Shannon's theorems: the strength of a simple idea

Alain Chenciner
Université Paris 7 & IMCCE
alain.chenciner@obspm.fr

Abstract

A corrupted word (or sentence), is correctly recognized as long as it differs less from the original word (or sentence) than from any other word (or sentence). Combined to the law of large numbers and its fundamental corollary, *the Asymptotic Equipartition Property*, this simple looking statement is at the root of the discovery by Claude Shannon of the limit $H < C$ imposed to any coding which allows a reliable transmission of information through a noisy channel, a limit which is almost achieved today by the turbocodes. In the course, after recalling the basic notions of probabilities, the *entropy* H of a source, and the *capacity* C of a channel will be defined, and Shannon's theorem will be proved in the simple case of *Bernoulli* sources

It is hoped that this elementary introduction will encourage the reader to enter the realm of *Ergodic Theory* and *Dynamical Systems*, in particular *Birkhoff's ergodic theorem*, which is a very strong version of the Law of Large Numbers, and *Kolmogorov's entropy* which plays a key role in classifications (it is not by chance that Kolmogorov was the first to recognize the importance of Shannon's work). A sketchy second part alludes to these topics.

Some quotations from the founders

The following quotations clearly define the mathematical setting. The first one is from Shannon [Sh] in 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.....

The second one is from Norbert Wiener [W], also in 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.

Contents

I – SHANNON'S THEOREM IN THE BERNOUILLI CASE

1 The weak Law of Large Numbers and Shannon's Asymptotic Equipartition Property

In this section, we prove in the simplest setting Shannon's Asymptotic Equipartition Property, a theorem on which rests the whole course.

1.1 Random variables with finite values

A random variable with finite values may be encountered under at least three equivalent disguises

– A *finite probability space*, that is a finite set $X = \{X_1, X_2, \ldots, X_N\}$ endowed with a probability law $\{P_1, P_2, \ldots, P_N\}$, where the P_i's (P_i=probability of X_i) are non negative real numbers whose sum is 1;

– A *finite measurable partition* $\Omega = C_1 + C_2 + \cdots + C_N$ of a space Ω endowed with a probability measure μ);

– A *random variable with finite values* $\xi : (\Omega, \mu) \to X = \{X_1, X_2, \ldots, X_N\}$.

The relations are straightforward: $C_i = \xi^{-1}(X_i),\ P_i = \mu(C_i)$, which means that the probability measure on X is the direct image $P = \xi_*\mu$ of μ.
The case of interest to us is $X = A^n$, where A is a finite alphabet, for example $A = \{a, b, c, \ldots, x, y, z\}$ or $A = \{0, 1\}$.

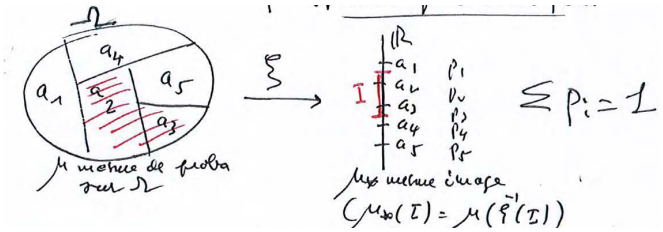

Figure 1 : Finite random variable.

Remark on the notations. In measure theory one writes $P(Y)$ for the probability of a subset $Y \subset X$, while in probability theory one writes $Pr\{\xi \in Y\}$.

Definition 1 *Two random variables $\xi, \eta : \Omega \to X$ with the same image law ($\mu_*\xi = \mu_*\eta$) are said to be* identically distributed.

Definition 2 *Two random variables $\xi, \eta : \Omega \to X$ are said to be independent if*

$$Pr(\xi \in Y,\, \eta \in Z) = Pr(\xi \in Y) Pr(\eta \in Z).$$

This definition can be immediately translated into a definition of the independence of two finite measurable partitions: $\Omega = C_1 + C_2 + \cdots + C_N$ and $\Omega = D_1 + D_2 + \cdots + D_K$ are independent if

$$\mu(C_i \cap D_j) = \mu(C_i)\mu(D_j).$$

Warning. The above definition extends immediately to the independence of a finite number of random variables. But, if the independence of a finite number of random variables implies their pairwise independence, figure 2 below shows that the converse is not true.

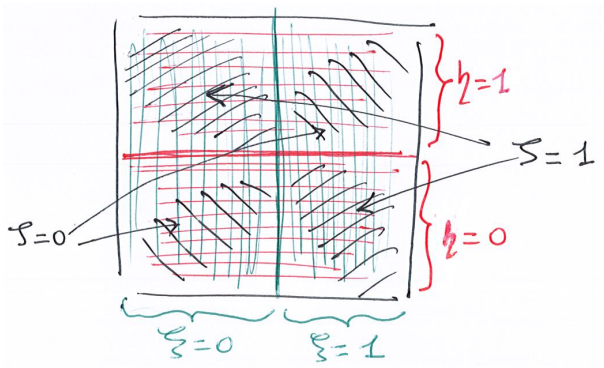

Figure 2 : Pairwise independent but globally dependent (μ is the area).

Example. The typical example of independent identically distributed (*iid*) random variables with finite values ξ_i is given by independent random draws of letters in a finite alphabet A endowed with a probability measure, for example $A = \{0, 1\}$, with 0 having probability p and 1 having probability $q = 1 - p$. Let $\Omega = A^n$, the set of words of length n, for some n, and define the measure $\mu = P_{p,q}$ by

$$P_{p,q}(a_1, a_2 \dots a_n) = p^\alpha q^{n-\alpha}\ ,$$

where α is the number of 0's in the word. The random variables $\xi_i : A^n \to \{0, 1\}$ are defined by

$$\xi_i(a_1 a_2 \dots a_n) = a_i.$$

As it will be important for our purpose to to be able to take the limit $n \to \infty$, we quickly introduce in the next section the space of infinite words with its probability laws corresponding to independence of the successive letters (the so-called *Bernouilli case*).

1.2 The game of "heads or tails" as a stochastic process

Let $X = \{0,1\}^{\mathbb{N}^*}$ be the set of *infinite* sequences

$$\omega = a_1 a_2 \dots$$

of 0's and 1's. As above, each such sequence can be thought of as an *infinite* sequence of "independent" coin tosses in a "heads or tails" game. It is the realization of a *stationary stochastic process without memory*: "stationary" means that the probability p that $a_i = 0$ and the probability $q = 1 - p$ that $a_i = 1$ are independent of the "time" i of the coin toss ; the independence (or absence of memory) means that the probability of a *cylinder*

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

is

$$\mu(A^{j_1 j_2 \dots j_k}_{i_1 i_2 \dots i_k}) = \mu(A^{j_1}_{i_1}) \mu(A^{j_2}_{i_2}) \dots \mu(A^{j_k}_{i_k}),$$

that is $p^\alpha q^{k-\alpha}$ if the sequence $j_1 j_2 \dots j_k$ contains α terms equal to 0.

Exercise 1 *1) A finite intersersection of cylinders is still a cylinder;*
2) the complement of a cylinder is a disjoint union of a finite number of cylinders;
3) a finite union of cylinders may also be written as a finite union union of disjoint cylinders;
4) deduce from 1),2),3) that the finite unions of disjoint cylinders form an algebra $\mathcal{G}$ of subsets of X (compare to the algebra of finite unions of disjoint intervals $[a_i, b_i[$ of $[0,1]$).

It is natural to define the tribe (= σ-algebra) $\mathcal{X}$ of measurable subsets as the one generated by the algebra $\mathcal{G}$ of finite unions of cylinders[1]. One says that the probability measure $\mu = P_{p,q}$ whose value on the cylinders was just given is the *product* of an infinity of copies of the measure (p, q) on $\{0,1\}$.

Apart from the countable unions of disjoint cylinders, producing non trivial elements of $\mathcal{X}$ is not so easy. In fact, the problem is the same as the one of producing a non trivial *Borelian* of the interval $[0,1] \subset \mathbb{R}$. The tribe $\mathcal{X}$ is indeed the *Borelian tribe* for the topology on X generated by the cylinders, that is the *infinite product* topology (see exercise 2): the probability of an element of $\mathcal{X}$ is defined as the unique extension of the probability we have defined for cylinders, in exactly the same way as the mesure of Borelians of $[0,1]$ is deduced from the measure (length) of intervals.

[1] Recall that, by definition, an algebra of subsets of a set X is a family closed under complements and finite unions while a σ-algebra is an algebra closed under countable unions. The σ-algebra $\mathcal{X} = \sigma(\mathcal{G})$ generated by the algebra $\mathcal{G}$ is the intersection of all the σ-algebras containing $\mathcal{G}$. While defining a probability space $(X, \mathcal{X}, \mu)$ one adds to the elements of the σ-algebra $\mathcal{X}$ all subsets of X of measure 0

Exercise 2 (the topological space $\{0,1\}^{\mathbb{N}^*}$ as a Cantor set). *One endows $\{0,1\}^{\mathbb{N}^*}$ with the product topology: a basis of open sets is formed by the cylinders. In other words, an open set is an arbitrary union of cylinders. Another definition is via the introduction of the distance $d(a_1a_2\ldots,b_1b_2\ldots)=\sum_{k=1}^{\infty}\frac{|a_k-b_k|}{2^k}$. Show that the map*

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

is a homeomorphism from $\{0,1\}^{\mathbb{N}^*}$ to the standard triadic Cantor set K. Show that K is of zero Lebesgue measure.

From $\{0,1\}^{\mathbb{N}^*}$ to the interval $[0,1]$: Let us now consider the map

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

As any element of $[0,1]$ possesses a *dyadic expansion* , this map is surjective. It is not injective: the inverse image of $\frac{1}{2}$ consists in $1000\ldots$ and $0111\ldots$, and same non-unicity phenomenon of the dyadic expansion occurs on the countable dense set of *dyadic numbers*, of the form $\frac{m}{2^k}$ where m and k are integers.

But, surprisingly, in the case of equiprobability ($p=q=1/2$) i.e. fair coin toss, one can show that f_2 is as good as a bijection from the measure point of view:

Proposition 1 (Steinhaus) *From the point of view of measure theory, the space of infinite sequences of bits (0's and 1') endowed with its borelian tribe and the measure corresponding to sequences of independent draws without bias, is equivalent to the interval $[0,1]$ endowed with its Borelian tribe and the Lebesgue measure: precisely, the map*

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

is an isomorphism of probability spaces[2] .

Proof. Thanks to a well-known property of tribes, its is enough to check measurability and preservation of the measure on generators of the Borelian tribe, that is on intervals, and even on intervals of the form $]\frac{p}{2^k},\frac{p+1}{2^k}]$. But, if $x=\sum_{i=1}^{k}\frac{a_i}{2^i}$ and $y=x+\frac{1}{2^k}$, one checks that $f_2^{-1}[x,y]=A_{12\ldots k}^{a_1a_2\ldots a_k}$ and hence that $P_{\frac{1}{2},\frac{1}{2}}\big(f_2^{-1}[x,y]\big)=\frac{1}{2^k}=|y-x|=\lambda([x,y])$.
On the other hand, the non injectivity of f_2 happens on a set of measure 0: let $\mathcal{D}$ be the subset of $\{0,1\}^{N^*}$ made by the sequences which after a certain rank contain only 1's; $\mathcal{D}$ is contained in a union of cylinders whose sum of

[2]Recall that a map $f:(X,\mathcal{A},\mu)\to(Y,\mathcal{B},\nu)$ from a probability space to another one is an isomorphism of probability spaces iff
1) it is a bijection modulo sets of measure 0
2) f and f^{-1} are measurable and preserve the measure.

probabilities may be chosen arbitrarily small (exercise) and hence it is of measure 0. Its complement $\{0,1\}^{N^*} \setminus \mathcal{D}$ is in bijection with the interval $[0,1[$ obtained by deleting a single point (hence of measure 0) to $[0,1]$. This ends the proof.

Corollary 2 *The map*

$$\delta = f_2 \circ f_3^{-1} : K \to [0,1]$$

is continuous and surjective.

This map δ takes the same values at the extremities of an interval of $[0,1] \setminus K$. Hence it can be extended into a map from $[0,1]$ to itself by giving it a constant value in each of the intervals of $[0,1] \setminus K$. This extension is a nice example of a function with *bounded variation* which is *not absolutely continuous* (i.e. not equal to the integral of its derivative, which exists and is equal to 0 Lebesgue-almost everywhere). Its graph, the *devil's staircase*, is illustrated on figure 3.

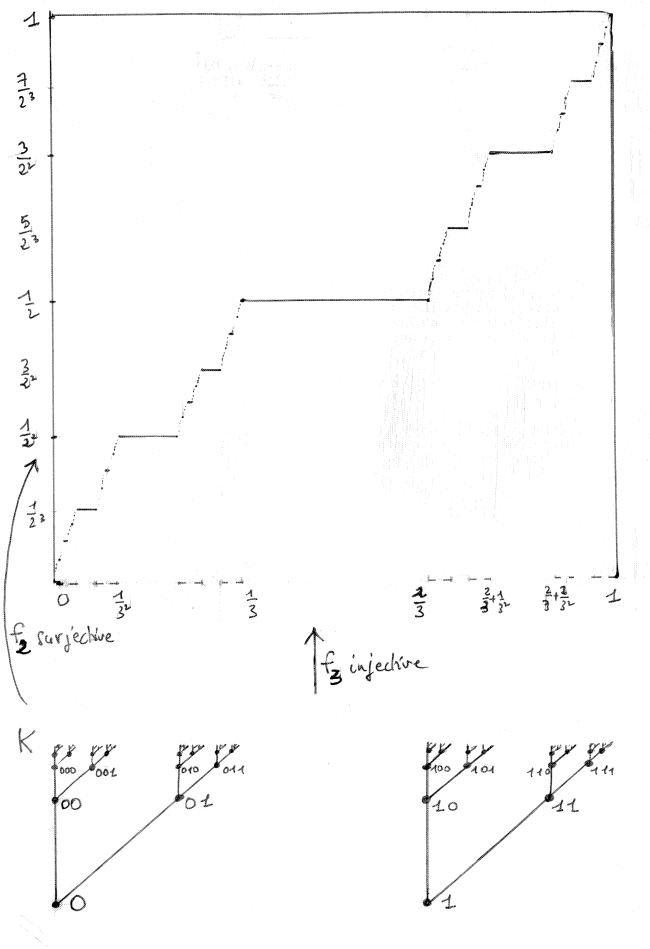

Figure 3 : the devil's staircase.

