un raffinement de celle, $\tilde{h} = \log r$ (c'est-à-dire $\tilde{h} = \frac{1}{n} \log N_n$ où N_n est le nombre total de messages de longueur n), donnée auparavant par Hartley. Les deux ne coïncident que lorsque les lettres de A sont équiprobables mais le premier théorème de Shannon, simple conséquence de la loi des grands nombres, montre qu'elle ne diffèrent en fait que par la prise en compte, dans la définition de Shannon, des seuls messages "signifiants", qui sont eux (approximativement) équiprobables. Les deux définitions sont donc essentiellement de même nature.

Mesure du nombre total de résultats "signifiants" (i.e. ayant une probabilité non infinitésimale) que peut fournir une expérience de n tirages au sort indépendants lorsque n est assez grand, h est une *mesure moyenne de notre incertitude* avant l'expérience ou, ce qui est équivalent, une *mesure moyenne de l'information que nous apporte une expérience.* Normalisée à 1 (si l'on prend le log en base r) dans le cas équiprobable où l'incertitude est totale, elle est d'autant plus proche de 0 que l'incertitude diminue et donc que l'expérience apporte peu. Par exemple, lorsqu'on lance une pièce dont les deux faces sont équiprobables, on ne peut rien dire *a priori* sur le résultat de l'expérience. Si au contraire les probabilités des deux faces sont distinctes, on s'attend à obtenir plus souvent celle qui a la probabilité la plus grande.

Plus précisément, montrons le

Lemme 20 *La fonction $h(p_1, \cdots, p_r)$, définie sur l'ensemble des lois de probabilité $P = (p_1, \cdots, p_r)$ sur A, est strictement concave et possède un unique maximum en $(\frac{1}{r}, \cdots, \frac{1}{r})$.*

Si l'on oublie la contrainte $\sum p_i = 1$, h est concave car sa dérivée seconde est la matrice diagonale $\mathrm{diag}(-\frac{1}{p_1}, \cdots, -\frac{1}{p_n})$ qui est définie négative. Mais la restriction un sous-espace affine d'une fonction concave est encore concave.

L'affirmation sur le maximum résulte d'un calcul élémentaire d'extremum lié : il doit exister un multiplicateur de Lagrange λ tel que, si $s(p_1, \cdots, p_r) = \sum p_i$,

$$\frac{\partial h}{\partial p_i}(p_1, \cdots, p_r) = \lambda \frac{\partial s}{\partial p_i}(p_1, \cdots, p_r), \text{ i.e. } \log \frac{1}{p_i} - 1 = \lambda \text{ for } i = 1, \cdots, r.$$

Les p_i doivent donc être tous égaux. Dans la figure ci-dessous on donne l'allure des graphes de h lorsque $r = 2$ et $r = 3$:

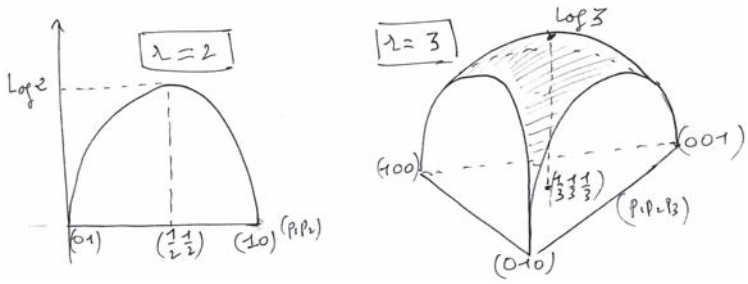


Figure 5 : Le graphe de l'entropie.

Donnons maintenant une autre démonstration, ne faisant pas appel au calcul différentiel mais seulement aux propriétés de convexité (concavité) résumées dans l'inégalité suivante, qui dit simplement que le *centre de gravité* d'un ensemble de masses appartient à l'*enveloppe convexe* de ces masses.

Proposition 21 (Inégalité de Jensen) *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction convexe. Quels que soient l'entier n, les points $x_1, \ldots, x_n \in \mathbb{R}$ et les poids $\lambda_1, \ldots, \lambda_n$ positifs et tels que $\sum_{k=1}^n \lambda_k = 1$, on a*

$$f\left(\sum_{k=1}^n \lambda_k x_k\right) \leq \sum_{k=1}^n \lambda_k f(x_k).$$

De plus, l'égalité n'est réalisée que si les x_k sont tous confondus.

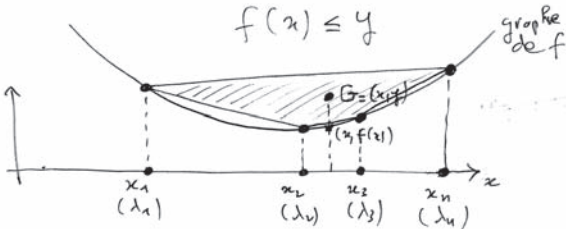


Figure 6 : L'inégalité de Jensen.

Afin de montrer que l'entropie est, au choix près de la normalisation, l'unique mesure raisonnable d'information, il nous faut en dégager une propriété fondamental qui nécessite la notion d'*entropie conditionnelle*.

2.1.2 Entropie conditionnelle

L'une des propriétés caractéristiques de l'entropie fait intervenir la notion d'entropie conditionnelle et il est bon de se souvenir de l'équivalence entre les notions d'espace de probabilité fini, de variable aléatoire valeurs finies et de partition finie. *Le langage des partitions, plus géométrique, est en effet est particulièrement bien adapté à la compréhension intuitive de la notion de probabilité conditionnelle.*
Etant donnés deux espaces de probabilité $A = \big(\{A_1, A_2, \dots, A_r\}, (p_1, p_2, \dots, p_r)\big)$ et $B = \big(\{B_1, B_2, \dots, B_s\}, (q_1, q_2, \dots, q_s)\big)$, les considérer comme partitions finies d'un même espace de probabilité $(\Omega, \mathcal{F}, \mu)$ revient à écrire leurs mesures sous la forme $p_k = \mu(A_k)$ et $q_l = \mu(B_l)$. On peut alors définir un espace de probabilité formé des *évènements conjoints* i.e. des couples d'un élément de A et d'un élément de B (souvent noté AB par les probabilistes),

$$A \vee B = \big(\{A_1 \cap B_1, \cdots, A_k \cap B_l, \cdots, A_r \cap B_s\}, (\pi_{11}, \cdots, \pi_{kl}, \cdots, \pi_{rs})\big),$$

où $\pi_{kl} = \mu(A_k \cap B_l)$, ainsi que des *lois de probabilité conditionnelle*

$$B|_{A_k} = \big(\{B_1, B_2, \dots, B_s\}, (q_{k1}, q_{k2}, \dots, q_{ks})\big),\ k = 1, 2, \dots, r,$$

et

$$A|_{B_l} = \big(\{A_1, A_2, \dots, A_s\}, (p_{1l}, q_{2l}, \dots, q_{rl})\big),\ l = 1, 2, \dots, s,$$

où les p_{kl} (probabilité de A_k si B_l est réalisé) et les q_{kl} (probabilité de B_l si A_k est réalisé) sont définis par

$$\pi_{kl} = p_k q_{kl} = p_{kl} q_l.$$

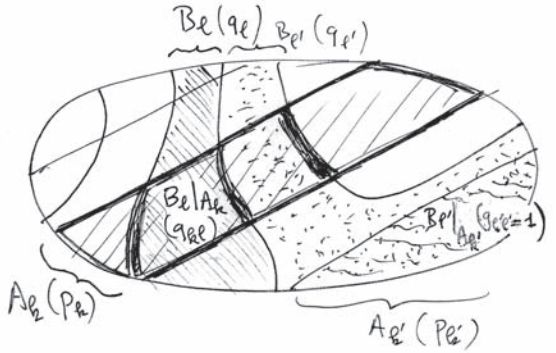


Figure 7 : Probabilités conditionnelles définies par deux partitions.