1.3 Expectation, variance

Being interested only in the very simple case of random variables with finite values, we need not speak of tribes, the tribe in this case being just the one generated by the corresponding partition.

Definition 3 *The* expectation *of the random variable* $\xi : (\Omega, \mu) \to \mathbb{R}$ *is its mean*

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

The deviation from the expectation is measured by the variance:

Definition 4 *The* variance *of a random variable is the expectation of the squared deviation from its mean*

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

If ξ takes the values $A_1, \cdots, A_r$, let $\Omega = C_1 + \cdots + C_r$ be the partition of Ω defined by the $C_i = \xi^{-1}(A_i)$. If $\mu(C_i) = p_i$, expectation and variance of ξ are given by the formulas

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

1.4 The weak Law of Large Numbers

The following elementary lemma implies the weak law of large numbers in the case of independent draws, also called the Bernouilli case:

Lemma 3 (Tschebishev 's inequality) *If* $\xi : (\Omega, \mu) \to \mathbb{R}$ *is a* positive *(i.e.*≥ 0*) random variable with finite values* $A_1, \ldots, A_r$*, and if* $\alpha > 0 \in \mathbb{R}$*, one has*

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

Proof.

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

Remark on the notations. In general, probabilists write the above inequality

$$Pr\{\xi \geq \alpha\} \leq \frac{E\xi}{\alpha}.$$

Applied to the positive random variable $(\xi - E\xi)^2$, Tschebishev's lemma becomes

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

Exercises.
1)Show that if $\xi_i, \xi_j : \Omega \to \mathbb{R}$ are *independent* random variables with finite values one has $E(\xi_i \xi_j) = E(\xi_i)E(\xi_j)$.
2) Show that the expectation of a sum of random variables $\xi_1, \cdots, \xi_n : \Omega \to \mathbb{R}$ with finite values is the sum of their expectations. Show that the same is true of their variances *if the ξ_i are two by two independent.*

We can now prove in the simplest case the weak law of large numbers, which founds the statistical interpretation of the notion of probability.
We consider random variables which are independent and identically distributed (*i.i.d.*); Here also the simplest example is given by the

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

Theorem 4 (Weak law of large numbers in the independent case) *If $\xi_1, \cdots, \xi_n : (\Omega, \mu) \to \mathbb{R}$ are* i.i.d. *random variables with finite values, whose expectation is m and variance is σ^2, one has:*

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

In particular, ϵ being fixed, this probability tends to 0 when n tends to $+\infty$.

Proof. Let $s_n = \xi_1 + \cdots + \xi_n$; one applies Tschebishev's inequality to the random variable $\left(\frac{s_n}{n} - E(\frac{s_n}{n})\right)^2$ and one uses the results of the above exercises.

1.5 The Asymptotic Equipartition Property (AEP) in the Bernoulli case

Corollary 5 (Asymptotic Equipartition Property in the Bernoulli case) *Under the above assumptions,*

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

where $\sigma^2 = \frac{1}{2} \sum_{i,j=1}^{r} p_i p_j (\log \frac{p_j}{p_i})^2$ and the probability is computed with the measure $P_{p_1,\cdots,p_r}$ on $\{A_1, \cdots, A_r\}^{\mathbb{N}^*}$ (or, this is equivalent, on $\{A_1, \cdots, A_r\}^n$). Hence, $\frac{1}{n} \log \frac{1}{p(a_1 \cdots a_n)}$ converges in probability to $h(p_1, \ldots, p_r) = \sum_{i=1}^{r} p_i \log \frac{1}{p_i}$ when n tends to infinity.

Proof. We apply the weak Law of Large Numbers theorem to the random variables

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

where the log can be taken in any basis and $p(a_i) = p_k$ if $a_i = A_k$. Writing $p(a_1 \cdots a_n) = p(a_1) \cdots p(a_n)$ for the probability of the cylinder $A_{1\cdots n}^{a_1 \cdots a_n}$ (which is also the probability of $a_1 \cdots a_n \in A^n$), one proves the theorem.

This corollary is of the utmost importance for Shannon's theory. It is best understood by introducing as in [CT] the ϵ-*typical set* $A_\epsilon^{(n)} \subset A^n$ defined by

$$A_\epsilon^{(n)} = \left\{ a_1 \dots a_n \in A^n,\ 2^{-n(h(p_1,\dots,p_r)+\epsilon)} \leq p(a_1 \dots a_n) \leq 2^{-n(h(p_1,\dots,p_r)-\epsilon)} \right\}.$$

One checks immediately that , if n is large enough,

$$Pr(A_\epsilon^{(n)}) \geq 1-\epsilon \text{ and } (1-\epsilon)2^{n(h(p_1,\dots,p_r)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(h(p_1,\dots,p_r)+\epsilon)}.$$

Figure 4, in which the size of elements represents their probability illustrates the interpretation of the corollary: as soon as n is large enough, one most probably will encounter only sequences (messages) $a_1 \cdots a_n$ whose probability is very close to 2^{-nh} (if the log is taken in basis 2) and the number of such messages is approximately 2^{nh}. Compared to the totality of the $2^{n \log r}$ possible messages of length $n.$, this maybe very small: If for example $h = 1/2 \log r$, that is one half of its maximal value (see next section), this represents 100 messages among 10000 !

Actually, It is easy to characterize most of these *typical* messages at the limit $n \to \infty$: according to the law of large numbers, they are the ones such that, for each $i = 1, 2, \dots, n$, the number n_i of occurences of the letter A_i satisfies $\lim_{n\to\infty} n_i/n = p_i$. Because of the possibility of large deviations from the mean, the messages with exactly $n_i = p_i n$ occurences of $A_i, i = 1, \dots, n$, are much less numerous but yet, thanks to *Stirling's formula* $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$, the logarithm of their number is asymptotically equivalent to nh when n tends to infinity:

$$\log \frac{n!}{(p_1 n)!(p_2 n!) \cdots (p_r n!)} \sim nh(p_1, p_2, \dots, p_n).$$

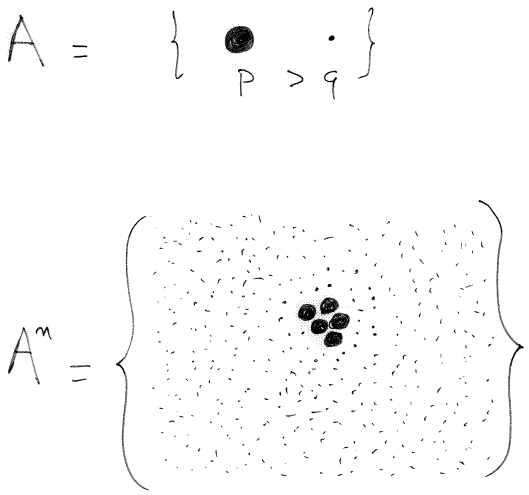

Figure 4 : The Asymptotic Equipartition Property.

2 The entropy of a finite probability space

2.1 From the definition of Hartley to the one of Shannon

Definition 5 (Shannon's entropy of a finite probability space) *The Shannon entropy of a finite set $A = \{A_1, \cdots, A_r\}$ endowed with probabilities $\{p_1, \cdots, p_r\}$, is the real number*

$$h(p_1, p_2, \cdots, p_r) = \sum_{i=1}^{r} p_i \log \frac{1}{p_i},$$

where the logs are most often taken in base 2.

This definition is a refinement of the one given by Hartley, $\tilde{h} = \log r$ (i.e. $\tilde{h} = \frac{1}{n} \log N_n$ where $N_n = r^n$ is the total number of messages of a given length n). Both definitions coincide only when all letters of A are equiprobables but the *Asymptotic Equipartition Property* (corollary 5) shows that their real difference is that in Shannon's definition only the *typical* messages, which are approximately equiprobable, are taken into account. Hence both definitions are essentially of the same nature.

Remarks.

1) entropy and Maxwell-Boltzmann's distribution. The above characterization of typical messages, based on Stirling's formula, is also at the root of the definition of Boltzmann's statistical distribution. Given n undistinguishable particles, each of which may be in r distinct energy states $E_1, E_2, \cdots, E_r$, a *macrostate* is defined by the r-tuple $(n_1, n_2, \cdots, n_r)$, where n_i is the number of particles having energy E_i. A given macrostate can be realized in $W = \frac{n!}{n_1! n_2! \cdots n_r!}$ ways by *microstates* in which the particles are labelled; hence, if all the microstates are supposed to be equiprobable, the probability of a given macrostate is proportional to W. Given a macrostate $(n_1, n_2, \cdots, n_r)$, let

$$E = \sum_{i=1}^{r} n_i E_i,\ p_i = \frac{n_i}{n},\ e = \frac{E}{n} = \sum_{i=1}^{r} p_i E_i$$

be respectively the total energy, the probability that a particle have energy E_i and the average energy per particle. The so-called *Maxwell-Boltzmann distribution* is obtained by maximizing the probability of a macrostate, that is of W with the constraint that the average energy per particle e be constant. Supposing that n is very big and replacing $\log W$ by its asymptotic equivalent $nh(p_1, p_2, \cdots, p_r)$, one is reduced to maximizing the entropy $h(p_1, p_2, \cdots, p_r)$ subject to the constraint $\sum_{i=1}^{r} p_i E_i = e$. Taking neperian logs, one gets the famous formula

$$p_i = \frac{e^{-\lambda E_i}}{\sum_{j=1}^{r} e^{-\lambda E_j}}, \quad \text{where } \lambda \text{ satisfies } \quad \frac{\sum_{j=1}^{r} E_j e^{-\lambda E_j}}{\sum_{j=1}^{r} e^{-\lambda E_j}} = e.$$

2) Entropy as a measure of information given by an experiment. As a mesure of the total number of *typical* (i.e. with a non infinitesimal probability) outputs which can result from n independent trials when n is big enough, h is an *average measure of our uncertainty* before the experiment; equivalently, it is an *average measure of the information which can be learnt from an experiment*. Maximal in the equiprobability case where uncertainty is complete, it approaches 0 when uncertainty vanishes, in which case an experiment brings no new information. For instance, when throwing a coin whose two faces are equiprobable, one cannot say anything *a priori* on the result of an experiment. If on the contrary the probabilities of the faces are different, one expects getting more often the one whose probability is the highest.

2.2 Elementary properties of the entropy function

Lemma 6 *The function $h(p_1, \cdots, p_r)$, defined on the set of all probabilitiy laws $P = (p_1, \cdots, p_r)$ on A, is strictly concave; it possesses a unique maximum at $(\frac{1}{r}, \cdots, \frac{1}{r})$.*

Forgetting at first the constraint $\sum p_i = 1$, h is concave because its second derivative is the diagonal matrix $\mathrm{diag}(-\frac{1}{p_1}, \cdots, -\frac{1}{p_n})$ which is negative definite. But the restriction to an affine subspace of a concave function remains concave.

The assertion concerning the maximum results from an elementary computation: there exists a Lagrange multiplier λ such that, if $s(p_1, \cdots, p_r) = \sum p_i$,

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

Hence all the p_i must be equal. Figure 5 shows the graph of h when $r = 2$ and $r = 3$:

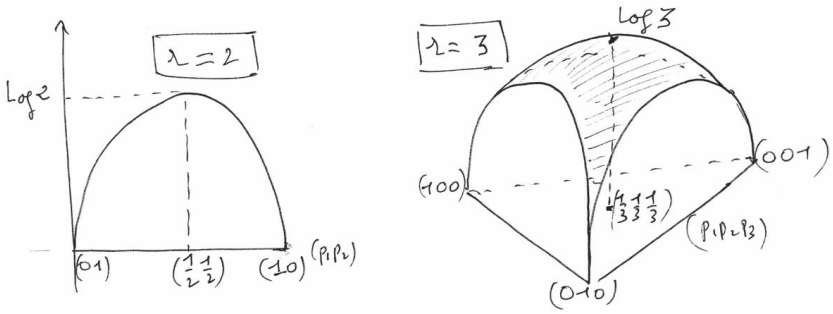

Figure 5 : Graph of entropy function

Let us give an alternative proof of the fact that the maximum occurs for equiprobability, not needing any differential calculus but using only the convexity (or concavity) properties. The following inequality expresses that the *center of mass* of a collection of point masses belongs to the *convex enveloppe* of these masses; applying it to the case $f(x) = \log(1/x), \lambda_k = p_k$, $x_k = 1/p_k r$, one gets that $\sum p_k \log(p_k r) \geq 0$, that is $h(p_1, \cdots, p_r) \leq \log r = h(1/r, \cdots, 1/r)$.

Proposition 7 (Jensen's inequality) *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. For any integer n, any points $x_1, \ldots, x_n \in \mathbb{R}$ and any positive weights $\lambda_1, \ldots, \lambda_n$ such that $\sum_{k=1}^n \lambda_k = 1$, one has*

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

Moreover, equality holds if and only if all the x_k are equal.

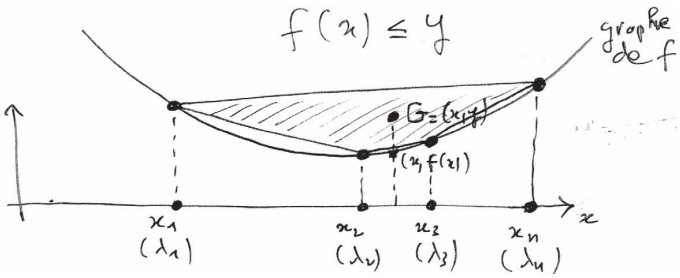

Figure 6 : Jensen's inequality

In order to show that, up to normalization, entropy is the unique reasonable measure of information, we must introduce a fundamental property for which we need the notion of *conditional entropy*.