Définition 22 *Si $A \mapsto H(A)$ est une fonction définie sur l'ensemble des espaces de probabilité finis $A = \big(A, (p_1, \cdots, p_r)\big)$, on note $H_{A_k}(B) = H(B|_{A_k})$ et on définit $H_A(B)$ (encore notée $H(B|A)$) comme l'espérance de la variable aléatoire $A_k \mapsto H_{A_k}(B)$ définie sur A :*

$$H_A(B) = H(B|A) = \sum_{k=1}^{r} p_k H_{A_k}(B).$$

*Si H est la fonction entropie, on appelle $H_A(B)$ l'*entropie conditionnelle *(ou "entropie de B si A")*

Des égalités $\pi_{kl} = p_k q_{kl}$ et $\sum_l q_{kl} = 1$, on déduit immédiatement que

Lemme 22 *L'entropie $H(p_1, \cdots, p_r) = \sum p_i \log \frac{1}{p_i}$ vérifie l'identité*

$$H(A \vee B) = H(A) + H_A(B).$$

Remarque 1. Dans le langage des variables aléatoires à valeurs finies $x : \Omega \to A$, $y : \Omega \to B$, etc..., les mesures sur $A, B, \cdots$ sont définies comme les images directes de la mesure μ sur Ω par $x, y, \cdots$ et les probabilistes utilisent la notation plus imagée

$$p_k = \mu(x = A_k),\ q_l = \mu(y = B_l),\ \pi_{kl} = \mu(x = A_k, y = B_l).$$

Il est naturel, dans ce cas, de noter $H(x)$ au lieu de $H(A)$, $H(y)$ au lieu de $H(B)$, $H(x, y)$ au lieu de $H(A \vee B)$ et $H_x(y)$ (ou $H(y|x)$) au lieu de $H_A(B)$ (ou $H(B|A)$).

Remarque 2. Voici maintenant comment G. Raisbeck ([Ra]) justifie l'introduction de l'entropie comme mesure de l'information attachée à une expérience : prenons pour Ω un ensemble fini (dont les éléments peuvent être considérés comme des "messages") ayant r éléments et pour mesure μ celle définie par l'équiprobabilité des éléments : $\mu = (\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r})$. Soit $x : \Omega \to \{0, 1\}$ une variable aléatoire et $p_0 = \frac{r_0}{r}$, $p_1 = \frac{r_1}{r}$ la probabilité image directe de μ sur $\{0, 1\}$ (i.e. r_0 et r_1 sont les cardinaux des ensembles $x^{-1}(0)$ et $x^{-1}(1)$). On pensera à cette variable aléatoire comme représentant la nature (à choisir entre les deux possibilités nommées "0" et "1") d'un message émis suivant la probabilité μ. Cette émission se faisant à partir de r messages équiprobables, elle fournit, si l'on prend connaissance du message, $\log r$ bits d'information, seule fonction à la normalisation près du choix de la base du logarithme (on choisira la base 2), qui soit additive par juxtaposition d'ensembles indépendants de messages équiprobables. L'information associée à l'émission d'un message dont on sait seulement auquel des deux groupes il appartient est la différence entre l'information complète $\log r$ et l'information partielle $\log r_0$ ou $\log r_1$ associée à l'émission d'un message parmi les r_0 de $x^{-1}(0)$ ou les les r_1 de $x^{-1}(1)$. Les deux cas se produisant dans les proportions respectives p_0 et p_1, il est naturel d'estimer l'information "moyenne" fournie par l'expérience à

$$H(x) = \log r - p_0 \log r_0 - p_1 \log r_1 = -p_0 \log p_0 - p_1 \log p_1.$$

On reconnait la propriété ci-dessus de l'entropie : posons

$$B = \{(B_1, \dots, B_r), (\frac{1}{r}, \dots, \frac{1}{r})\},\ A = \{(0,1), (p_0, p_1)\},$$

et

$$A \vee B = \{(i,j), (\pi_{ij}),\ i \in A, j \in B\},$$

où $\pi_{ij} = 1$ si j appartient au paquet r_i et 0 sinon. Comme espaces de probabilité, $A \vee B$ et B sont isomorphes (exercice). L'entropie $H(A \vee B) = \log r$ correspondant à un tirage qui fournit la reconnaissance précise de l'élément est la somme de l'entropie $H(A) = p_0 \log \frac{1}{p_0} + p_1 \log \frac{1}{p_1}$ correspondant à un tirage indiquant seulement auquel des deux sous-ensembles il appartient et de l'entropie conditionnelle $H_A(B) = p_0 \log r_0 + p_1 \log r_1$.

Dans la section qui suit, nous montrons que les deux propriétés de l'entropie que nous avons mises en évidence suffisent essentiellement à la caractériser.

2.1.3 Caractérisation de l'entropie d'un espace de probabilité fini

La caractérisation donnée ici, à peine différente de celle que donne Shannon, est tirée de [Kh1] dont je conserve essentiellement les notations à ceci près que j'ai noté $H_{A_k}(B)$ ce qu'il note $H_k(B)$.

L'une comme l'autre considèrent une fonction définie à la fois sur tous les espaces de probabilité finis (la tribu, formée de tous les sous-ensembles, est sous-entendue).

Théorème 23 *Considérons, pour tout entier r, une fonction*

$$H(A) = H(p_1, p_2, \dots, p_r)$$

définie sur l'ensemble des espaces de probabilité finis $(A, p_1, \dots, p_r)$. On suppose que les fonctions H sont continues et vérifient :
1) Pour chaque r, $H(p_1, p_2, \dots, p_r)$ atteint son maximum en $(\frac{1}{r}, \frac{1}{r}, \dots \frac{1}{r})$.
2) Si A et B sont deux espaces de probabilité finis et si, comme ci-dessus, B est de plus muni de lois probabilité conditionnelle par rapport aux A_k, on a

$$H(A \vee B) = H(A) + H_A(B).$$

3) $H(p_1, p_2, \dots, p_r, 0) = H(p_1, p_2, \dots, p_r)$.
Alors il existe une constante positive λ telle que

$$H(p_1, p_2, \dots, p_r) = \lambda \sum_{k=1}^{r} p_k \log \frac{1}{p_k}.$$

Remarque. Chez Shannon, la première condition est remplacée par la condition que la fonction $H(\frac{1}{r},\dots,\frac{1}{r})$ soit monotone croissante en r, condition qui découle immédiatement de 1) et 3). D'autre part, il formule la condition 2) en termes de "choix successifs".

Esquisse de démonstration. *(i)* Posons

$$L(r) = H(\frac{1}{r},\frac{1}{r},\dots,\frac{1}{r}).$$

Si $A^{(1)}, A^{(2)}$ sont 2 copies de $A = \big(\{A_1, A_2, \dots, A_r\}, (\frac{1}{r},\frac{1}{r},\dots,\frac{1}{r})\big)$ indépendantes (au sens où la probabilité de l'événement conjoint $A^{(1)}_{k_1}A^{(2)}_{k_2}$ est $\frac{1}{r^2}$), la propriété 2) s'écrit $H(A^{(1)} \vee A^{(2)}) = H(A^{(1)}) + H(A^{(2)})$, c'est-à-dire $L(r^2) = 2L(r)$. On montre de même que $L(r^m) = mL(r)$ pour tout entier $m > 0$. C'est maintenant un exercice classique que de montrer qu'une fonction monotone $L(r)$ qui vérifie la condition ci-dessus est nécessairement de la forme $L(r) = \lambda \log r$.

(ii) La fonction H étant, pour tout r, supposée continue, il suffit de démontrer la formule dans le cas où $p_1, p_2, \dots, p_r$ sont tous rationnels, ce qui permet de les écrire $p_k = \frac{g_k}{g}$, où les g_k sont des entiers positifs de somme $\sum_{k=1}^r g_k = g$. Soit B l'espace de probabilité

$$B = \big(\{B_1^{(1)},\dots,B_1^{(g_1)},\dots,B_r^{(1)},\dots,B_r^{(g_r)}\}, (\frac{1}{g},\dots,\frac{1}{g},\dots,\frac{1}{g},\dots,\frac{1}{g})\big).$$

On définit les probabilités conditionnelles

$$P(B_l^{(i)}|A_k) = 0 \text{ si } l \neq k \text{ et } P(B_k^{(i)}|A_k) = \frac{1}{g_k}.$$

On a donc $H_{A_k}(B) = H(\frac{1}{g_k},\frac{1}{g_k},\dots,\frac{1}{g_k}) = \lambda \log g_k$ et

$$H_A(B) = \sum_{k=1}^r p_k H_{A_k}(B) = \lambda \sum_{k=1}^r p_k \log p_k + \lambda \log g.$$

Evaluons enfin $H(A \vee B)$: seuls les $g = \sum_{k=1}^r g_k$ événements $(A_k, B_k^{(i)})$ ont une probabilité non nulle (égale à $p_k \times \frac{1}{g_k} = \frac{1}{g}$). Donc $H(A \vee B) = L(g) = \lambda \log g$, ce qui conclut la démonstration.

Dorénavant, nous poserons

$$H(A) = H(p_1, p_2, \dots, p_r) = \sum_{k=1}^r p_k \log \frac{1}{p_k}, \quad H_A(B) = \sum_{k=1}^r p_k H_{A_k}(B).$$

2.2 L'inégalité de Shannon

2.2.1 L'approche classique basée sur les propriétés de convexité

Proposition 24 *$H_A(B) \leq H(B)$. En particulier, $H(A \vee B) \leq H(A) + H(B)$: l'entropie d'un choix simultané est inférieure ou égale à la somme des entropies des choix individuels. De plus, l'inégalité ne devient une égalité que lorsque A et B sont indépendants, c'est-à-dire lorsque pour tous k, l, on a $\pi_{kl} = p_k q_l$.*

Démonstration. On écrit

$$H_A(B) = \sum_k p_k \sum_l q_{kl} \log \frac{1}{q_{kl}},\ H(B) = \sum_k p_k \sum_l q_{kl} \log \frac{1}{q_l},$$

où dans l'expression de $H(B)$ on a utilisé l'égalité $q_l = \sum_k \pi_{kl} = \sum_k p_k q_{kl}$. L'inégalité énoncée dans la proposition suit du lemme suivant, qui affirme que pour chaque k, on a l'inégalité $\sum_l q_{kl} \log \frac{1}{q_{kl}} \leq \sum_l q_{kl} \log \frac{1}{q_l}$, l'égalité ne pouvant avoir lieu que si pour chaque l on a $q_{kl} = q_l$, c'est-à-dire $\pi_{kl} = p_k q_l$. La différence $H(B) - H_A(B)$ est donc une somme indicée par k de termes positifs ou nuls et elle ne peut s'annuler que si chacun des termes s'annule, c'est-à-dire si A et B sont indépendants.

Lemme 25 (Inégalité de Gibbs) *Si $P = (p_1, \ldots, p_r)$ et $Q = (q_1, \ldots, q_r)$ sont deux mesures de probabilité sur un même ensemble fini A, on a*

$$\sum_{k=1}^{r} p_k \log \frac{1}{q_k} \geq \sum_{k=1}^{r} p_k \log \frac{1}{p_k}.$$

De plus, l'égalité n'a lieu que si $p_k = q_k$ pour tout k.

Démonstration. On applique l'inégalité de Jensen à la fonction $x \mapsto \log \frac{1}{x}$ et aux points $x_k = \frac{q_k}{p_k}$ munis des poids p_k.

Remarque. *Seule la moyenne (l'espérance) $H_A(B)$ des $H_{A_k}(B)$ est $\leq H(B)$. Il est tout à fait possible que la réalisation d'un événement A_k particulier augmente l'incertitude sur le résultat du tirage de B* ; autrement dit, il est possible que, pour certaines valeurs de k, on ait $H_{A_k}(B) > H(B)$: ce sera par exemple le cas si la réalisation de A_k rend les B_l équiprobables, c'est-à-dire si, pour tout l, on a $q_{kl} = \frac{1}{s}$. La figure ci-dessous, tirée de [M2], en est un exemple (la mesure μ sur le carré Ω est la mesure de Lebesgue).

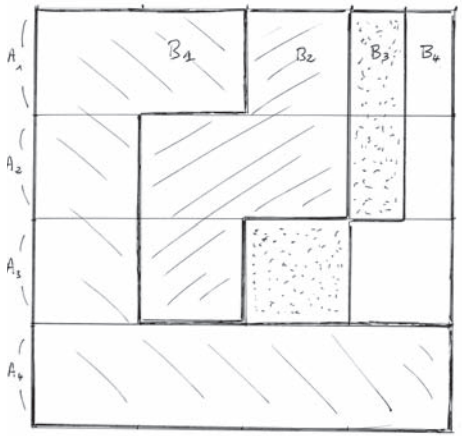


Figure 8 : Un exemple (les A_k sont des bandes horizontales d'égale épaisseur).

Exercice. Montrer l'inégalité de Shannon en maximisant $\sum_{kl} \pi_{kl} \log \frac{1}{\pi_{kl}}$ sous les contraintes $\sum_k \pi_{kl} = q_l$ et $\sum_l \pi_{kl=p_k}$.

2.2.2 Une inégalité de Shannon précisée

Lors d'une série de conférences en 2006, Misha Gromov a donné une version précisée de l'inégalité de Shannon, version dont je n'ai trouvé trace dans aucun des livres de théorie de l'information que j'ai consultés mais qui semble connue, au moins dans dans un cas particulier, des spécialistes de mécanique statistique (voir [Ru]).

Définition 23 *Une partition $C = \{C_1, \cdots, C_m\}$ d'un ensemble Ω est dite plus fine qu'une partition $D = \{D_1, \cdots, D_n\}$ du même ensemble si, pour tout i, il existe j tel que $C_i \subset D_j$. On dit aussi que D est plus grossière que C.*

Voici une définition de la partition $A \vee B$ formée des intersections $A_k \cap B_l$ qui appelle une définition symétrique :

Définition 24 *Etant données deux partitions A et B d'un même ensemble Ω, la partition $A \vee B$ est la plus grossière qui soit plus fine que A et que B ;*
la partition $A \wedge B$ est la plus fine qui soit plus grossière que A et que B.

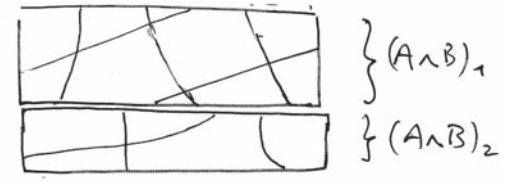


Figure 9 : Un exemple dans lequel $A \wedge B$ n'est pas trivial.

Proposition 26 (Inégalité de Shannon précisée) *Si A et B sont des partitions finies de l'espace de probabilité $(\Omega, \mathcal{F}, \mu)$, on a*

$$H(A \vee B) + H(A \wedge B) \leq H(A) + H(B).$$

L'égalité n'a lieu que si dans chacun des morceaux de $A \wedge B$, les partitions induites par A et B sont indépendantes.

Démonstration. Il suffit d'appliquer l'inégalité de Shannon a chaque élément de la partition $A \wedge B$. Précisément, Notons $A \wedge B = C = \{C_1, \cdots, C_m, \cdots, C_p\}$. Par hypothèse, chaque C_m est réunion de certains A_k et également réunion de certains B_l. Adoptons pour les éléments de A et B une numérotation à deux indices, le premier indiquant l'élément de C dans lequel l'élément considéré est contenu :

$$A = \{A_{11}, \cdots, A_{1\alpha_1}, A_{21}, \cdots, A_{2\alpha_2}, \cdots, A_{p1}, \cdots, A_{p\alpha_p}\},$$

$$B = \{B_{11}, \cdots, B_{1\beta_1}, B_{21}, \cdots, B_{2\beta_2}, \cdots, B_{p1}, \cdots, B_{p\beta_p}\}.$$

On note de même p_{mi} et q_{mj} mesures (i.e. les probabilités) de A_{mi} et B_{mj}. Notons enfin $\rho_m = \sum_{i=1}^{\alpha_m} p_{mi} = \sum_{j=1}^{\beta_m} q_{mj}$ la mesure de C_m et π_{mij} celle de $A_{mi} \cap B_{mj}$, tout en remarquant que si $m' \neq m$, l'intersection $A_{m'i} \cap B_{m''j}$ est vide. On a

$$H(A \vee B) = \sum_m \sum_{1 \leq i \leq \alpha_m} \sum_{1 \leq j \leq \beta_m} \pi_{mij} \log \frac{1}{\pi_{mij}}.$$

Mais pour m fixé, les A_{mi} d'une part, les B_{mj} d'autre part, définissent deux partitions C_m^A et C_m^B de C_m et l'on a

$$H(C_m^A \vee C_m^B) = \sum_{1 \leq i \leq \alpha_m} \sum_{1 \leq j \leq \beta_m} \frac{\pi_{mij}}{\rho_m} \log \frac{\rho_m}{\pi_{mij}}.$$

Puisque $\sum_{1 \leq i \leq \alpha_m} \sum_{1 \leq j \leq \beta_m} \pi_{mij} = \rho_m$, on en déduit que

$$H(A \vee B) = \sum_m \rho_m H(C_m^A \vee C_m^B) + H(A \wedge B).$$

De même,

$$H(A) = \sum_m \rho_m H(C_m^A) + H(A \wedge B),\ H(B) = \sum_m \rho_m H(C_m^B) + H(A \wedge B),$$

et donc

$$H(A)+H(B)-H(A\vee B)-H(A\wedge B) = \sum_m \rho_m \big(H(C_m^A)+H(C_m^B)-H(C_m^A\vee C_m^B)\big).$$

L'inégalité suit de l'application de l'inégalité de Shannon aux partitions C_m^A et C_m^B de C_m pour $1 \leq m \leq p$. Il en est de même de la conclusion concernant les cas d'égalité puisque le deuxième membre est une combinaison linéaire à coefficients positifs de termes positifs ou nuls.