2.3 Conditional entropy

Let us first remember the equivalence between the notions of finite probability space, finitely valued random variable and finite partition. *The language of partitions, more geometric, is indeed extremely well suited to intuitive understanding of the notion of conditional probability.*
Two probability spaces

$$A = \big(\{A_1, A_2, \ldots, A_r\}, (p_1, p_2, \ldots, p_r)\big) \text{ and } B = \big(\{B_1, B_2, \ldots, B_s\}, (q_1, q_2, \ldots, q_s)\big)$$

being given, let us consider them as finite partitions of one and the same probability space (Ω, μ); this amounts to writing their measures in the form $p_k = \mu(A_k)$ and $q_l = \mu(B_l)$. One can then define

1) a probability space consisting in *joint events* i.e. couples of an element of A and an element of B (often noted AB by the probabilists),

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

where $\pi_{kl} = \mu(A_k \cap B_l)$; the entropy $H(A \vee B)$ (also noted $H(A, B)$) is called the *joint entropy* of A and B.

2) conditional probability laws

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

and

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

where the p_{kl} (probability of A_k if B_l is realized) and the q_{kl} (probability of B_l if A_k is realized) are defined by $\pi_{kl} = p_k q_{kl} = p_{kl} q_l$.

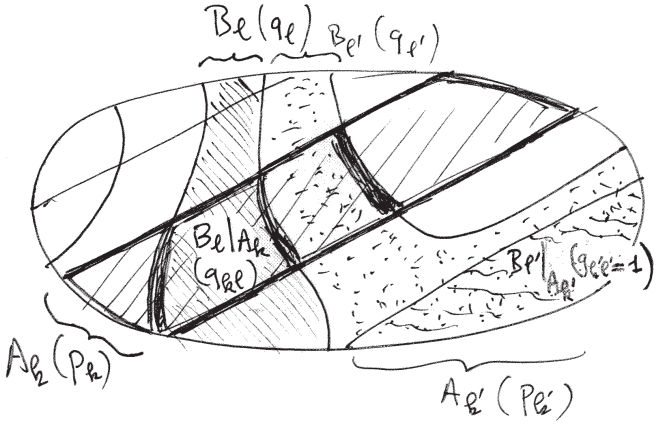

Figure 7 : Conditional probabilities defined by two partitions.

Definition 6 *If $A \mapsto H(A)$ is a function defined on the set of all finite probability spaces $A = \big(A, (p_1, \cdots, p_r)\big)$, one notes $H_{A_k}(B) = H(B|_{A_k})$ and one defines $H_A(B)$ (also noted $H(B|A)$) as the expectation of the random variable $A_k \mapsto H_{A_k}(B)$ defined on A :*

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

If H is the entropy, $H_A(B)$ is the conditional entropy (or entropy of B if A).

From the equalities $\pi_{kl} = p_k q_{kl}$ and $\sum_l q_{kl} = 1$, one deduces immediately the

Lemma 8 *The entropy $H(p_1, \cdots, p_r) = \sum p_i \log \frac{1}{p_i}$ satisfies the "chain rule"*

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

Remark 1. It follows that the equality $H(A \vee B) = H(A) + H(B)$, that is $H(A \vee B)$ maximum, is equivalent to the independence of A and B.

Remark 2. In the language of finitely valued random variables $x : \Omega \to A$, $y : \Omega \to B$, etc..., measure on $A, B, \cdots$ are defined as direct images by $x, y, \cdots$ of the measure μ on Ω and probabilists use the more figurative notation

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

It is then natural to note $H(x)$ instead of $H(A)$, $H(y)$ instead of $H(B)$, $H(x, y)$ instead of $H(A \vee B)$ and $H_x(y)$ (or $H(y|x)$) instead of $H_A(B)$ (or $H(B|A)$).

Remark 3. Here is how G. Raisbeck ([Ra]) justifies the introduction of entropy as a measure of the information attached to an experiment : let us take for Ω a finite set (whose elements may be considered as "messages") with r elements and for measure μ the one defined by equiprobability of elements : $\mu = (\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r})$. Let $x : \Omega \to \{0, 1\}$ be a random variable and $p_0 = \frac{r_0}{r}$, $p_1 = \frac{r_1}{r}$ be the direct image of μ on $\{0, 1\}$ (i.e. r_0 and r_1 are the cardinals of the sets $x^{-1}(0)$ and $x^{-1}(1)$). One shall think of this random variable as representing the nature (to be chosen among two possibilities named "0" and "1") of a message emited along the probability μ. As this emission is made from r equiprobable messages, reading the message provides $\log r$ information bits: this is indeed the only function, up to normalization amounting the choice basis for the logarithms (we shall choose the basis 2), which is additive under the juxtaposition of independent sets of equiprobable messages. The information associated to the emission of a message of which the only thing known is the group to which it belongs is the difference between the complete information $\log r$ and the partial information $\log r_0$ or $\log r_1$ associated to the emission of a message among the r_0 of $x^{-1}(0)$ or the r_1 of $x^{-1}(1)$. As the two cases occur in the respective proportions p_0 and p_1, it is natural to estimate the "average" information produced by the experiment as being

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

One recognizes the property of entropy mentioned above: let us set

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

and

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

where $\pi_{ij} = 1$ if j belongs to the subset r_i and 0 otherwise. As probability spaces, $A \vee B$ and B are isomorphic (exercise). The entropy $H(A \vee B) = \log r$ corresponding to a trial which gives a complete knowledge of the element is the sum of the entropy $H(A) = p_0 \log \frac{1}{p_0} + p_1 \log \frac{1}{p_1}$ corresponding to a trial indicating only to which of the two subsets it belongs and of the conditional entropy $H_A(B) = p_0 \log r_0 + p_1 \log r_1$.

In the following section, we show that the two properties of the entropy we have singled out suffice to characterize it.

2.4 Characterisation of the entropy of a finite probability space

The characterization given below, almost the same as the one given by Shannon, comes from [Kh1]. The notations are the same as the ones of [Kh1] except that I note $H_{A_k}(B)$ instead of $H_k(B)$.

Both authors consider a function defined on all finite probability spaces endowed with the tribe formed of all subsets.

Theorem 9 *Let, for any integer r,*

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

be a function defined on the set of all finite probability spaces $(A, p_1, \ldots, p_r)$. One supposes that the functions H are continuous and that they verify:
1) For each r, $H(p_1, p_2, \ldots, p_r)$ attains its maximum at $(\frac{1}{r}, \frac{1}{r}, \ldots \frac{1}{r})$.
2) If A and B sare two finite probability spaces and if, as above, B comes with conditional probabilities with respect to the A_k, the chain rule holds:

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

3) $H(p_1, p_2, \ldots, p_r, 0) = H(p_1, p_2, \ldots, p_r)$.
Then there exists a positive constant λ such that

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

Remark. Shannon replaces the first condition by the condition that the function $H(\frac{1}{r}, \ldots, \frac{1}{r})$ be monotone increasing in r, a condition which follows immediately from 1) and 3). On the other hand, he formulates condition 2) in terms of "successive choices".

Sketch of proof. *(i)* Let us set

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

If $A^{(1)}, A^{(2)}$ are two copies of $A = \big(\{A_1, A_2, \ldots, A_r\}, (\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r})\big)$ indépendent (i.e. such that the probability of the simultaneaous event $A^{(1)}_{k_1} A^{(2)}_{k_2}$ is $\frac{1}{r^2}$), property 2) reads $H(A^{(1)} \vee A^{(2)}) = H(A^{(1)}) + H(A^{(2)})$, that is $L(r^2) = 2L(r)$. One shows in the same way that $L(r^m) = mL(r)$ pour any integer $m > 0$. It is now a classical exercise to show that a monotonous function $L(r)$ which satisfies the above condition is necessarily of the form $L(r) = \lambda \log r$.
(ii) As for all r H is continuous, it suffices to show the formula in the case where $p_1, p_2, \ldots, p_r$ ar all rational numbers; one can then write $p_k = \frac{g_k}{g}$, where the g_k are positive integers whose sum is $\sum_{k=1}^{r} g_k = g$. Let B be the probability space

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

One defines the conditional probabilities

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

Hence $H_{A_k}(B) = H(\frac{1}{g_k}, \frac{1}{g_k}, \ldots, \frac{1}{g_k}) = \lambda \log g_k$ and

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

Finally, in the evaluation of $H(A \vee B)$: only the $g = \sum_{k=1}^{r} g_k$ events $(A_k, B_k^{(i)})$ have a non zero probability (equal to $p_k \times \frac{1}{g_k} = \frac{1}{g}$). Hence $H(A \vee B) = L(g) = \lambda \log g$, which concludes the proof.

From now on we shall set

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

Remarks.
1. Tverberg's characterization of entropy. In [Tv], continuity is not assumed but only integrability on the interval $[0, 1]$ of the function $H(x, 1-x)$; the function H is supposed to be symmetric in its arguments and to satisfy

$$H(x_1, x_2, \cdots, x_{n-1}, u, v) = H(x_1, x_2, \cdots, x_n) + x_n H\left(\frac{u}{x_n}, \frac{v}{x_n}\right),$$

which is condition 2) of theorem 9 in the case when B has 2 elements and the conditional probabilities $pr(B|A_k) = 0$ are equal to 0 if $k = 1, \cdots, n-1$ and (u, v) if $k = n$. Tverberg's proof is based on the following functional equation satisfied by the function $f(x) = H(x, 1-x)$:

$$f(x) + (1-x) f\left(\frac{u}{1-x}\right) = f(u) + (1-u) f\left(\frac{x}{1-u}\right)$$
for all $0 \leq x < 1$, $0 \leq u < 1$ such that $x + u \leq 1$.

2) the homological nature of entropy. In [BB], Baudot and Bennequin define a Hoschild type *information homology* for which the 1-cocycle condition is exactly the chain-rule identity; they use Tverbeg's functional equation above to prove that the entropy function generates the 1 dimensional cohomology group.

2.5 Shannon's inequality

2.5.1 Classical approach based on convexity properties

Proposition 10 *$H_A(B) \leq H(B)$. In particular, $H(A \vee B) \leq H(A) + H(B)$: the entropy of a simultaneous choice is less or equal to the sum of the entropies of the individual choices. Moreover, the inequality becomes an equality only when A and B are independent, that is when for all k, l, one has $\pi_{kl} = p_k q_l$.*

Proof. One writes

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

where in the expression of $H(B)$ one used the equality $q_l = \sum_k \pi_{kl} = \sum_k p_k q_{kl}$. The inequality in the proposition follows from the following lemma which affirms

that for each k, one has the inequality $\sum_l q_{kl} \log \frac{1}{q_{kl}} \leq \sum_l q_{kl} \log \frac{1}{q_l}$, equality occuring only if for each l one has $q_{kl} = q_l$, that is $\pi_{kl} = p_k q_l$. Hence the difference $H(B) - H_A(B)$ is a sum indexed by k positive (or zero) terms and it can be zero only if each term is zero, that is if A and B are independent.

Lemma 11 (Gibbs's inequality) *If $P = (p_1, \ldots, p_r)$ and $Q = (q_1, \ldots, q_r)$ are two probability measures on the same finite set A, one has*

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

Moreover, equality occurs only if $p_k = q_k$ for each k.

Proof. One applies Jensen's inequality to the function $x \mapsto \log \frac{1}{x}$ and the points $x_k = \frac{q_k}{p_k}$ endowed with weights p_k.

Remark. *It is only the average (the expectation) $H_A(B)$ of the $H_{A_k}(B)$ which is less than $H(B)$. It may well happen that the realization of some particular event A_k increases the incertainty on the result of the drawing of B* ; in other words, it is possible that, for some values of k, one has $H_{A_k}(B) > H(B)$: for instance, it will be the case if the realization of A_k makes the B_l equiprobable, that is if, for all l, one has $q_{kl} = \frac{1}{s}$. Figure 8 below, taken from [MK], is an example (the measure μ on the square Ω is Lebesgue).

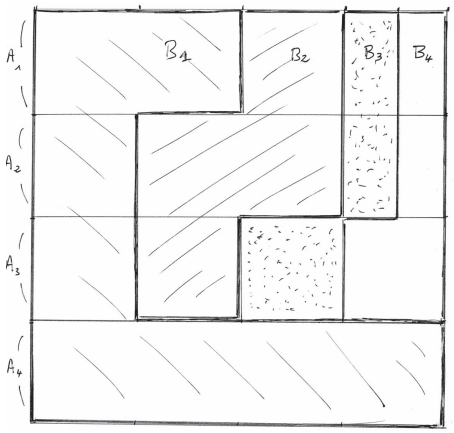

Figure 8 : An example (the A_k are horizontal bands of equal thickness).

Exercise. Prove Shannon's inequality via maximisation $\sum_{kl} \pi_{kl} \log \frac{1}{\pi_{kl}}$ under the constraints $\sum_k \pi_{kl} = q_l$ and $\sum_l \pi_{kl} = p_k$.

2.5.2 A more precise Shannon's inequality

This version of Shannon's inequality was given by Misha Gromov during a series of lectures in 2006. I have found mention of it in none of the books of Information Theory that I consulted but it seems to be known among specialists of Statistical Mechanics. (see [Ru]).

Definition 7 *A partition $C = \{C_1, \cdots, C_m\}$ of a set Ω is said to be finer than a partition $D = \{D_1, \cdots, D_n\}$ of the same set if, for all i, there exists j such that $C_i \subset D_j$. On says also thart D is coarser than C.*

Here is a definition of the partition $A \vee B$ consisting in the intersections $A_k \cap B_l$ which calls for a symmetric definition of $A \wedge B$:

Definition 8 *Given two partitions A and B of a set Ω, the partition $A \vee B$ is the coarser partition which is finer than A and B ; the partition $A \wedge B$ is the finer partition which is coarser than A and B.*

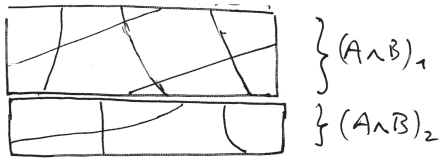

Figure 9 : An example such that $A \wedge B$ is non trivial.

Proposition 12 (Shannon's symmetric inequality) *If A and B are finite partitions of the probability space $(\Omega, \mathcal{F}, \mu)$, one has*

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

Equality occurs only if in each piece of $A \wedge B$, the partitions induced by A and B are independent.

Proof. It is enough to apply Shannon's inequality to each element of the partition $A \wedge B$. Precisely, let $A \wedge B = C = \{C_1, \cdots, C_m, \cdots, C_p\}$. By hypothesis, each C_m is the union of some A_k's and also the union of some B_l's. We may attach to the elements of A and B a double index, the first one indicating the element of C in which the element is contained :

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

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

One notes in the same way p_{mi} and q_{mj} the measures (i.e. the probabilities) of A_{mi} and B_{mj}. Finally, let $\rho_m = \sum_{i=1}^{\alpha_m} p_{mi} = \sum_{j=1}^{\beta_m} q_{mj}$ be the measure of C_m and π_{mij} the one of $A_{mi} \cap B_{mj}$. Note that if $m' \neq m$, the intersection $A_{m'i} \cap B_{m''j}$ is empty. One has

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

But for m fixed, the A_{mi}'s on the one hand, the B_{mj}'s on the other hand, define two partitions C_m^A and C_m^B of C_m and one has

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

As $\sum_{1\leq i\leq\alpha_m}\sum_{1\leq j\leq\beta_m}\pi_{mij}=\rho_m$, one deduces that

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

In the same way,

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

and hence

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

To conclude, one applies Shannon's inequality to the partitions C_m^A and C_m^B of C_m for $1\leq m\leq p$. The characterization of equality follows from the fact that the right hand side is a linear combination with positive coefficients of positive terms.

2.5.3 A trivial special case and a nice example of equality

If each one of the partitions A and B is made of equiprobable pieces, i.e. if $p_1=\cdots=p_r=\frac{1}{r}$ and $q_1=\cdots=q_s=\frac{1}{s}$, Shannon's inequality, in its precised version, is an immediate consequence of the following upper bound of the cardinal $|A\vee B|$ of $A\vee B$ in terms of $r=|A|, s=|B|$ and $p=|A\wedge B|$. Indeed, from the first characteristic property of the entropy fiunction, one deduces

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

Lemma 13 *The following inequality holds:*

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

Proof. The partition $A\vee B$ is the one defined by the intersections $A_k\cap B_l$; as such an intersection is empty as soon as A_k and B_l do not belong to the same piece of the partition $A\wedge B$, a majoration of $|A\vee B|$ is obtained by looking for the maximum of $\sum_{m=1}^p\alpha_m\beta_m$, where the notations are the one of the preceding section. This is still a problem of constrained extrema, with constraints $\sum_{m=1}^p\alpha_m=r$ and $\sum_{m=1}^p\beta_m=s$. Hence there must exist two Lagrange multipliers λ and μ such that, for all m, $\beta_m=\lambda$ and $\alpha_m=\mu$, that is $\alpha_m=\frac{r}{p}$ and $\beta_m=\frac{s}{p}$ which implies $\sum_{m=1}^p\alpha_m\beta_m=\frac{rs}{p}$.

An example of equality (Gromov). As the ambient space one takes the vector space $(F_2)^N$ of dimension N on the field F_2 with 2 elements, each element being given the same probability $\frac{1}{2^N}$. One calls A and B two partitions, each one consisting in *parallel affine subspaces*. The precised form of Shannon's inequality, in this case an equality, reduces to an identity of codimensions

of affine subspaces. Indeed, if a and b are the respective codimensions of the affine subspaces A_k and B_l, these subspaces have respectively 2^{N-a} and 2^{N-b} elements; hence

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

and the entropies are nothing but the codimensions:

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

Now, the partitions $A \vee B$ and $A \wedge B$ are also partitions into affine subspaces, respectively those defined by the intersections $A_k \cap B_l$ and those generated by the couples $A_k \cup B_l$ with non empty intersection. I leave to the reader the pleasure of the conclusion.

Of course, one of the virtues of such an example is to show that it was conceptually a mistake to forget one of the terms of the inequality, but in order to make this remark it was enough to consider the case where $A = B$.

2.5.4 The Asymptotic Equipartition Property as a tool for proofs

The proof of Shannon's inequality using the law of large numbers is not simpler than the classical proof based on convexity. Nevertheless it brings forth the beautiful idea of using the law of large numbers as a tool which reduces the general case to the case of equiprobability. In the same way as Shannon's definition reduces to Hartley's definition if one considers only sufficiently long "typical" messages, Gromov noticed that the Asymptotic Equipartition Property allows reducing the proof of proposition 12 to the case above where each of the partitions A and B consists in equiprobable pieces.

Technically, one "almost" reduces the general case to the equiprobable case by replacing Ω by Ω^N with N large enough and "forgetting the non typical elements". More precisely, if $(\Omega, \mathcal{F}, \mu)$ is a probability space, one defines on Ω^N the probability measure $\mu^{\otimes N}$ corresponding to the independence of coordinates and one argues as follows:

(i) To the partitions A and B of Ω correspond the following partitions A' and B' of Ω^N : an element $A_{a_1 \ldots a_N}$ of A' is defined as the set of all N-tuples $(x_1, ..., x_N)$ such that $x_i \in A_{a_i}$ for i=1,...,N. the measure $p_{a_1 \cdots a_N}$ of $A_{a_1 \ldots a_N}$ is the product $p_{a_1} \cdots p_{a_N}$ of the measures of the A_{a_i}'s ; an explicit computation shows that $H(A') = NH(A)$.

(ii) One shows easily that $A' \vee B' = (A \vee B)'$ andt $A' \wedge B' = (A \wedge B)'$. Hence, it is enough to prove the inequality for the partitions A' and B'.

The Asymptotic Equipartition Property asserts the existence of a subset $Z_A \subset \Omega^N$ whose measure is arbitrarily close to 1 if N is large enough which consists in elements $A_{a_1 \ldots a_N}$ of A' such that

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

It follows that the probability of each of these elements lies between $2^{-N(H(A)+\epsilon)}$ and $2^{-N(H(A)-\epsilon)}$, which implies that their number is at most $2^{N(H(A)+\epsilon)}$. The same count as was done in the equiprobable case (one counts elements of $A'' \vee B''$ as a function of the number of elements of $A'' \wedge B''$ and one extremises) shows that if one considers the restrictions A'' and B'' of A' and B' to the intersection of Z_B and Z_A one has

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

(iv) It remains to compare the entropy of $A'' \vee B''$, etc to the one of $A' \vee B'$, etc. A quick estimation (to be checked) shows that in the above estimation, replacing A'' by A' etc amounts to add to the 2ϵ an error of the form $cste(\epsilon + (1/N)\epsilon \log(1/\epsilon))$. Making ϵ tend to 0 gives the conclusion.

2.5.5 Applications of Shannon's inequality (Gromov)

Proposition 14 (A universal inequality for finite subsets of $\mathbb{Z}^n$) *The cardinal $|Y|$ of a finite subset Y of the lattice $\mathbb{Z}^n$ satisfies*

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

where the product is taken over all the lines S parallel to some coordinate axis which contain at least one point of Y. Equality occurs if and only if Y is a product of subsets of $\mathbb{Z}$.

Proof. Let A^i be the partition of Y whose pieces are the intersections of Y with the lines parallel to the ith axis of coordinates, and $A^{12\cdots i}$ the partition whose pieces are the affine subspaces parallel to the i-plane generated by the first i axes of coordinates. Endowing Y with the equiprobability of its elements, one has

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

Indeed, as each point of Y belongs to n lines in S, one has $\sum_S |Y \cap S| = n|Y|$ and the proposition is equivalent to the inequality

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

But, since the partition $A^1 \vee A^2$ is the partition into atoms and the partition $A^1 \wedge A^2$ is a priori finer than A^{12},

$$H(A^1) + H(A^2) \geq H(A^1 \vee A^2) + H(A^1 \wedge A^2) \geq \log |Y| + H(A^{12}).$$

It follows that

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

In the same way, as $A^{12} \vee A^3$ is the partition into atoms and $A^{12} \wedge A^3$ is a priori finer than A^{123}, one gets

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

Adding one by one the terms $H(A^i)$ and noticing that $A^{12\cdots n}$ is trivial and hence has entropy 0, one gets the required inequality. I leave to the reader the characterization of the cases of equality.

Remark. The difference between the two terms of the inequality is a nice measure of the "dispersion" of the subset, i.e. of its "distance" to the shape of a k-dimensional rectangle.

This inequality implies (and is stronger than) the discrete version of the *Loomis-Whitney inequality*:

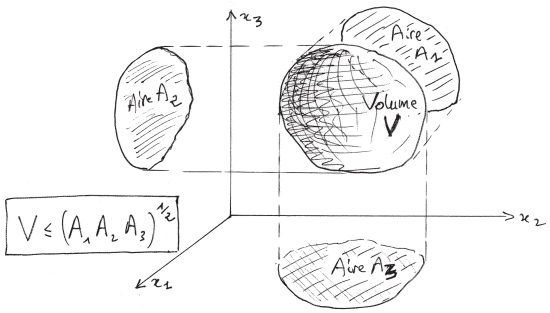

Figure 10 : Isoperimetric Loomis-Whitney inequality in dimension 3.

Proposition 15 *$|Y| < \Pi |Y_i|^{1/(n-1)}$ where the Y_i's are the projections of the n coordinate hyperplanes.*

Proof. We must show that $(n-1)) \log |Y| < \sum \log |Y_i|$, which is implied by the stronger inequality $(n-1) \log |Y| < \sum H(m_i)$ where $H(m_i)$ is the entropy of the image by the projection on the ith coordinate axis of measure on $|Y|$ which gives the same probability $1/|Y|$ to each point (i.e. the mass $(1/|Y|)|Y \cap S|$ is given to the point which is the projection of the fibre S). This inequality can be written $-(n-1) \log |Y| > \sum -H(m_i)$ or $\log |Y| > \sum (\log |Y| - H(m_i))$ (n terms). But this last sum is neither but the sum of the coentropies of the projections (or relative entropies of Y with repsect to the partitions defined by the fibres), that is the sum of the expectations of the functions which to a point in one of the n quotients associates the entropy of the corresponding fibre: i.e. the sum of the $(|Y \cap S|/|Y|) \log |Y \cap S|$. Finally, the inequality amounts to $\log |Y| > \sum (|Y \cap S|/|Y|) \log |Y \cap S|$ which, by exponentiation, gives the discrete Loomis-Whitney inequality.

3 Data compression

3.1 Using the typical set to define a code

Given a finite probability space $(A,p) = \big(\{A_1, A_2, \cdots, A_r\}, (p_1, p_2, \cdots, p_r)\big)$, let $A^n_\epsilon \subset A^n$ be the *typical subset*, that is the subset of words $a_1 \cdots a_n$ such that

$$H(p) - \epsilon \leq \frac{1}{n} \log \frac{1}{p(a_1 \cdots a_n)} \leq H(p) + \epsilon,$$

where, in the Bernouilli case, $p(a_1 \cdots a_n) = p(a_1) \cdots p(a_n)$. The Asymptotic Equipartition Property asserts that, if n is large enough,

$$Pr(A^n_{(\epsilon)}) \geq 1 - \epsilon.$$

We define in the following way a coding of the elements of A^n:

1) One orders the elements of A^n_ϵ and one associates to each of them their order number in base 2 as code word. Moreover, we prefix these code words by zero in order to recognize them. This takes at most $n(H+\epsilon)+2$ bits.
2) One codes in the same way the elements of $A^n \setminus A^n_{(\epsilon)}$ and we prefix their code words by 1. This takes at most $n \log r + 2$ bits.
Such a code is 1-1 and easy to decode and all typical sequences have short codes. The expectation of the length of a code word satisfies

$$EL = \sum_{\omega \in A^n} p(\omega) L(\omega) \leq Pr(A^n_{(\epsilon)}) \big(n(H+\epsilon)+2\big) + Pr(A^n \setminus A^n_{(\epsilon)})(n \log r + 2),$$

that is

$$EL \leq n(H + \epsilon').$$

We have proved the

Theorem 16 *Let a finite probability space (A,p) and a real number $\epsilon > 0$ be given. There exists a one to one binary coding of words of length n such that the expectation $E(L(X))$ of the length of a code word satisfies*

$$EL(X)) \leq n(H(p) + \epsilon).$$

In this computation, the lengths of the code words are treated as if they were taking only the two values $n(H+\epsilon)+2$ and $n \log r + 2$ (which would be the case if we were adding the ad hoc number of zeroes at the end of each code word). In fact, even taking accurately into account the length of each code word, one cannot do better for any one to one code (see [CT] Chapter 5 for a study of prefix codes, Kraft inequality, optimal code lengths, etc...): the Min expected code length in bits per symbol is the entropy; roughly said, if n is large enough, the source A^n can be considered as having in the average 2^{nH} equiprobable elements.

4 Data transmission

4.1 Mutual information and channel capacity

Figure 11 below summarizes Shannon's main theorems, which give a theoretical limit to the possibilities of transmitting messages through a noisy channel with an arbitrarily small error.