2.2.3 Un cas particulier trivial et un bel exemple d'égalité.

Si les partitions A et B sont formées chacune de morceaux équiprobables, i.e. si $p_1 = \cdots = p_r = \frac{1}{r}$ et $q_1 = \cdots = q_s = \frac{1}{s}$, l'inégalité de Shannon, sous sa forme précisée, se déduit immédiatement de la majoration suivante du nombre $|A \vee B|$ d'éléments de $A \vee B$ en fonction de $r = |A|, s = |B|$ et $p = |A \wedge B|$.

Lemme 27 *On a la majoration*

$$|A \vee B| \leq \frac{rs}{p}.$$

En effet, de la première propriété caractéristique de l'entropie, on déduit que

$$H(A) = \log r,\ H(B) = \log s,\ H(A \wedge B) = \log p,\ H(A \vee B) \leq \log \frac{rs}{p},$$

qui n'est autre que l'inégalité cherchée.

Démonstration du lemme. Puisque la partition $A \vee B$ est celle définie par les intersections $A_k \cap B_l$ et qu'une telle intersection est vide dès que A_k et B_l n'appartiennent pas au même moreceau de la partition $A \wedge B$, une majoration de $|A \vee B|$ s'obtient en cherchant le maximum de $\sum_{m=1}^{p} \alpha_m \beta_m$, où les notations sont celles de la section précédente. C'est encore un problème d'extrema liés, les contraintes étant $\sum_{m=1}^{p} \alpha_m = r$ et $\sum_{m=1}^{p} \beta_m = s$. Il doit donc exister deux multiplicateurs de Lagrange λ et μ tels que, pour tout m, on ait $\beta_m = \lambda$ et $\alpha_m = \mu$, ce qui équivaut à $\alpha_m = \frac{r}{p}$ et $\beta_m = \frac{s}{p}$ et implique $\sum_{m=1}^{p} \alpha_m \beta_m = \frac{rs}{p}$, d'où la conclusion.

Un exemple d'égalité (Gromov). On prend pour espace ambiant l'espace vectoriel $(F_2)^N$ de dimension N sur le corps à deux éléments F_2 dans lequel on donne la même probabilité $\frac{1}{2^N}$ à chaque élément ; on appelle A et B deux partitions formées chacune de *sous-espaces affines parallèles*. L'inégalité de Shannon sous sa forme précisée, qui est dans ce cas une égalité, se réduit à une identité de codimensions de sous-espaces affines. En effet, si a et b sont respectivement la codimension des sous-espaces affines A_k et celle des B_l, ces sous-espaces ont respectivement 2^{N-a} et 2^{N-b} éléments et donc

$$r = 2^a,\ p_1 = \cdots = p_r = \frac{1}{2^a},\ s = 2^b,\ q_1 = \cdots = q_s = \frac{1}{2^b}\cdot$$

On en déduit que, si la base des log est 2, les entropies ne sont autres que les codimensions :

$$H(A) = a, \quad H(B) = b.$$

Il reste à remarquer que les partitions $A \vee B$ et $A \wedge B$ sont également des partitions en sous-espaces affines, respectivement ceux définis par les intersections des A_k et des B_l et ceux engendrés par les couples d'un A_k et d'un B_l ayant une intersection non vide. Je laisse au lecteur le plaisir de conclure.

L'une des vertus de cet exemple est bien entendu de montrer qu'il était conceptuellement déraisonnable d'oublier un des termes de l'(in)égalité, mais pour voir cela il suffisait de considérer le cas où $A = B$.

2.2.4 L'approche de Gromov basée sur la loi des grands nombres

La démonstration de l'inégalité de Shannon utilisant la loi des grands nombres n'est pas plus simple, au contraire, que la démonstration classique utilisant la convexité mais elle me semble très intéressante par cette idée de faire de la loi des grands nombres un outil permettant de ramener le cas général à un cas simple. De même que la définition de Shannon se confond avec celle de Hartley si l'on ne considère que les messages suffisamment longs "signifiants", Gromov a en effet remarqué que la loi des grands nombres permet de restreindre la démonstration de la proposition au cas, étudié ci-dessus, où les partitions A et B sont chacune formée de morceaux équiprobables. Techniquement, on se ramène "presque" au cas équiprobable en remplaçant Ω par Ω^N avec N assez grand et en "oubliant les éléments insignifiants". Plus précisément , si $(\Omega, \mathcal{F}, \mu)$

est un espace de probabilité, on définit sur Ω^N la mesure de probabilité produit tensoriel $\mu^{\otimes N}$ correspondant à l'indépendance des coordonnées et l'on raisonne comme suit :

(i) Aux partitions A et B de Ω correspondent des partitions A' et B' de Ω^N définies comme suit : un élément $A_{a_1...a_N}$ de A' est defini comme l'ensemble des $(x_1, ..., x_N)$ tels que $x_i \in A_{a_i}$ pour i=1,...,N. La mesure $p_{a_1\cdots a_N}$ de $A_{a_1...a_N}$ est le produit $p_{a_1} \cdots p_{a_N}$ des mesures des A_{a_i} ; on en déduit par un calcul explicite que $H(A') = NH(A)$.

(ii) On montre sans peine que $A' \vee B'=(A \vee B)'$ et $A' \wedge B' = (A \wedge B)'$. Il suffit donc de prouver l'inégalité pour les partitions A' et B'.

(iii) La forme faible du premier théorème de Shannon, obtenue par application de la loi faible des grands nombres aux variables aléatoires identiquement distribuées et indépendantes $\Omega^N \to R$ définies par $x = x_1...x_N \mapsto \log(1/p_{a_i})$ si $x \in A_{a_1...a_N}$, affirme l'existence, pour tout $\epsilon > 0$, d'un sous-ensemble Z_A de Ω^N de mesure arbitrairement proche de 1 (si N assez grand) qui est réunion d'éléments $A_{a_1...a_N}$ de A' tels que

$$|(1/N)\log(1/p_{a_1\cdots p_{a_N}}) - H(A)| < \epsilon.$$

Chacun de ces éléments a donc une probabilité comprise entre $2^{-N(H(A)+\epsilon)}$ et $2^{-N(H(A)-\epsilon)}$ (si les log sont en base 2), ce qui implique qu'il y en a au plus $2^{N(H(A)+\epsilon)}$. Un comptage analogue à celui qu'on fait dans le cas équiprobable (on compte les éléments de $A'' \vee B''$ en fonction du nombre d'éléments de $A'' \wedge B''$ et on extrémise) montre que si on considère les restrictions A'' et B'' de A' et B' à l'intersection de Z_B et Z_A on a

$$H(A'' \vee B'') + H(A'' \wedge B'') < N\left(H(A) + H(B) + 2\epsilon\right).$$

(iv) il reste à comparer l'entropie de $A'' \vee B''$, etc à celle de $A' \vee B'$, etc. Une estimation rapide (à vérifier) montre que dans l'estimation précédente, le remplacement de A'' par A' etc conduit à ajouter au 2ϵ une erreur de la forme. $cste(\epsilon + (1/N)\epsilon\log(1/\epsilon))$. Il reste à faire tendre ϵ vers 0.

2.2.5 Applications de l'inégalité de Shannon (Gromov)

Proposition 28 (Une inégalité universelle sur les parties finies de $\mathbb{Z}^n$) *Le cardinal $|Y|$ d'un sous-ensemble fini Y du réseau $\mathbb{Z}^n$ vérifie*

$$|Y|^{|Y|} \geq \Pi_S |Y \cap S|^{|Y \cap S|},$$

où le produit est pris sur l'ensemble de toutes les droites S parallèles à un axe de coordonnées et contenant un des points de Y, l'égalité ayant lieu si et seulement si Y est un produit de sous-ensembles de $\mathbb{Z}$.