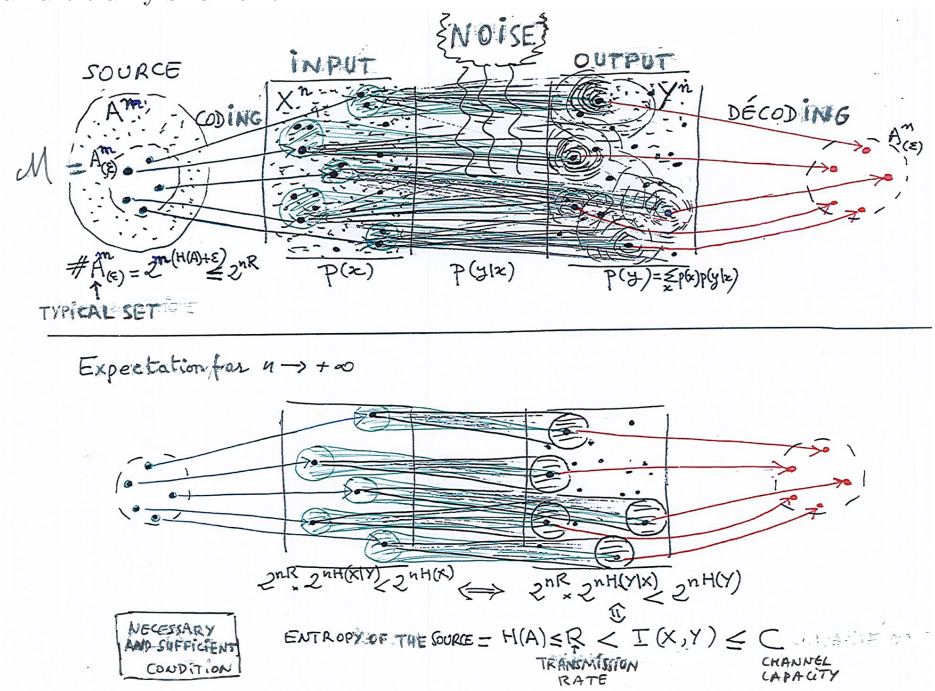

Figure 11 : A summary

In the simplest model, a "source" emits messages., these messages are coded (the "input"), then transmitted through a "channel" and then read as an "output". Supposing that the coding is also a random process, i.e. supposing that the letters in A are coded by randomly and independently chosen symbols in X, input and output messages are elements of probability spaces, say $(X^n, p(x))$ and $(Y^n, q(y))$. The possible errors due to the transmission by the channel are represented by conditional probabilities $q_x(y) = q(y|x)$ which give the probability to receive the message y when x is emitted. Then $q_x(y) = \frac{\pi(x,y)}{p(x)}$, where $\pi(x, y)$ is the probability of the joint event (x, y).

The main point is that for long enough messages-, the Asymptotic Equipartition Property allows one to forget about non typical messages, whose probability of being sent is asymptotically 0. This means that, be it for the source messages, the set of inputs (obtained by coding from the source messages) or the set of outputs (transmitted through the channel), one can replace the total number $2^{n \log |A|}$ (resp. $2^{n \log |X|}$ or $2^{n \log |Y|}$) of possible messages of length n by 2^{nh}, where h is the entropy of the corresponding alphabet A, X or Y.

The two types of fans represent respectively the probability of the ouputs originating from a given input or the probability of the inputs from which a given ouput may originate. When n tends to infinity (lower part of the figure), the probabilities that a given typical output comes from 2 different typical inputs (and similarly the probability that two typical outputs originate from the same typical input) vanishes, provided a certain "sphere packing property" holds; this leads to the the definitions of *mutual information* and *channel capacity* in terms of which Shannon's theorem is stated: for each typical input sequence of length n (large), there are in average approximately $2^{nH(Y|X)}$ most probable Y-output sequences, all of them having approximately the same probability. Indeed, the number of typical messages received is by definition 2^{nh} where h is the expectation $h = \sum_x p(x)H(Y|_x)$ of the random variable which to an input $x = x_1x_2\cdots x_n$ associates the entropy $H(Y|_x) = \sum_y p(y|x)log\frac{1}{p(y|x)}$ of the random variable $y = y_1y_2\cdots y_n$ for the conditional probability law corresponding to independant draws $p(y|x) = \prod_{i=1}^n p(y_i|x_i)$. Hence

$$h = \sum_{x,y} p(x)p(y|x)\log\frac{1}{p(y|x)} = H(Y|X).$$

But the total number of most probable Y-sequences of length n (large) is approximately $2^{nH(Y)}$. Hence the maximal number of disjoint most probable Y-output sequences is likely to be asymptotically the quotient $2^{nH(Y)}/2^{nH(Y|X)} = 2^{n(H(Y)-H(Y|X))}$.

Dually, for each typical Y-output sequence, there are approximately $2^{nH(X|Y)}$ typical X-input sequence, all of them having approximately the same probability. The total number of most probable input sequences of length n (large) being approximately $2^{nH(X)}$, the maximal number of disjoint most probable X-input sequences is likely to be approximately the quotient $2^{nH(X)}/2^{nH(X|Y)} = 2^{n(H(X)-H(X|Y))}$.

The two above estimates are in fact the same and they deserve a name:

Definition 9 (Mutual information I) *The following equivalent expressions define the mutual information $I(X,Y) = I(Y,X)$ of two random variables:*

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X \vee Y) \\ &= H(X \vee Y) - H(X|Y) - H(Y|X) \\ &= \sum_{x,y} \pi(x,y) log \frac{\pi(x,y)}{p(x)q(y)}. \end{aligned}$$

Remarks. 1) It follows from Shannon's inequality (see 2.5) that $I(X,Y) \geq 0$ (and even that $I(X,Y) \geq H(X \wedge Y)$).
2) Mutual information is a special case of the *relative entropy* or *Kullback – Leibler "distance"* (which in spite of its name is not a distance because non

symmetric) between two probability laws on X, defined as the expectation for the probability law p of the random variable $\log \frac{p(X)}{q(X)}$:

$$D(p|q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Indeed, $I(X,Y)$ is the relative entropy between the joint probability $\pi(x,y)$ and the product probability $p(x)q(y)$.

Finally, optimizing the choice of the probability law $p(x)$ on X, we get a fundamental characterization of a channel:

Definition 10 (Capacity C of a channel)

$$C = \max_p I(X,Y),$$

where the max is taken over all possible probabilities laws $p(x)$ on X .

Exemples in case $X = A$ and $Y = B$ are finite alphabets.

1) *Noisy Typewriter:* here $A = B = A_1, \cdots, A_r$ and the letter A_k has probability $1/2$ of being transmitted as A_k and probability $1/2$ of being transmitted as A_{k+1} (with the convention that $A_{r+1} = A_1$). One easily shows that $C = \log r - 1$.

2) *Binary Symmetric channel:* $A = B = \{0,1\}$ with conditional probabilities

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

where $0 < \alpha < 1$ is given. Show that the capacity is $1 - \alpha \log \frac{1}{\alpha} - (1-\alpha) \log \frac{1}{1-\alpha}$.

3) *Binary Erasure channel:* $A = \{0,1\}, B = \{0,e,1\}$ with conditional probabilities

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

Show that the capacity is $1 - \alpha$.

4) In november 1948, in a short note titled *A case of efficient coding for a very noisy channel*, Shannon considers the case when $p = \frac{1+\epsilon}{2}$, $q = \frac{1-\epsilon}{2}$ with ϵ small. The capacity of such a channel is approximately $K\epsilon^2$ where k is a constant. The theoretical code proposed by Shannon (only theoretical because the integer n must be very large) is the following: one repeats n times each symbol 0 or 1 and one decodes each sequence of n symbols at the majority. Each such sequence follows a *binomial law*, i.e. the probability of receiving k times 0 if 0 was sent is $\binom{n}{k} p^k (1-p)^{n-k}$ and analogously with p and q exchanged if 1 is sent. By the de Moivre-Laplace theorem, when n tends to infinity, one obtains

$$\lim_{n \to \infty} Pr \left\{ A \leq \frac{k - pn}{\sqrt{npq}} \leq B \right\} = \frac{1}{2\pi} \int_A^B e^{-\frac{z^2}{2}} dz.$$

The probabiity of error when sending 0 is the probability that the number k of 0's which are received satisfies $-\infty \leq k - pn \leq \frac{n}{2} - pn = -\frac{n\epsilon}{2}$, that is

$$-\infty \leq \frac{k-pn}{\sqrt{npq}} \leq \frac{-n\frac{\epsilon}{2}}{\sqrt{n\frac{1-\epsilon^2}{4}}} = -\sqrt{\epsilon}\sqrt{n}\sqrt{1-\epsilon^2} \sim -\epsilon\sqrt{n}.$$

Hence this probability is asymptotically given by the Gaussian integral

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-\epsilon n}{\sqrt{1-\epsilon^2}}}$$

Hence a bound on the error is obtained by bounding below $\epsilon\sqrt{n}$; , for example to get an error below 10^{-3} one must have $\epsilon\sqrt{n} \geq 3.1$ which is of the same order as the theoretical limit which will be given by Shannon's theorem.

4.2 The channel coding theorem

It is the remarkable fact discovered by Shannon that the rough hint given by the above analysis turns out to be a rigorous estimate of the possibilities of asymptotically error-free transmission by a noisy channel.

In order to state a precise theorem we still need a few definitions; note that having in mind a coding of the typical set, we deal only with fixed length codes.

Definition 11 ((M,n)-Code) *An (M,n) code for the channel $\{X^n, p(y|x), Y^n\}$ is an index set $\{1,2,\cdots,M\}$, an encoding function $\{1,2,\cdots,M\} \to X^n$, yelding codewords $X^n(1),\cdots,X^n(M)$ and a decoding function $g : Y^n \to \{1,2,\cdots,M\}$ (or $g(y)$="error").*

In the case we are considering, M will be approximately the number $2^{nH(A)}$ of typical messages from the source, all these typical messages being approximately equiprobable.

Definition 12 (Probability of error) *If f is the coding and g the decoding, we set*

$$\lambda_w^{(n)}(\mathcal{C}) = Pr(\big\{g(Y^n) \neq w \quad and \quad X^n = f(w)\big\},$$

$$\lambda^{(n)}(\mathcal{C}) = \max_{w\in\{1,2,\cdots,M\}} \lambda_w^{(n)}(\mathcal{C}), \quad \overline{\lambda^{(n)}(\mathcal{C})} = \frac{1}{M}\sum_{w=1}^{M} \lambda_w^{(n)}(\mathcal{C}).$$

Definition 13 (Rate of a code) *The rate of an $\{M,n\}$ code is $R = \frac{\log M}{n}$ bits per transmission. It is said to be achievable if there exists a sequence of $\{2^{nR},n\}$ codes such that the maximal probability error $\lambda^{(n)}$ tends to zero when n tends to infinity.*

We now prove Shannon's theorem in the simple setting of Bernouilli (i.e. discrete memoryless) sources and channel where the successive symbols sent are independent. We closely follow [CT].

Theorem 17 (Channel Coding Theorem) *All rates below capacity C are achievable. Conversely, any sequence $\mathcal{C}_i$ of $(2^{n_i R})$ codes with $\lambda(\mathcal{C}_i)$ tending to zero when n_i tends to infinity must have $R \leq C$.*

Recalling that there are approximately $2^{nH(A)}$ typical messages of length n which have a non asymptotically vanishing probability of being sent, we get

Corollary 18 (Source-channel coding theorem) *There exists a source channel code with λ tending to zero when n tends to infinity if $H(\mathcal{A}) < C$. Conversely, if $H(\mathcal{A}) > C$, the probability of error is bounded away from 0, and it is not possible to send the process over the channel with arbitrary low probability of error.*

Proof. We recapitulate the process of coding and sending the coded message. Only the last point (decoding) is new:

1) A $(2^{nR}, n)$ code is chosen randomly according to the probability law $p(x)$ on X. This means that the code

$$\mathcal{C} = \begin{pmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \cdots & \cdots & \cdots & \cdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{pmatrix}$$

is given the probability

$$Pr(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p\big(x_i(w)\big).$$

2) The sender and the recipient both know the code and the conditional probabilities $p(y|x)$ of the channel.

3) The message W is chosen according to equiprobability:

$$Pr(W = w) = 2^{-nR},\ w = 1, 2, \ldots, 2^{nR}.$$

4) If $x(w) = x_1(w)x_2(w)\cdots x_n(w)$ is sent, $y = y_1 y_2 \cdots y_n$ is receivend according to the probability

$$p(y|x(w)) = \prod_{i=1}^{n} p(y_i|x_i(w)).$$

5) A coded message $x(w)$ being sent which is received as y, we decode it as $g(y) = \hat{w}$ if

(i) the pair $(x(\hat{w}), y)$ is *joint typical* (i.e. if $x(\hat{w})$ is typical in X^n, y is typical in Y^n and the pair is typical in $X^n \vee Y^n$);

(ii) $\hat{w}$ is the only message with this property.

If $\hat{w} \neq w$ or if there is no such $\hat{w}$, one decodes as $g(y) = error$.

Computing the average error. (for details, see [CT] pages 198 ...)
This is the main point. The heuristic reasoning is the same as the one we did at the beginning, the only difference being the use for the mutual information of the formula $I(X,Y) = H(X) + H(Y) - H(X \vee Y)$: choosing a couple at random in $X^n \vee Y^n$ according to the probability defined by the probability on X and the conditional probabilities on Y characterizing the channel, the probability that it be typical goes to 1 when n tends to infinity. On the other hand, taking independently a random typical message in X^n and a random typical message y in Y^n, the probability that the couple (x,y) be joint typical is approximately $2^{nH(X\vee Y)} / \left(2^{nH(X)} \times 2^{nH(Y)}\right) = 2^{-nI(X,Y)}$. Hence the probability that another pair (x',y) be typical is approximately this probability times the number of candidate messages, that is $\left(2^{nR} - 1\right) \times 2^{-nI(X,Y)}$ which goes to zero when n goes to infinity as soon as $R < I(X,Y)$.

To turn this into a proof, the bright idea (which belongs to Shannon) is to look first at the average

$$\overline{\lambda^{(n)}}(\mathcal{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w^{(n)}(\mathcal{C})$$

over all source words w of the probability of error for a given code, *and then to average it over all possible codes* $\mathcal{C}$, that is to compute the double average

$$\sum_{\mathcal{C}} Pr(\mathcal{C}) \overline{\lambda^{(n)}}(\mathcal{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} Pr(\mathcal{C}) \lambda_w^{(n)}(\mathcal{C}).$$

Now, because of the average over all possible codes $\mathcal{C}$, the sum $\sum_{\mathcal{C}} Pr(\mathcal{C}) \lambda_w^{(n)}(\mathcal{C})$ is *independent of* w, hence equal for example to $\sum_{\mathcal{C}} Pr(\mathcal{C}) \lambda_1^{(n)}(\mathcal{C})$. It follows that this double average amounts to the probability (computed over all random codes) of getting an error when decoding the output obtained when the word 1 is coded and sent through the channel, that is the probability

$$Pr(\mathcal{E}_1) = \sum_{\mathcal{C}} Pr(\mathcal{C}) \lambda_1^{(n)}(\mathcal{C})$$

over all random codes of the event $\mathcal{E}_1$ which corresponds to decoding an error after having sent the codeword $f(1)$. The event $\mathcal{E}_1$ can be decomposed into

$$\mathcal{E}_1 = E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}} ,$$

where, by definition, the event E_w occurs if the codeword $f(w)$ and the output y_1 received after sending the codeword $f(1)$ are *joint typical*.

Now, because the codewords and the codes are chosen randomly, it follows from the Asymptotic Equipartition Property that

i) the probability of E_1^c, that is the probability that $(f(1), y_1)$ are not joint typical, tends to zero when n tends to infinity;

ii) as $f(1)$ and $f(w)$ are independent if $w \neq 1$, so are $f(w)$ and y_1. As when n tends to infinity, there are asymptotically $2^{nH(X)} \times 2^{nH(Y)}$ couples

of a typical element of X^n and a typical element of Y^n, the probability that the pair $(f(w), y_1)$ is joint typical is asymptotically equivalent to the quotient $2^{nH(X\vee Y)}/\big(2^{nH(X)}\times 2^{nH(Y)}\big) = 2^{-nI(X,Y)}$. As there are $2^{nR}-1$ possibilities for $w \neq 1$, the total probability of the events $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ is equivalent[3] to $2^{n(R-I(X,Y))}$ when n tends to infinity. From i) and ii) it follows that the error can be made arbitrarily small when n is large enough as soon as $R < I(X,Y)$. Maximizing among the probability laws on X, one gets the necessary condition that the rate of the code be less than the capacity of the channel: $R < C$.

Finally, a simple trick of throwing away the worst half of the code words in the best code turns the estimate on the average error into an estimate on the maximum error (see [CT] page 202).

Proof of the converse. Let

$$\mathcal{M} = \{1, 2, \cdots, 2^{nR}\}$$

be the set of input messages. As these are supposed to be almost equiprobable, we have approximately $H(\mathcal{M}) = nR$. Recall that, by definition of the information, $H(\mathcal{M}) = H(\mathcal{M}|Y^n) + I(\mathcal{M}, Y^n)$.

Lemma 19 *The following inequalities hold*
1) $I(\mathcal{M}, Y^n) \leq I(X^n, Y^n)$ *(Data processing inequality),*
2) $I(X^n, Y^n) \leq nI(X,Y)$ *(chain rule).*

Intuitively, *1)* asserts that no processing of X^n (here replacing it by $\mathcal{M}$) can increase the information that Y^n contains about X^n.

It follows that $nR = H(\mathcal{M}) \leq H(\mathcal{M}|Y^n) + nC$, which shows that the problem is to estimate the conditional entropy $H(\mathcal{M}|Y^n)$. Now, if the channel was perfect, i.e. if it was doing no mistake, the decoding would be exact and this would imply $H(\mathcal{M}|Y^n) = 0$. In the general case, the converse of the theorem is a consequence of the

Lemma 20 (Fano 1952) $H(\mathcal{M}|Y^n) \leq 1 + \overline{\lambda}^{(n)} nR$.

Indeed, it follows that

$$R \leq C + \overline{\lambda}^{(n)} R + \frac{1}{n},$$

hence the proof of the converse when $n \to \infty$.

Proof of Lemma 19
1) Given $w \in \mathcal{M}, x \in X$ and $y \in Y$, the following holds:

$$p\big((w,y)|x\big) = p(w|x)p(y|x).$$

Indeed, *x being given, w and y are conditionally independent.* By definition of conditional probabilities, after multiplying by the common denominator $p(x)$, this equality can be written $p(w,x,y) = p(w,x)p(y|x)$, that is

$$p(w,x,y) = p(w)p(x|w)p(y|x),$$

[3] for a proof with all the epsilons, see [CT].

or is in a symmetric form

$$p(w,x,y)p(x) = p(w,x)p(x,y).$$

In [CT], a triple $\mathcal{M} \leftrightarrow X^n \leftrightarrow Y^n$ satisfying the above equation is called a *Markov chain*. This property implies the *data processing inequality*: indeed,

$$\begin{aligned} I(Y^n|\mathcal{M} \vee X^n) &= I(\mathcal{M}, Y^n) + I\big((X^n, Y^n)|\mathcal{M}\big) \\ &= I(X^n, Y^n) + I(\mathcal{M}, Y^n|X^n). \end{aligned}$$

On the one hand, as any information quantity, $I\big((X^n, Y^n)|\mathcal{M}\big) \geq 0$; on the other hand, the conditional independence of $\mathcal{M}$ and Y^n implies $I(\mathcal{M}, Y^n|X^n) = 0$ and the conclusion follows: $I(\mathcal{M}, Y^n) \leq I(X^n, Y^n)$, which is the very intuitive statement that no processing of X^n (in the present case, replacing X^n by $\mathcal{M}$) can increase the information that Y^n contains about X^n.

2) The chain rule, which says that a memoryless discrete channel does not gain in transmission if it is used several times, is a consequence of the following identities: because of the independance of the codewords which are sent, the same is true of the code words which are received. In other words, the random variables $Y_1, Y_2, \cdots Y_n$ are i.i.d. (independent and identically distributed). Hence $Y^n = Y_1 \vee Y_2 \vee \cdots \vee Y_n$, and one shows by induction that

$$\begin{aligned} H(Y^n) &= \sum_{i=1}^{n} H(Y_i|Y_1 \vee Y_2 \vee \cdots \vee Y_{i-1}) \leq \sum_{i=1}^{n} (Y_i), \\ H(Y^n|X^n) &= \sum_{i=1}^{n} H(Y_i|Y_1 \vee Y_2 \vee \cdots \vee Y_{i-1} \vee X^n) = \sum_{i=1}^{n} H(Y_i|X_i), \end{aligned}$$

the last equality resulting from the fact that X_i depends only on Y_i. It follows that

$$\begin{aligned} I(X^n, Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i=1}^{n} I(X_i, Y_i) = nI(X, Y). \end{aligned}$$

Proof of Fano's lemma 20 Let

$$E : \mathcal{M} \times Y^n \to \{0, 1\}$$

be the random variable defined by

$$\begin{aligned} &E(w, y) = 0 \text{ if } g(y) = w, \\ &E(w, y) = 1 \text{ if } g(y) \neq w \text{ or if there is a decoding error,} \end{aligned}$$

where g is the decoding function.

One has by definition, $Pr(E=1) = \overline{\lambda}^{(n)}$ and $Pr(E=0) = 1 - \overline{\lambda}^{(n)}$. Moreover, as the knowledge of w and y determines $g(y)$ and hence E, one has

$$H(E|\mathcal{M} \vee Y^n) = 0.$$

Hence,

$$H(\mathcal{M} \vee E|Y^n) = H(\mathcal{M}|Y^n) + H(E|\mathcal{M} \vee Y^n) = H(\mathcal{M}|Y^n).$$

On the other hand, exchanging the roles of $\mathcal{M}$ and E,

$$H(\mathcal{M} \vee E|Y^n) = H(E|Y^n) + H(\mathcal{M}|E \vee Y^n) \leq H(E) + H(\mathcal{M}|E \vee Y^n).$$

But as E takes only two values, one has $H(E) \leq 1$, hence

$$H(\mathcal{M} \vee E|Y^n) \leq 1 + H(\mathcal{M}|E \vee Y^n).$$

Finally, one computes

$$\begin{aligned} H(\mathcal{M}|E \vee Y^n) &= \sum_y Pr(0,y) H(\mathcal{M}|(0,y)) + \sum_y Pr(1,y) H(\mathcal{M}|(1,y)) \\ &= (1 - \overline{\lambda}^{(n)}) \sum_y Pr(y|0) H(\mathcal{M}|(0,y)) + \overline{\lambda}^{(n)} \sum_y Pr(y|1) H(\mathcal{M}|(1,y)). \end{aligned}$$

By definition of the conditional entropy as an average, the first term is equal to $(1 - \overline{\lambda}^{(n)}) H(\mathcal{M}|(E=0) \vee Y^n)$ and hence vanishes because $E = 0$ means that the knowledge of y determines w.

The second term is equal to $\overline{\lambda}^{(n)} H(\mathcal{M}|(E=1) \vee Y^n) \leq \overline{\lambda}^{(n)} H(\mathcal{M}) = \overline{\lambda}^{(n)} nR$ because the elements of $\mathcal{M}$ are supposd to be equiprobable. This concludes the proof.

Shannon's theorems admit generalizations to more realistic sources, for example Markov sources, where the probability of emitting a letter (or a word) depends on the letters (words) which have been emitted before. More generally, they are valid for the general case of *ergodic sources*. We need first introduce the language of *dynamical systems*[4].

5 Random draws and Bernouilli shifts

The stochastic properties of an infinite sequence of independent draws (with respective probabilities p, q of $0, 1$), are nicely reflected in the dynamical properties of a single object, the *Bernoulli shift*)

$$T : (\{0,1\}^{\mathbb{N}^*}, \mathcal{F}, P_{p,q}) \to (\{0,1\}^{\mathbb{N}^*}, \mathcal{F}, P_{p,q}), \quad T(a_1a_2a_3\ldots) = (a_2a_3a_4\ldots).$$

Forgetting a_1, this map is surjective but not injective: the inverse image of any element consists in a pair of elements. It preserves any of the probability measures $P = P_{p,q}$ on the Borel tribe $\mathcal{F}$ of $\{0,1\}^{\mathbb{N}^*}$: indeed, the inverse image $T^{-1}(A)$ of the cylinder $A = A^{j_1 j_2 \ldots j_k}_{i_1 i_2 \ldots i_k}$ is the cylinder $A^{j_1 j_2 \ldots j_k}_{i'_1 i'_2 \ldots i'_k}$, where $i'_n = i_n + 1$; as te process is stationary, it has the same probability $p^{k_0}q^{k_1}$ as A and one concludes by lemma 4.

Orbits and dynamics. An *orbit* $\{\omega, T\omega, T^2\omega, \ldots, T^n\omega, \ldots\}$ of T is a dynamical description of the sequence of draws ω and the language and methods of of the *theory of dynamical systems* – for which such an orbit is the discrete version of a an integral curve – is remarkably pertinent in the description of this type of stationary processes.

Exercise 3 *When $p = q = \frac{1}{2}$, the translation of T in the world of the interval $[0, 1]$ is the mapping $x \mapsto 2x(mod.1) = 2x - [2x]$ which is easily shown directly to preserve Lebesgue measure ([x] means the integer part of x).*

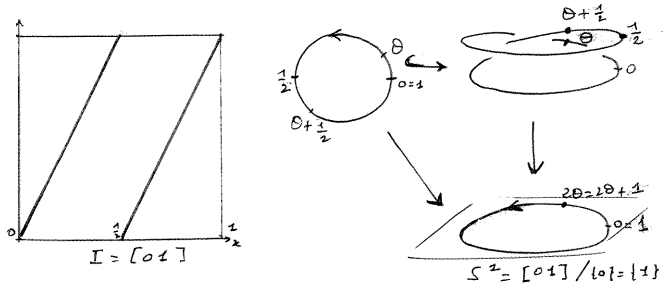

Figure 12 : The map $x \mapsto 2x$ on the interval and on the circle.

[4]for details and proofs, see [C]

From simply infinite sequences to doubly infinite sequences:
It is often more pleasant to work with invertible transformations and, in our case, this is accomplished by the consideration of a double infinity of draws, both in the past and in the future. Correspondingly, one defines a bimeasurable bijection $T : \{0,1\}^{\mathbb{Z}} \to \{0,1\}^{\mathbb{Z}}$ which preserves all the probability measures $P = P_{p,q}$ by setting

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

Exercise 4 *One supposes that $p = q = \frac{1}{2}$. Show that if $g_2 : \{0,1\}^{\mathbb{Z}} \to [0,1]^2$ is defined by*

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

the direct image of P by g_2 is the Lebesgue measure on $[0,1]^2$ and that g_2 conjuguates measurably $T : \{0,1\}^{\mathbb{Z}} \to \{0,1\}^{\mathbb{Z}}$ to the transformation $\tau : [0,1]^2 \to [0,1]^2$ defined by

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

where $[2y]$ denotes the largest integer $\leq 2y$ (clearly, $2y - [2y]$ is nothing but $2y(mod 1)$). Explain why dynamicists call τ the "baker's transformation".

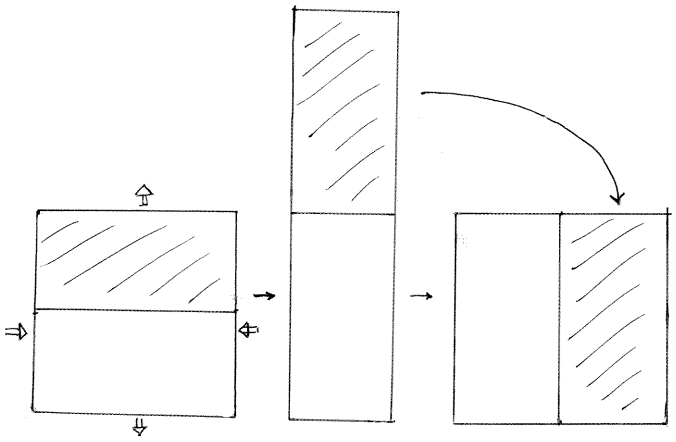

Figure 13 : The baker's transformation.

Stationary stochastic processes and shifts : dictionary
Up to now we have supposed that the successive draws were independent and this was reflected by the choice of the probability measure $P = P_{p,q}$ on $\{0,1\}^{\mathbb{Z}}$. In information theory , one sends messages whoses structure is in general not so simple: in any natural language, the probability that some letter follows another one depends on the two letters and this leads to the considération of more complex processes, for example the *Markov chains* which we shall define.