Démonstration. Soit A^i la partition de Y dont les morceaux sont les intersections de Y avec les droites parallèle au $i$ème axe de coordonnées, et

$A^{12\cdots i}$ la partition dont les morceaux sont les sous-espaces affines parallèles au i-plan engendré par les i premiers axes de coordonnées. Munissant Y de la loi d'équiprobabilité de tous ses éléments, on a

$$H(A^1)+\cdots+H(A^n) = \sum_S \frac{|Y\cap S|}{|Y|}\log\frac{|Y|}{|Y\cap S|} = n\log|Y| - \sum_S \frac{|Y\cap S|}{|Y|}\log|Y\cap S|$$

puisque, chaque point de Y appartenant à n droites S, on a $\sum_S |Y\cap S| = n|Y|$. La proposition équivaut donc à l'inégalité

$$H(A^1)+\cdots+H(A^n) \geq (n-1)\log|Y|.$$

$$\text{Mais,}\quad H(A^1)+H(A^2) \geq H(A^1\vee A^2)+H(A^1\wedge A^2) \geq \log|Y| + H(A^{12}),$$

puisque la partition $A^1\vee A^2$ est la partition en atomes et que la partition $A^1\wedge A^2$ est a priori plus fine que A^{12}. On a donc

$$H(A^1)+H(A^2)+H(A^3) \geq \log|Y| + H(A^{12}\vee A^3) + H(A^{12}\wedge A^3),\quad \text{i.e.}$$

$$H(A^1)+H(A^2)+H(A^3) \geq 2\log|Y| + H(A^{123})$$

puisque $A^{12}\vee A^3$ est la partition en atomes et que $A^{12}\wedge A^3$ est a priori plus fine que A^{123}. Ajoutant un à un les termes $H(A^i)$ et remarquant que $A^{12\cdots n}$ est triviale et donc d'entropie nulle, on obtient l'inégalité cherchée. Je laisse au lecteur le soin de caractériser les cas d'égalité.

Remarque. La différence entre les deux termes de l'inégalité est une jolie mesure de la "dispersion" du sous-ensemble, i.e. de sa "distance" à la forme d'un "rectangle" k-dimensionnel.

L'inégalité ci-dessus implique très simplement la version discrète (et en fait plus forte) de l'*inégalité de Loomis-Whitney* :

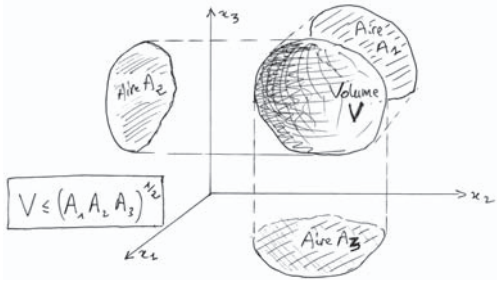


Figure 10 : L'inégalité isopérimétrique de Loomis-Whitney en dimension 3.

Proposition 29 $|Y| < \Pi|Y_i|^{1/(n-1)}$ *où les* Y_i *sont les projections sur les* n *hyperplans de coordonnées.*

En effet, cela équivaut à $(n-1))\log|Y| < \sum\log|Y_i|$ qui est impliqué par l'inégalité plus forte $(n-1)\log|Y| < \sigma H(m_i)$ où $H(m_i)$ est l'entropie de la mesure image de la mesure de Y (toujours muni de la loi d'équiprobabilité qui

donne la probabilité $1/|Y|$ à chaque point) par la projection sur le ième hyperplan de coordonnées (masse $(1/|Y|)|Y \cap S|$ sur le point qui est la projection de la fibre S). Cette inégalité s'écrit encore $-(n-1)\log|Y| > \sigma - H(m_i)$ ou $\log|Y| > \sigma(\log|Y| - H(m_i))$ (n termes). Mais cette dernière somme n'est autre que celle des coentropies des projections (ou entropies relatives de Y par rapport aux partitions définies par les fibres), c'est-à-dire ici la somme des l'espérances des fonctions qui à un point d'un des n quotients associe l'entropie de la fibre correspondante : i.e. la somme des $(|Y \cap S|/|Y|)\log|Y \cap S|$. L'inégalité se réduit donc à $\log|Y| > \sigma(|Y \cap S|/|Y|)\log|Y \cap S|$, qui par exponentiation donne l'inégalité de Loomis-Whitney discrète.

2.3 De l'entropie de Shannon à celle de Kolmogorov

Soit $A = (A_1, \cdots, A_r)$ un ensemble fini muni d'une loi de probabilité $(p_1, \cdots, p_r)$. Rappelons que, dans le cas de tirages indépendants, l'entropie $\sum_{i=1}^r p_i \log \frac{1}{p_i}$ peut être définie comme la limite $\lim_{n\to\infty} \frac{1}{n}\log\frac{1}{p(a_1\cdots a_n)}$ pour $P_{p_1,\cdots,p_r}$-presque toute suite $a_1\cdots a_n\cdots \in A^{\mathbb{N}^*}$. C'est une conséquence immédiate du théorème ergodique appliqué à la variable aléatoire

$$f : A^{\mathbb{N}^*} \to \mathbb{R}, \quad f(a_1a_2\cdots a_n\cdots) = \log\frac{1}{p_{a_i}}$$

(Rappelons les notations : $p_{a_i} = p_j$ si $a_i = A_j$ et, plus généralement, $p_{a_1\cdots a_n}$ est la mesure du cylindre $A_{1\cdots n}^{a_1\cdots a_n}$; il sera commode d'identifier A à l'ensemble $\{1, 2, \cdots, r\}$ et de noter également $A_{1\cdots n}^{a_1\cdots a_n} = A_{1\cdots n}^{j_1 j_2\cdots j_n}$ si $a_k = A_{j_k}$.)
Si l'on munit maintenant $A^{\mathbb{N}^*}$ d'une mesure de Markov ergodique $\mu = P_{M;p_1,\cdots,p_r}$ définie par des probabilités initiales p_i et des probabilités conditionnelles p_{ij}, on a pour μ-presque toute suite $a_1\cdots a_n\cdots \in A^{\mathbb{N}^*}$,

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{1}{\mu(a_1a_2\cdots a_n)} = \sum_{i,j=1}^r p_ip_{ij}\log\frac{1}{p_{ij}}.$$

On applique en effet le théorème ergodique à la fonction

$$g : \Omega \to \mathbb{R}, \quad g(a_1a_2\ldots a_n\ldots) = \log\frac{1}{p_{a_1a_2}},$$

ce qui donne

$$\frac{1}{n}\log\frac{1}{\mu(a_1a_2\ldots a_n)} = \frac{1}{n}\log\frac{1}{\mu(a_1)} + \frac{1}{n}\sum_{i=0}^{n-1} g(T^i(a_1a_2\ldots))$$

et

$$\int_\Omega g(x)d\mu(x) = \sum_{i,j=1}^r p_ip_{ij}\log\frac{1}{p_{ij}}.$$

La fonction g est en effet constante sur les atomes de la partition $\Omega = \sum_{i,j=1}^r A_{12}^{ij}$: elle vaut $\log\frac{1}{p_{ij}}$ sur A_{12}^{ij} et $\mu(A_{12}^{ij}) = p_ip_{ij}$.

2.3.1 L'entropie d'une chaîne de Markov

Ce qui précède nous amène à définir l'entropie d'une chaîne de Markov par la formule

$$H = \sum_{i,k=1}^{r} p_i p_{ik} \log \frac{1}{p_{ik}}.$$

Avec les notations utilisées pour définir l'entropie conditionnelle, H est l'espérance de la variable aléatoire $A_i \mapsto H_{A_i}(A) = \sum_{k=1}^{r} p_{ik} \log \frac{1}{p_{ik}}$. C'est celle d'un tirage suivant celui de la lettre A_i, la probabilité de tirer A_k étant la probabilité conditionnelle p_{ik}. Autrement dit, H n'est autre que l'entropie conditionnelle $H_A(A^2)$, qui se réduit bien à la définition déjà donnée dans le cas de tirages indépendants (encore appelé le cas "Bernouilli").

2.3.2 L'entropie comme quantité d'information moyenne par symbole

Soit μ une mesure de probabilité sur $A^{\mathbb{N}^*}$ (ou $A^{\mathbb{Z}}$) invariante par le décalage (par exemple $P_{p_1,\dots,p_r}$ dans le cas Bernouilli, $P_{M;p_1,\dots,p_r}$ dans le cas Markov). Soit $H_\mu^{<n>}$ l'entropie de l'espace de probabilité fini A^n muni de la mesure de probabilité définie par la mesure des cylindres, c'est-à-dire de la mesure de probabilité image directe par la projection canonique $\pi_n : A^{N^*} \to A^n$ (ou $A^{\mathbb{Z}} \to A^n$), $\pi(\cdots a_i \cdots) = a_1 a_2 \cdots a_n$ de la mesure μ :

$$H_\mu^{<n>} = \sum_{j_1 j_2 \dots j_n \in \{1,2,\dots,r\}^n} \mu(A_{12\dots n}^{j_1 j_2 \dots j_n}) \log \frac{1}{\mu(A_{12\dots n}^{j_1 j_2 \dots j_n})}.$$

L'entropie $H_\mu^{<n>}$ mesure l'information fournie par l'émission d'une suite de n symboles successifs (ou n expériences successives).