Definition 14 (discrete source) *A (stationary) discrete source is a shift-invariant probability measure defined on the Borelian tribe of $\{0,1\}^{\mathbb{Z}}$ (resp. of $A^{\mathbb{Z}}$, where $A = \{A_1, A_2, \ldots, A_r\}$ is a finite alphabet).*

This is equivalent to giving a stationary stochastic process with values in $\{0,1\}$ (resp. in A) in the following sense:

Definition 15 *A discrete time stochastic process on the probability space $(\Omega, \mathcal{F}, \mu)$ with values in the topological space X is a sequence $\{\xi_n\}_{n\in\mathbb{Z}}$ of random variables $\xi_n : \Omega \to X$.*

To such a process on associates a probability measure ν on $X^{\mathbb{Z}}$ by setting

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

where $\tilde{\xi}(\omega) = \ldots \xi_{-2}(\omega)\xi_{-1}(\omega)\xi_0(\omega)\xi_1(\omega)\xi_2(\omega)\ldots$, that is by defining this measure as the direct image $\nu = \tilde{\xi}_*\mu$ of μ by the map $\tilde{\xi}$. The processus is said to be *stationary* if ν is invariant by the shift T. If the random variables ξ_i are independent, equally distributed and with values in the finite alphabet$X = (A_1, \cdots, A_r)$ with image probability measure $(p_1, \cdots, p_r)$, the measure $\nu = \tilde{\xi}$ on $X^{\mathbb{Z}}$ coincides with $P_{p_1,\cdots,p_r}$.

6 Ergodicity and Birkhoff's theorem

6.1 Ergodicity

Independence of the coin tosses in a "heads or tails" game implies "forgetting of the initial condition": each toss ignores the result of all the former tosses; to this corresponds a very strong property of the Bernouilli shifts, called *ergodicity*[5]:

Definition 16 (Ergodicity) *Let $(X, \mathcal{X}, \mu, T)$ be a measured dynamical system*[6]*. One says that T (or that the dynamical system) is ergodic if every set $A \in \mathcal{X}$ which is invariant*[7] *by T satisfies $\mu(A) = 0$ or $\mu(A) = 1$. When T is given, one says also that the invariant measure μ is ergodic.*

Exercise 5 *Show that the map T is ergodic if and only if any one of the following properties is satisfied:*
1) every measurable T-invariant function $f : X \to \mathbb{C}$ is a.e. constant;
2) There exists $p \geq 1$, such that every T-invariant function $f \in L^p(X, \mathbb{C})$ is a.e. constant.

[5] for a more detailed presentation, see [C]

[6] That is a probability space $(X, \mathcal{X}, \mu)$ and a measure preserving map T from this space to itself.

[7] here, the precise meaning of "A is invariant" is $\mu(A \Delta T^{-1}(A) = 0$.

6.2 Mixing

In order to prove ergodicity of the Bernoulli shifts, we shall prove that they are *mixing*, a strictly stronger property:

Definition 17 (Mixing) *Let $(X, \mathcal{X}, \mu, T)$ be a measured dynamical system. One says that T is mixing if for any $A, B \in \mathcal{X}$,*

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

Definition 18 (Weak mixing) *Let $(X, \mathcal{X}, \mu, T)$ be a measured dynamical system. One says that T is weak mixing if for any $A, B \in \mathcal{X}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \left|\mu(A \cap T^{-k}(B)) - \mu(A)\mu(B)\right| = 0.$$

Exercise 6 *Mixing implies weak mixing and weak mixing implies ergodicity.*

Exercise 7 *T is mixing if and only if for every $f, g \in L^2(X, \mathcal{X}, \mu)$,*

$$\lim_{n\to\infty} \int_X f \cdot (g \circ T^n) d\mu = \left(\int_X f d\mu\right)\left(\int_X g d\mu\right);$$

it is weak mixing if and only if for every $f, g \in L^2(X, \mathcal{X}, \mu)$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \left|\int_X f \cdot (g \circ T^k) d\mu - \left(\int_X f d\mu\right)\left(\int_X g d\mu\right)\right| = 0.$$

Theorem 21 *Bernouilli shifts T on $\{0,1\}^{\mathbb{N}^*}$ or $\{0,1\}^{\mathbb{Z}}$ are mixing (and hence ergodic) for any one of the product probability measures $\mu = P_{p,q}$.*

Proof. It is enough to check the defining property on the algebra $\mathcal{G}$ of finite union of cylinders which generates the Borelian tribe[8]. But, given two finite unions of cylinders A_0 and B_0, the set of indices associated to A and $T^{-n}(B_0)$ are disjoint as soon as n is large enough, and this implies that $\mu(T^{-n}(A_0) \cap B_0) = \mu(A_0)\mu(B_0)$. The end of the proof is left to the reader.

Corollary 22 *The map $x \mapsto 2x(mod.1) : [0,1] \to [0,1]$ and the baker map $\tau : [0,1]^2 \to [0,1]^2$ are mixing, and hence ergodic, for the Lebesgue measure.*

6.3 Birkhoff's ergodic theorem

Originating in Poincaré"s recurrence theorem, the *ergodic theorem* was proved by Birkhoff in 1931. An important generalization implying Birkhoff's theorem, the *subadditive ergodic theorem*, was given by Kingman in 1968. The nice proof of Kingman's theorem by Avila and Boschi (2009) is explained in [C].

[8] for details, see [C]

Theorem 23 *Let $(X, \mathcal{X}, \mu, T)$ be a measured dynamical system. For every function $f \in L^1(X, \mathcal{X}, \mu)$, the limit of "Birkhoff sums"*

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

exists for μ-almost every $x \in X$ and it defines a function $f^* \in L^1(X, \mathcal{X}, \mu)$ satisfying $f^* \circ T = f^*$ (μ-a.e.) and $\int_X f(x)\, d\mu(x) = \int_X f^*(x)\, d\mu(x)$.

Remark. *If T is invertible, the functions f^* and $\bar{f}^*$ respectively defined by T and T^{-1} coincide almost everywhere.*
Indeed, suppose that the a.e. T-invariant set $Y = \{x \in X,\ f^* > \overline{f}^*\}$ has positive measure. Applying Birkhoff's theorem to the restrictions of T and T^{-1} to Y one gets

$$\int_Y f^* d\mu = \int_Y f d\mu = \int_Y \overline{f}^* d\mu\,,$$

hence $\int_Y (f^* - \overline{f}^*) d\mu = 0$. But this is a contradiction because $f^* - \overline{f}^*$ is strictly positive on Y.

Corollary 24 *Under the same hypotheses, if moreover T is ergodic, f^* is a constant, equal to $\int_X f d\mu$.*

In words, this means that if T is ergodic, the *time average*, that is the limit of the *Birkhoff sums* exists almost everywhere and is equal to the integral, that is to the *spatial average*. If for example f is the characteristic function $\mathcal{X}_A$ of a measurable subset $A \in \mathcal{X}$, the corollary asserts that, for almost every x, the proportion of "time" the orbit of x spends in A coincides with the measure (the probability) of A ($n \in \mathbb{N}^*$ or $n \in \mathbb{Z}$ should indeed be interpreted as a discrete time, the unit of time corresponding to one iteration of T).

6.4 Applications: strong forms of the Law of Large Numbers

Applied to the Bernoulli shifts, corollary 24 says that the statistical structure of almost all sequences is the same, which is a strong form of the so-called *strong law of large numbers*. In what follows, we consider only the case of random variables with finite values.

Theorem 25 (Stong law of large numbers in the independent case) *If $f_1, \cdots, f_n, \cdots : (X, \mathcal{X}, \mu) \to \mathbb{R}$ are independent and identically distributed random variables whose values are $A_1, \cdots, A_r$ with probabilities $p_1, \cdots, p_r$, one has*

$$Pr\left\{ \lim_{n\to\infty} \frac{1}{n}(f_1 + \cdots + f_n) = \sum_{i=1}^{r} p_i A_i \right\} = 1.$$

Proof. We apply corollary 24 to the generalized shift)

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

and to the functions $f_i(a_1 \cdots a_n \cdots) = a_i = f_1(T^{i-1}(a_1, \cdots, a_n, \cdots))$, which are a universal model of i.i.d. random variables. The conclusion follows because, on the one hand $f_1(x) + \cdots + f_n(x) = \sum_{k=0}^{n-1} f_1(T^k(x))$, on the other hand the integral of f_1 on $\{A_1, \cdots, A_r\}^{\mathbb{N}^*}$ is equal to $\sum_{i=1}^{r} p_i A_i$.

Applying the law of large numbers to the random variables

$$f_i : \{A_1, \ldots, A_r\}^{\mathbb{N}^*} \to \mathbb{R},\ f_i(a_1 \cdots a_n \cdots) = \log \frac{1}{p_{a_i}},\ \text{where } p_{a_i} = p_j \text{ if } a_i = A_j,$$

one gets

Corollary 26 (Strong form of the Asymptotic Equipartition Property) *Given $\xi_1, \cdots, \xi_n, \cdots : (\Omega, \mathcal{F}, \mu) \to \mathbb{R}$, independent and identically distributed random variables with values in $\{\alpha_1, \cdots, \alpha_r\}$ and image probability measure $(p_1, \cdots, p_r)$, one has*

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

The following proposition justifies the adjectives “”weak and “strong”:

Proposition 27 *The strong law implies the weak law*

The proof is of a classical nature in probability theory: *“convergence with probability 1 implies convergence in probability”*.

Proof. Let X_n, $n \in \mathbb{N}$, be a sequence of random variables converging with probability 1 to the random variable X:

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

The complement of the subset L of measure 1 defined in this formula is

$$L^c = \cup_\epsilon \{\omega \in \Omega,\ |X_n(\omega) - X(\omega| \geq \epsilon \text{ for an infinity of } n\}.$$

It suffices in fact to take the union on the countable set of rational ϵ's, which shows that L is measurable. The proposition is then consequence of

Lemma 28 *Suppose that*

$$\mu\{\omega \in \Omega,\ |X_n(\omega) - X(\omega| \geq \epsilon \textit{ for an infinity of } n\} = 0\,;$$

Then

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

By definition, if this last property is satisfied for every $\epsilon > 0$, X_n converges to X in probability.

Proof. Let $G_n = \{\omega \in \Omega,\ |X_n(\omega) - X(\omega| \geq \epsilon\}$. The set of those ω which belong to G_n for an infinity of values of n is by definition $\limsup G_n$; it can be defined by the formula

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

From this follows that

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

Indeed, $\limsup G_n$ is the intersection of the $U_n = \cup_{k=n}^{\infty} G_k$ which are a decreasing sequence of subsets (i.e. $U_{n+1} \subset U_n$). Hence $\lim_{n\to\infty} \mu(U_n) = \mu(\limsup G_n)$. But $G_n \subset U_n$ hence $\mu(G_n) \leq \mu(U_n)$ and finally

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

The hypothesis of the lemma is that $\mu(\limsup G_n) = 0$. One deduces that $\limsup \mu(G_n) \leq 0$ and hence, as μ has positive values, that $\lim \mu(G_n) = 0$.

An example of a more precise result Consider in $(\{0,1\}^{\mathbb{N}^*}, \mathcal{B}, P_{p,q})$ the cylinder A defined by $a_1 = a_2 = \ldots = a_{1000} = 0$.
The Birkhoff sum $\frac{1}{n}\sum_{k=0}^{n-1} \mathcal{X}_A\big(T^k(x)\big)$, where T is the shift, represents the frequency with which one gets $a_{k+1} = a_{k+2} = \ldots = a_{k+1000} = 0$ when k varies from 0 to n. the theorem asserts that, for almost every sequence, this frequency tends to a limit equal to p^{1000}, when n tends to $+\infty$. This occurs for any cylinder, that is for any finite motive of 0's and 1's and it is far from exhausting the content of the theorem as the function f could depend on an arbirtrary number of coordinates.

7 Beyond independence: a glance at Markov chains

If the sequences $a_1 a_2 \ldots a_n$ we consider are sentences written in some language, the probability of an individual letter is not independent of the preceding one (or more generally of the preceding ones). In other words, the random variable $a_1 a_2 \ldots a_n \mapsto a_i$ is not independent of the random variable $a_1 a_2 \ldots a_n \mapsto a_{i-1}$. Let $A = \{A_1, \ldots, A_r\}$ be a finite alphabet. The simplest probability laws on A^n taking this into account are the so-called *Markov chains with one step memory* characterized by the *initial probabilities* $(p_1, \ldots, p_r)$ and the *conditional probabilities* $p_{ij}, 1 \leq i, j \leq r$, which are the probabilities that A_j follows A_i, the conditions being that these numbers be all non negative and such that $\sum_{i=1}^r p_i = 1$ and, for $i = 1, \ldots, r$, $\sum_{j=1}^r p_{ij} = 1$ (a matrix (p_{ij}) with non negative entries and the eigenvector $(1, 1, \ldots, 1)$ with eigenvalue 1 is called *stochastic*). The probability of the sequence $A_{i_1} A_{i_2} \cdots A_{i_n}$ is by definition the product

$p_{i_1}p_{i_1i_2}p_{i_2i_3}\dots p_{i_{n-1}i_n}$. Let us describe in more details the case when $A=\{0,1\}$ has only 2 elements. The conditional probabilities are conveniently represented by the graph in figure 14:

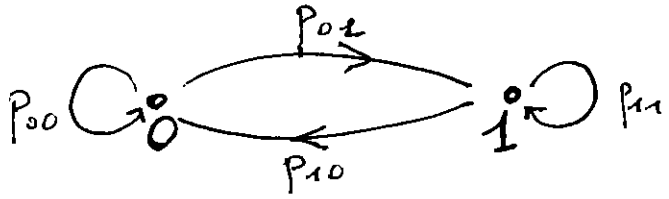

Figure 14 : A one step Markov chain.

A straightforward calculation shows that, given a sequence $a_1a_2\dots a_n\in A^n$, the probabilities $p_0^{(k)}$ and $p_1^{(k)}$ that a_k be 0 or 1 are given by the matrix identity

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

In particular, if $(p_0p_1)=(p_0p_1)M$, the probabilities that $a_k=0$ or $a_k=1$ are respectively p_0 and p_1 independently of k. One then deduces from the probabilities of finite sequences in $\{0,1\}^n$ a probability measure P_{M,p_0,p_1} on the set $\{0,1\}^{\mathbb{N}^*}$ of infinite sequences endowed with the tribe generated by the *cylinders* defined by fixing the values (0 or 1) of a finite number of terms $a_{k_1},a_{k_2},\dots,a_{k_n}$; the above condition is equivalent to the stationarity of this measure, i.e. its invariance under the *shift*

$$a_1a_2\dots a_n\dots\mapsto a_2a_3\dots a_{n+1}\dots$$

The following lemma, a simple consequence of the fixed point theorem for contractions, implies, under the given hypotheses, the *ergodicity* of this measure, that is (recall 6.1) the fact that any measurable subset of $\{0,1\}^{N^*}$ invariant under the shift has measure 0 or 1.