1) le cas Bernouilli : si $\mu = P_{p_1,\dots,p_r}$, un calcul direct montre que

$$H_\mu^{<n>} = nH.$$

2) le cas Markov : si $\mu = P_{M;p_1,\dots,p_r}$, posons

$$p_{i;j_1 j_2 \dots j_n} = p_{ij_1} p_{j_1 j_2} \cdots p_{j_{n-1} j_n}.$$

C'est la probabilité conditionnelle pour que, A_i étant réalisé, une suite de n tirages fournissent les résultats $A_{j_1}, A_{j_2}, \dots, A_{j_n}$. Comme dans le cas $n = 1$, on définit l'entropie de la chaîne itérée n fois par

$$H^{(n)} = \sum_{i=1}^{r} p_i H_i^{(n)} = \sum_{i,j_1,j_2,\dots,j_n=1}^{r} p_i p_{i;j_1 j_2 \dots j_n} \log \frac{1}{p_{i;j_1 j_2 \dots j_n}}.$$

Lemme 30 *On a les identités*

$$nH = H^{(n)} = H_\mu^{<n+1>} - \sum_{i=1}^{r} p_i \log p_i$$

Démonstration. $H^{(n)}$ est l'entropie associée à une suite de n tirages successifs. On raisonne par récurrence sur n. Supposons que le système soit dans l'état A_i et décomposons une suite de $n+1$ tirages en un premier tirage (événement A) suivi d'une suite de n tirages (événement B). Ces deux événements ne sont pas indépendants et la formule $H(A \vee B) = H(A) + H_A(B)$ s'écrit :

$$H_i^{(n+1)} = H_i + \sum_{k=1}^{r} p_{ik} H_k^{(n)}.$$

Mais $(p_1, p_2, \dots, p_r)$ étant un vecteur de probabilité associé à la matrice M des probabilités conditionnelles, on a $\sum_{i=1}^{r} p_i p_{ik} = p_k$ et donc

$$H^{(n+1)} = \sum_{i=1}^{r} p_i H_i^{(n+1)} = H + H^{(n)},$$

ce qui démontre la première identité. La deuxième résulte d'un calcul explicite.

2.3.3 L'entropie d'une source discrète

Rappelons qu'une source discrète est la donnée d'un alphabet fini $A = \{A_1, \dots A_r\}$ (par exemple $\{0, 1\}$) et d'une mesure de probabilité μ sur $\Omega = A^{\mathbb{N}^*}$ (ou $A^{\mathbb{Z}}$) invariante par le décalage T. *La probabilité d'obtenir un certain résultat lors d'un tirage dépend a priori de toute l'histoire passée, ce qui nous force à considérer des suites de tirages arbitrairement longues pour définir une entropie.* Les calculs de la section précédente montrent que la définition de l'entropie donnée dans le lemme suivant généralise bien celle donnée dans les cas où $\mu = P_{p_1,\dots,p_r}$ est "de Bernouilli et ceux où $\mu =$P$_{M;p_1,\dots,p_r}$ est "de Markov".

Lemme 31 (McMillan) *La "quantité moyenne d'information par symbole" $\frac{1}{n} H_\mu^{<n>}$ tend vers une limite $H_\mu = H_\mu(T)$ lorsque la longueur n des messages tend vers l'infini :*

$$H_\mu(T) = \lim_{n\to\infty} \frac{1}{n} H_\mu^{<n>} = \inf_n (\frac{1}{n} H^{<n>})$$

est par définition l'entropie de la source.

Démonstration. Comme dans la démonstration du lemme 19, on décompose l'émision d'une suite de $n+m$ symboles en deux événements dont le deuxième dépend du premier : l'émission X_n de la suite des n premiers symboles suivie de celle Y_m des m derniers. On a, en posant $u_n = H_\mu^{<n>}$,

$$u_{n+m} = H(X_n \vee Y_m) = H(X_n) + H_{X_n}(Y_m) \leq H(X_n) + H(Y_m) = u_n + u_m.$$

Cette "sous-additivité" de la suite $u_n = H_\mu^{<n>}$ est la clé de la preuve : Soit $v = \inf_n(\frac{u_n}{n})$. Par définition de v, quel que soit $\epsilon > 0$, il existe $N > 0$ tel que $u_N < N(v + \epsilon)$. Mais la division euclidienne permet d'écrire tout entier n sous

la forme $n = kN + r$ avec $k \geq 0$ et $1 \leq r \leq N-1$. Par la sous-additivité, ceci implique $u_n \leq u_{kN} + u_r \leq ku_N + \rho_N$, où $\rho_N = \sup_{1 \leq r \leq N-1}(u_r)$. Finalement,

$$\limsup_{n\to\infty}(\frac{u_n}{n}) \leq \limsup_{k\to\infty}(\frac{ku_N + \rho_N}{kN}) = \frac{u_N}{N} \leq v + \epsilon.$$

On en déduit que la suite $\frac{u_n}{n}$ converge vers v, ce qui démontre le lemme.

Dans le paragraphe suivant, j'indique brièvement la remarquable généralisation qu'a donnée Kolmogorov de l'entropie de Shannon – McMillan dans le cas où le décalage T est remplacé par n'importe quelle transformation préservant la mesure d'un espace de probabilité dans lui-même.

2.3.4 L'entropie de Kolmogorov

Définition 25 (Entropie d'une partition finie) *Soit $(\Omega, \mathcal{F}, \mu)$ un espace de probabilité et $\mathcal{E}$ une partition finie $\Omega = A_1 + A_2 + \ldots + A_r$ (à laquelle il n'est pas interdit de penser comme à une sous-algèbre finie de $\mathcal{F}$). L'entropie de $\mathcal{E}$ est*

$$H_\mu(\mathcal{E}) = \sum_{i=1}^{r} \mu(A_i) \log \frac{1}{\mu(A_i)}.$$

Notations. Etant donnée une transformation $T : \Omega \to \Omega$ et une partition $\mathcal{E}$, on note $T^{-1}\mathcal{E}$ l'algèbre formée des $T^{-1}(A_i)$, $A_i \in \mathcal{E}$. Etant données des partitions finies $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(m)}$ de Ω, on note $\vee_{i=1}^m \mathcal{E}^{(i)}$ la partition dont les atomes sont les intersections $A_{k_1}^{(1)} \cap A_{k_2}^{(2)} \cap \ldots \cap A_{k_m}^{(m)}$, où $A_{k_i}^{(i)}$ est un atome de $\mathcal{E}^{(i)}$.

Définition 26 *L'entropie $H_\mu(\mathcal{E}, T)$ d'une partition $\mathcal{E}$ par rapport à une transformation $T : \Omega \to \Omega$ préservant la mesure et l'entropie $H_\mu(T)$ de la transformation T sont respectivement définies par*

$$H_\mu(\mathcal{E}, T) = \limsup_{n\to\infty} \frac{1}{n} H_\mu\big(\vee_{k=0}^{n-1} T^{-k}\mathcal{E}\big), \quad H_\mu(T) = \sup_{\mathcal{E}} H_\mu(\mathcal{E}, T),$$

où le sup est pris sur toutes les partitions finies $\mathcal{E}$ de Ω.

Explication (voir [B2]) : un élément $A_{k_1}^{(1)} \cap A_{k_2}^{(2)} \cap \ldots \cap A_{k_m}^{(m)}$ de la partition $\vee_{i=1}^m \mathcal{E}^{(i)}$ peut être interprété comme la réalisation de m expériences, correspondant aux m partitions $\mathcal{E}^{(i)}$. Etant donnée une partition $\mathcal{E}$, notons $A = \{A_1, \ldots, A_r\}$ l'ensemble des atomes de la partition et $x : \Omega \to A$ la variable aléatoire qui, à un élément ω de Ω, fait correspondre l'atome A_i auquel il appartient. Puisque T préserve la mesure μ, les mesures images de μ par les variables aléatoires $x \circ T^n$ sont toutes les mêmes (n est un entier naturel, ou même un entier relatif si T est inversible). Autrement dit, les expériences correspondant aux partitions $T^{-n}(\mathcal{E})$ ont la même structure probabiliste et peuvent donc être toutes considérées comme des réalisations, a priori non indépendantes, d'une même expérience.

Le cas où $T : A^{\mathbb{N}^*} \to A^{\mathbb{N}^*}$, $A = \{A_1, A_2, \dots, A_r\}$, est un décalage et $\mathcal{E}$ la partition en cylindres élémentaires

$$\Omega = \{\omega = \dots a_1 a_2 \dots | \, a_1 = A_1\} + \{\omega | \, a_1 = A_2\} + \dots + \{\omega | \, a_1 = A_r\},$$

éclaire cette affirmation : les atomes de la partition $\vee_{k=0}^{n-1} T^{-k}\mathcal{E}$ sont les cylindres définis par la fixation de $a_1, a_2, \dots, a_n$; les expériences sont indépendantes s'il s'agit d'un décalage de Bernoulli, elles ne le sont pas pour un décalage de Markov.

Exercice 17 *Montrer que, dans le cas d'un décalage de Bernoulli T, pour tout n l'entropie $H_\mu\big(\vee_{k=0}^{n-1} T^{-k}\mathcal{E}, T\big)$ vaut n fois l'entropie $\sum_{k=1}^{r} p_k \log \frac{1}{p_k}$ de l'espace de probabilité fini $(A, p_1, \dots p_n)$ à partir duquel est construite la mesure invariante sur $A^{\mathbb{N}^*}$.*

Exercice 18 *Même exercice que le précédent mais en remplaçant décalage de Bernoulli et son entropie par décalage de Markov et son entropie, définie en 2.3.1.*

Nous admettrons le théorème suivant, dû à Kolmogorov (voir [B2, CFS]), qui implique imédiatement que l'entropie d'un décalage de Bernoulli est $\sum_{k=1}^{r} p_k \log \frac{1}{p_k}$ et celle d'un décalage de Markov $\sum_{i,k=1}^{r} p_i p_{ik} \log \frac{1}{p_{ik}}$:

Théorème 32 *Si T est inversible et s'il existe une partition finie $\mathcal{E}$ "génératrice" i.e. telle que la partition $\vee_{n=-\infty}^{n=+\infty} T^{-n}(\mathcal{E})$ ait pour atomes les éléments de Ω,*

$$H_\mu(T) = H_\mu(\mathcal{E}, T).$$

2.4 Le théorème de Shannon-McMillan-Breiman sur les sources discrètes ergodiques

2.4.1 L'entropie comme espérance

Considérons une source discrète quelconque, c'est-à-dire un alphabet fini A et une mesure de probabilité μ sur $\Omega = A^{\mathbb{N}^*}$ (ou $\Omega = A^{\mathbb{Z}}$), invariante par le décalage T. Nous avons déjà considéré l'espace de probabilité fini A^n muni de la probabilité définie par la mesure

$$p_{j_1 j_2 \dots j_n} = \mu(A_{j_1} A_{j_2} \dots A_{j_n}) := \mu(A_{12\dots n}^{j_1 j_2 \dots j_n})$$

des cylindres de longueur n. Son entropie, que nous avons noté $H_\mu^{<n>}$, est par définition, l'espérance de la variable aléatoire $\xi_n : A^n \to \mathbb{R}$ définie par

$$\xi_n(A_{j_1} A_{j_2} \dots A_{j_n}) = \log \frac{1}{\mu(A_{j_1} A_{j_2} \dots A_{j_n})}.$$

On a donc, par définition de l'entropie d'une source générale :

$$H_\mu(T) = \lim_{n\to\infty} E\left(\frac{1}{n}\xi_n\right).$$

2.4.2 La réalisation de l'espérance ou l'équipartition asymptotique

Nous avons déjà rencontré le théorème de Shannon dans les cas Bernouilli et Markov. Son énoncé, qui remplace l'espérance par la variable elle-même, a été généralisé à des sources ergodiques quelconques par McMillan et Breiman. Je renvoie à [B2] pour la démonstration (assez technique).

Théorème 33 (Shannon-McMillan-Breiman) *L'entropie d'une source discrète ergodique (T, μ) vérifie : pour μ-presque tout $\omega = \ldots a_1 a_2 \ldots a_k \ldots \in \Omega$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu(a_1 a_2 \ldots a_n)} = H_\mu(T).$$

Autrement dit,

$$\mu \left\{ \omega = \ldots a_1 a_2 \ldots a_k \ldots \in \Omega, \quad \lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu(a_1 a_2 \ldots a_n)} = H_\mu(T) \right\} = 1.$$

Voici maintenant la forme faible de cet énoncé, plus proche des énoncés initialement donnés par de Shannon.

Corollaire 34 *Pour tout $\epsilon > 0$, on a*

$$\lim_{n\to\infty} \mu \left\{ \omega = \ldots a_1 a_2 \ldots a_k \ldots \in \Omega, \quad \left| \frac{1}{n} \log \frac{1}{\mu(a_1 a_2 \ldots a_n)} - H_\mu(T) \right| \geq \epsilon \right\} = 0.$$

Rappelons l'interprétation du corollaire (voir la figure 2) : par définition de la limite, il existe pour tout $\epsilon > 0$ un entier $n(\epsilon)$ ayant la propriété suivante : dès que $n \geq n(\epsilon)$, l'ensemble A^n peut être décomposé en deux morceaux : un "petit" (précisément d'un cardinal de l'ordre de 2^{nH}, où $H = H_\mu(T)$) sous-ensemble de suites de longueur n à peu près équiprobables, et le complémentaire dont la probabilité est $\leq \epsilon$. En particulier, *il est la plupart du temps possible de traiter les suites de n symboles (n grand) comme s'il n'y en avait que 2^{nH}, chacune ayant la probabilité 2^{-nH}.*

On déduit de cet énoncé une affirmation voisine, également donnée par Shannon : considérons le nombre minimum de messages de longueur n dont la réunion ait une probabilité $\geq 1-\delta$. On obtiendrait ce nombre en choisissant les messages de longueur n par ordre de probabilité décroissante et en s'arrêtant lorsque la borne $1 - \delta$ est atteinte.

Définition 27 *L'information essentielle (c'est le terme utilisé dans [M2]) $H_{\mu,\delta}^{<n>}$ d'une source discrète est définie par la formule suivante, dans laquelle $|E|$ désigne le cardinal de E :*

$$H_{\mu,\delta}^{<n>} = \log \min\{|E|; E \subset A^n, \mu(E) \geq 1 - \delta\}.$$

Théorème 35 *Soit μ une source discrète ergodique. Quels que soient $\epsilon > 0$ et $0 < \delta < 1$, il existe un entier N tel que, quel que soit $n \geq N$,*

$$\left| \frac{1}{n} H_{\mu,\delta}^{<n>} - H_\mu(T) \right| \leq \epsilon.$$

Démonstration. 1) Le réel ϵ et l'entier n étant donnés, notons

$$E_1(n,\epsilon) = \left\{ (a_1, a_2, \dots, a_n) \in A^n, \ \left| \frac{1}{n} \log \frac{1}{\mu(a_1 a_2 \dots a_n)} - H_\mu(T) \right| \leq \epsilon \right\}.$$

On déduit du Corollaire 34 l'existence d'un entier $n_0 = n_0(\epsilon, \delta)$ tel que,

$$\forall n \geq n_0, \ \mu\big(E_1(n,\epsilon)\big) \geq 1 - \delta.$$

Les éléments $(a_1, a_2, \dots, a_n)$ de $E_1(n,\epsilon)$ vérifiant chacun

$$\mu(a_1 a_2 \dots a_n) \geq 2^{-n(H_\mu(T)+\epsilon)},$$

on en déduit que $|E_1(\epsilon,\delta)| \leq 2^{n(H_\mu(T)+\epsilon)}$ et donc que

$$\forall n \geq n_0(\epsilon,\delta), \ \frac{1}{n} H_{\mu,\delta}^{<n>} \leq H_\mu(T) + \epsilon.$$

2) Réciproquement, soit $E \subset A^n$ tel que $|E| \leq 2^{n(H_\mu(T)-\epsilon)}$. On a

$$\mu(E) = \mu\left(E \cap E_1(n, \frac{\epsilon}{2})\right) + \mu\left(E \cap E_1^c(n, \frac{\epsilon}{2})\right).$$

Choisissons $\delta' > 0$ tel que $\delta + \delta' < 1$. Le premier terme du second membre est majoré par $2^{n(H_\mu(T)-\epsilon)} \times 2^{-n(H_\mu(T)-\frac{\epsilon}{2})} = 2^{-n\frac{\epsilon}{2}}$ et donc par $\frac{\delta'}{2}$ dès que $n \geq n_1(\epsilon,\delta')$; le deuxième est majoré par $\mu(E_1^c(n,\frac{\epsilon}{2}))$ et donc par $\frac{\delta'}{2}$ dès que $n \geq n_0(\frac{\epsilon}{2}, \frac{\delta'}{2})$. On en déduit que, pour n assez grand, $\mu(E) \leq \delta' < 1 - \delta$. Il en résulte que

$$\forall n \geq \sup\big(n_1(\epsilon,\delta'), n_0(\frac{\epsilon}{2}, \frac{\delta'}{2})\big), \ \frac{1}{n} H_{\mu,\delta}^{<n>} \geq H_\mu(T) - \epsilon,$$

ce qui termine la démonstration.

L'interprétation de ce théorème est que, pour n assez grand, non seulement il est possible de ne prendre en compte qu'environ $2^{nH_\mu(T)}$ messages "typiques" de longueur n parmi les $2^{n \log |A|}$ messages possibles avec une probabilité arbitrairement petite de rencontrer un message en dehors de ceux-ci mais que ce nombre ne peut être essentiellement diminué et ce même au prix d'une augmentation de la marge d'erreur δ autorisée.

En attendant que la suite soit rédigée, reportez-vous pour l'intuition au livre de McKay et pour les mathématiques à celui de Cover et Thomas. Ce qui suit n'est qu'une ébauche

3 La transmission de l'information par des canaux discrets avec bruit

Le modèle le plus primitif est le suivant : une "entrée" ou "source" émet des messages. Ceux-ci sont transmis par un "canal" et recueillis à la "sortie". L'entrée et la sortie sont définis comme des espaces de probabilité, respectivement $(X = A^I, \mathcal{F}_X, P)$ et $(Y = B^J, \mathcal{F}_X, P)$ dont les éléments sont des messages (finis ou infinis suivant le choix des exposants I et J), c'est-à-dire des suites de mots dans des alphabets finis A et B rspectivement. Les erreurs que peut faire le canal sont représentées par la donné de probabilités conditionnelles $p_x(y)$ donnant la probabilité de recevoir le message y lorsque x a été émis.

3.1 La capacité d'un canal

Définition 28 *La capacité C d'un canal est définie par*

$$C = \max I(X,Y) = \max(H(X) - H_Y(X)) = \max(H(Y) - H_X(Y)),$$

où le max est pris sur toutes les sources possibles.

Exercices dans le cas où $X = A$ et $Y = B$ sont des alphabets finis.
1) Regarder dans McKay l'exemple de la machine à écrire approximative (noisy typewriter), où $A = B = A_1, \cdots, A_r$ et la lettre A_k a une probabilité $1/2$ d'être transmise comme A_k et une probabilité $1/2$ d'être transmise comme A_{k+1} (on pose $A_{r+1} = A_1$). Montrer que $C = \log r - 1$.

2) Calculer la capacité du canal binaire symétrique : $A = B = \{0,1\}$ et les probabilités conditionnelles sont

$$p_{00} = 1 - \alpha, p_{01} = \alpha, p_{10}, p_{11} = 1 - \alpha,$$

où $0 < \alpha < 1$ est donné.

3) Calculer la capacité du canal binaire à effacement : $A = \{0,1\}, B = \{0,e,1\}$ et les probabilités conditionnelles sont

$$p_{00} = 1 - \alpha, p_{0e} = \alpha, p_{01} = 0, p_{10} = 0, p_{1e} = \alpha, p_{11} = 1 - \alpha.$$

3.2 Le théorème de Shannon en l'absence de bruit

Etant donnée une source d'entropie H (bits par symbole) et un canal de capacité C (bits par seconde), il est possible de coder le message de façon à transmettre en moyenne $\frac{C}{H} - \epsilon$ symbols par seconde. Il n'est pas possible de transmettre un message à un taux moyen supérieur à $\frac{C}{H}$.

3.3 Le théorème de Shannon en présence de bruit

Idée fondamentale de coder les messages de façon à ce que les mots du code soient suffisamment éloignés les uns des autres : d'après van Lint "a misprint in a long(!) word is recognized because the word is changed into something that resembles the correct word more than it resembles any other word we know". Si un canal discret et une source discrète ont respectivement la capacité C et l'entropie par seconde H :
– si $H \leq C$, il existe un codage tel que la source sera transmise avec une fréquence d'erreurs arbitrairement petite.
– si $H > C$, il existe un codage tel que l'ambiguïté soit inférieure à $H - C + \epsilon$. Il n'existe pas de codage qui donne une ambiguïté inférieure à $H - C$. La preuve est une superbe idée de moyenne sur tous les codes possibles.

3.4 Exemples de canaux discrets

3.5 Exemples de codes correcteurs

4 Codage discret et bande finie

4.1 Fonctions périodiques et séries de Fourier

4.2 Le théorème d'échantillonnage de Shannon

(la possibilité de représenter les signaux dont la largeur de bande est finie par un échantillonnage fini)

4.3 Au-delà de Shannon

Les méthodes d'analyse en fréquence de Laskar

References

[AA] V. Arnold & A. Avez, Problèmes ergodiques de la mécanique classique, Gauthier-Villars 1967

[B1] P. Billingsley, Probability and measure, Wiley 1979

[B2] P. Billingsley, Ergodic theory and information,

[CFS] I.P. Cornfeld, S.V. Fomin & Ya. G. Sinai, Ergodic theory, Springer 1982

[CS] J.H. Conway & N.J.A. Sloane, Sphere Packings, Lattices and Groups, Springer 1993 (troisième édition 1999) Chapitre 3

[CT] T.C. Cover & J.A. Thomas, Elements of Information Theory, Wiley 1991

[F] W. Feller, An introduction to probability theory and its applications, vol. 1, Wiley

[G] M. Gromov, Entropy and Isoperimetry for Linear and non-Linear Group Actions, preprint mai 2007

[KH] A. Katok & B. Hasselblatt, Introduction to the Modern Theory of Dynamical Systems, Cambridge University Press 1995, section 4.1

[Kh1] A.I. Khinchin, Mathematical foundations of information theory, Dover 1957

[Ko] N. A. Kolmogorov, Foundations of the theory of probabilities, Chelsea 1960 (première édition, en allemand en 1933 sous le titre Grundbegriffe der Wahrscheinlichkeitrechnung)

[Ku] S. Kullback, Information theory and statistics, Wiley 1959, Dover 1968

[L] J.H. Van Lint, Introduction to Coding Theory, Springer 1982 (troisième édition 1999) Chapitres 1 et 2

[M1] R. Mañé, Ergodic theory and Differentiable Dynamics, Springer 1987

[M2] D. MacKay, Information theory, inference, learning algorithms, Cambridge University press 2004

[Pe] K. Petersen, Ergodic Theory, Cambridge University Press 1983

[Po] H. Poincaré, Les méthodes nouvelles de la mécanique céleste, tome III, chapitre XXVI "Stabilité à la Poisson", Gauthier-Villars 1899, retirage librairie scientifique et technique Albert Blanchard 1987

[R] W. Rudin, Real and complex analysis, McGraw-Hill 1966

[Ra] G. Raisbeck, Information theory: An Introduction for Scientists and Engineers, MIT ...

[Ru] D. Ruelle, Statistical mechanics, rigorous results, ?????

[Se] J. Segal, Le Zéro et le Un, Histoire de la notion scientifique d'information au $20^{ème}$ siècle, Syllepse 2003

[Sh] C. Shannon, A Mathematical Theory of Communication, Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, July, October 1948

[Si] Y Sinai, Probability theory, an introductory course (Moscou, 1985-1986, Springer 1992)

[St] H. Steinhaus, Les probabilités dénombrables et leur rapport à la théorie de la mesure, Fund. Math. 4, p. 286-310 (1923)

[Str] Strang, Introduction to applied mathematics, Wellesley - Cambridge press, 1986

[Sv] Sveshnikov, Problems in probability theory, mathematical statistics, theory of random functions, Dover, chapitre 5 sur l'entropie et la théorie de l'information

[W] N. Wiener, Cybernetics, Hermann 1948

[YY] A.M. Yaglom & I.M. Yaglom, Probabilité et information, $2^{ème}$ édition, Dunod 1969

Sur le web : L'article de Shannon ([Sh]) de 1948 :
http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html
Le court cours de David MacKay et le livre [M2] correspondant (attention, interdit d'imprimer le livre) :
http://www.inference.phy.cam.ac.uk/mackay/info-theory/course.html
Des liens intéressants (bios en particulier) sur
http://www.answers.com/topic/information-theory