Lemma 29 (Perron-Frobenius) *Let M be a 2×2 matrix with non negative coefficients such that the vector with coordinates $(1,1)$ is invariant. Suppose there exists an integer s such that the matrix M^s has all its coefficients, noted $p_{ij}^{(s)}$, strictly positive. The there exists a unique probability (p_0,p_1) on $\{0,1\}$ such that*
1) $(p_0\,p_1)M=(p_0\,p_1)$,
2) $\lim_{s\to\infty}p_{ij}^{(s)}=p_j,\ j=0,1$.

The interpretation of the hypotheses is that for any pair $i,j\in\{0,1\}$, there is a positive probability that if $a_k=i$, then $a_{k+s}=j$. The interpretation of the conclusion is that, *1)* the measure is stationary (i.e. invariant under the shift), *2)* the conditions $a_k=i$ and $a_{k+N}=j$ are asymptotically independent when $N\to\infty$: the conditional probability $p_{ij}^{(N)}$ depends less and less on i.

This last property implies mixing, hence ergodicity: indeed, let us consider first elementary cylinders $A_i^j=\{(a_1\cdots a_n\cdots)\in\{0,1\}^{\mathbb{N}^*},a_i=j\}$ fixing a single

term of the sequence; we see that

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

(one supposes $n + i_1 > i_2$) tends to $p_{j_2} p_{j_1} = P_{M,p_0,p_1}(A_{i_1}^{j_1}) P_{M,p_0,p_1}(A_{i_2}^{j_2})$ when n tends to $+\infty$. The full proof follows by considering first finite unions of arbitrary cylinders, which proves the mixing property on the elements of the algebra generated by the cylinders, then showing that this implies the same for the σ-algebra.
All this generalizes to Markov chains associated to any finite alphabet.

8 From Shannon's entropy to Kolmogorov's entropy

Let $A = (A_1, \cdots, A_r)$ be a finite set endowed with a probability measure $(p_1, \cdots, p_r)$. Recall that, in the case of independent draws, the entropy $\sum_{i=1}^r p_i \log \frac{1}{p_i}$ may be defined as the limit $\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{p(a_1 \cdots a_n)}$ for $P_{p_1,\cdots,p_r}$-almost every sequence $a_1 \cdots a_n \cdots \in A^{\mathbb{N}^*}$. This is a direct consequence of the ergodic theorem applied to the random variable

$$f : A^{\mathbb{N}^*} \to \mathbb{R}, \quad f(a_1 a_2 \cdots a_n \cdots) = \log \frac{1}{p_{a_i}}$$

(Recall the notations : $p_{a_i} = p_j$ if $a_i = A_j$ and, more generally, $p_{a_1 \cdots a_n}$ is the measure of the cylinder $A_{1\cdots n}^{a_1 \cdots a_n}$; it will be convenient to identify A to the set $\{1, 2, \cdots, r\}$ and to use the notation $A_{1\cdots n}^{a_1 \cdots a_n} = A_{1\cdots n}^{j_1 j_2 \cdots j_n}$ if $a_k = A_{j_k}$.)
If we endow $A^{\mathbb{N}^*}$ with an ergodic Markov measure $\mu = P_{M;p_1,\cdots,p_r}$ defined by initial probabilities p_i and conditional probabilities p_{ij}, one gets that for μ-almost every sequence $a_1 \cdots a_n \cdots \in A^{\mathbb{N}^*}$,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu(a_1 a_2 \cdots a_n)} = \sum_{i,j=1}^r p_i p_{ij} \log \frac{1}{p_{ij}}.$$

Indeed, one applies the ergodic theorem to the function

$$g : \Omega \to \mathbb{R}, \quad g(a_1 a_2 \ldots a_n \ldots) = \log \frac{1}{p_{a_1 a_2}},$$

which leads to

$$\frac{1}{n} \log \frac{1}{\mu(a_1 a_2 \ldots a_n)} = \frac{1}{n} \log \frac{1}{\mu(a_1)} + \frac{1}{n} \sum_{i=0}^{n-1} g(T^i(a_1 a_2 \ldots)$$

and

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

The function g is indeed constant on the atoms of the partition $\Omega = \sum_{i,j=1}^r A_{12}^{ij}$: it is equal to $\log \frac{1}{p_{ij}}$ on A_{12}^{ij} and $\mu(A_{12}^{ij}) = p_i p_{ij}$.

8.1 The entropy of a Markov chain

The considerations above lead us to define the entropy of a Markov chain by the formula

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

With the notations used to define conditional entropy, H is the expectation of the random variable $A_i \mapsto H_{A_i}(A) = \sum_{k=1}^{r} p_{ik} \log \frac{1}{p_{ik}}$. It is the expectation of a draw following the one of the letter A_i, the probability the the result be A_k being the conditional probability p_{ik}. In other words, H is simply the conditional entropy $H_A(A^2)$, which, in the independent (= Bernoulli) case reduces to the definition we have given.

8.2 Entropy as the mean information content by symbol

Let μ be a probability measure on $A^{\mathbb{N}^*}$ (or $A^{\mathbb{Z}}$) which is invariant under the shift (for example $P_{p_1,\dots,p_r}$ in the Bernoulli case, $P_{M;p_1,\dots,p_r}$ in the Markov case). Let $H_\mu^{<n>}$ be the entropy of the finite set A^n endowed with the probability measure defined by the measure of cylinders, that is of the direct image under the canonical projection $\pi_n : A^{N^*} \to A^n$ (or $A^{\mathbb{Z}} \to A^n$), $\pi(\cdots a_i \cdots) = a_1 a_2 \cdots a_n$ of the measure μ :

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

The entropy $H_\mu^{<n>}$ is a measure of the information obtained from the emission of a sequence of n successive symbols (or n successive experiments).

1) the Bernouilli case : if $\mu = P_{p_1,\dots,p_r}$, a direct computation shows that

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

2) the Markov case : if $\mu = P_{M;p_1,\dots,p_r}$, we set

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

This is the conditional probability that, A_i being realized, a sequence of n draws result in $A_{j_1}, A_{j_2}, \dots, A_{j_n}$. As in the case $n = 1$, one defines the entropy of the chain iterated n times by

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

Lemma 30 *The following identities hold true:*

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

Proof. $H^{(n)}$ is the entropy associated to a sequence of n successive draws. Reasoning by induction on n, we suppose that the system is in the state A_i and we decompose a sequence of $n+1$ draws into a first draw (called the event A) followed by a sequence of n draws (called the event B). These two events are not independent and the formula $H(A \vee B) = H(A) + H_A(B)$ becomes :

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

But $(p_1, p_2, \ldots, p_r)$ being a probability vector associated to the matrix M of conditional probabilities, one has $\sum_{i=1}^{r} p_i p_{ik} = p_k$ and hence

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

which proves the first identity. The second identity results from an explicit computation.

8.3 The entropy of a discrete source

A discrete source is the data of a finite alphabet $A = \{A_1, \ldots A_r\}$ (for example $\{0, 1\}$) and a probability measure μ on $\Omega = A^{\mathbb{N}^*}$ (ou $A^{\mathbb{Z}}$) which is invariant under the shift T. *The probability of a given result of a draw depending a priori of the whole past history, we need consider arbitrarily long sequences in order to define an entropy.*

Computations made in the former section show that the definition of entropy given in the following lemma generalizes the ones given in the cases when $\mu = P_{p_1,\ldots,p_r}$ is Bernouilli or when $\mu = P_{M;p_1,\ldots,p_r}$ is Markov.

Lemma 31 (McMillan) *The "mean information content by symbol" $\frac{1}{n} H_\mu^{<n>}$ tends to a limit $H_\mu = H_\mu(T)$ when the length n of the sequence (the message) tends to infinity :*

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

is by definition the entropy of the source.

Proof. One decomposes as above the emission of a sequence of $n+m$ symbols into two events, the second of which depends on the first: the emission X_n of the n fist symbols followed by the one Y_m of the m last symbols.
Noting $u_n = H_\mu^{<n>}$, one has

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

this "subadditivity" of the sequence $u_n = H_\mu^{<n>}$ is the key of the proof : Let $v = \inf_n(\frac{u_n}{n})$. By definition of v, given any $\epsilon > 0$, there exists $N > 0$ such that

$u_N < N(v+\epsilon)$. But euclidean division allows to write every integer n in the form $n = kN + r$ with $k \geq 0$ and $1 \leq r \leq N-1$. By subadditivity, this implies $u_n \leq u_{kN} + u_r \leq ku_N + \rho_N$, where $\rho_N = \sup_{1 \leq r \leq N-1}(u_r)$. Finally,

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

One deduces that the sequence $\frac{u_n}{n}$ converges to v, which proves the lemma.

In the next section, I briefly alludes to the remarkable generalization of Shannon – McMillan's entropy given by Kolmogorov in case the shift T is replaced by any measure preserving transformation of a probability space into itself.

8.4 Kolmogorov's entropy

Definition 19 (Entropy of a finite partition) *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $\mathcal{E}$ a finite partition $\Omega = A_1 + A_2 + \ldots + A_r$ (to which we can think as a finite subalgebra of $\mathcal{F}$). The entropy of $\mathcal{E}$ is*

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

Notations. Given a transformation $T : \Omega \to \Omega$ and a partition $\mathcal{E}$, we note $T^{-1}\mathcal{E}$ the algebra of subsets formed by the $T^{-1}(A_i)$, $A_i \in \mathcal{E}$. Given finite partitions $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(m)}$ of Ω, we note $\vee_{i=1}^m \mathcal{E}^{(i)}$ the partition whose atoms are the intersections $A_{k_1}^{(1)} \cap A_{k_2}^{(2)} \cap \ldots \cap A_{k_m}^{(m)}$, where $A_{k_i}^{(i)}$ is an atom of $\mathcal{E}^{(i)}$.

Definition 20 *The entropy $H_\mu(\mathcal{E}, T)$ of a partition $\mathcal{E}$ with respect to a measure preserving transformation $T : \Omega \to \Omega$ and the entropy $H_\mu(T)$ of the transformation T are respectively defined by*

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

where the sup is taken among all finite partitions $\mathcal{E}$ of Ω.

Explanation (see [B2]) : an element $A_{k_1}^{(1)} \cap A_{k_2}^{(2)} \cap \ldots \cap A_{k_m}^{(m)}$ of the partition $\vee_{i=1}^m \mathcal{E}^{(i)}$ may be considered as the realization of m experiments, corresponding to the m partitions $\mathcal{E}^{(i)}$. Given a partition $\mathcal{E}$, let us denote by $A = \{A_1, \ldots, A_r\}$ the set of atoms of the partition and by $x : \Omega \to A$ the random variable which, to an element $\omega \in \Omega$, associates the atom A_i to which it belongs. As T preserves the measure μ, the image measures of μ by the random variables $x \circ T^n$ are all the same (n is an integer or a relative integer if T is invertible). In other words, the experiments corresponding to the partitions $T^{-n}(\mathcal{E})$ have all the same probabilistic structure and hence they can be considered as realizations, a priori not independent, of one and the same experiment.

The case when $T : A^{\mathbb{N}^*} \to A^{\mathbb{N}^*}$, $A = \{A_1, A_2, \ldots, A_r\}$, is a shift and $\mathcal{E}$ is the partition into elementary cylinders

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

sheds light on this assertion : the atoms of the partition $\vee_{k=0}^{n-1} T^{-k}\mathcal{E}$ are the cylinders defined by fixing $a_1, a_2, \ldots, a_n$; the experiments are independent in the Bernoulli case, they are not in the Markov case.

Exercise 8 *Show that, in the case of a Bernoulli shift T, for any integer n the entropy $H_\mu\big(\vee_{k=0}^{n-1} T^{-k}\mathcal{E}, T\big)$ is n times the entropy $\sum_{k=1}^r p_k \log \frac{1}{p_k}$ of the finite probability space $(A, p_1, \ldots p_n)$ from which the invariant measure on $A^{\mathbb{N}^*}$ is defined.*

Exercise 9 *Same exercise, replacing Bernoulli by Markov*

We shall admit the following theorem, due to Kolmogorov (see [B2, CFS]); it immediately implies that the entropy of a Bernoulli (resp. Markov) shift is $\sum_{k=1}^r p_k \log \frac{1}{p_k}$ (resp. $\sum_{i,k=1}^r p_i p_{ik} \log \frac{1}{p_{ik}}$) :

Theorem 32 *If $T : (\Omega, \mathcal{F}, \mu)$ is invertible and if there exists a finite partition $\mathcal{E}$ which is generating in the sense that the partitions $\vee_{i=-n}^{n} T^{-i}(\mathcal{E})$, $n = 1, \cdots, \infty$, genrate the σ-algebra $\mathcal{F}$, one has*

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

An analogous statement holds true in the non invertible case if one replaces $\vee_{i=-n}^{n} T^{-i}(\mathcal{E})$ by $\vee_{i=0}^{n} T^{-i}(\mathcal{E})$.

8.5 The Shannon-McMillan-Breiman theorem on ergodic discrete sources

8.5.1 Entropy as expectation

Let us consider a discrete source, that is a finite alphabet A and a probability measure μ on $\Omega = A^{\mathbb{N}^*}$ (or $\Omega = A^{\mathbb{Z}}$) which is invariant under the shift T. The finite probability space A^n is endowed with the probability defined by the measure

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

of the cylinders of length n. Its entropy $H_\mu^{<n>}$, is by definition, the expectation of the random variable $\xi_n : A^n \to \mathbb{R}$ defined by

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

Hence, by definition of the entropy of a general source,

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

8.5.2 The Asymptotic Equipartition Property

Replacing the expectation by the random variable itself, this property has been proved by McMillan and Breiman to hold true for general ergodic sources. The quite technical proof can be found in [B2].

Theorem 33 (Shannon-McMillan-Breiman) *The entropy of an ergodic discrete source (T, μ) satisfies the strong Asymptotic Equipartition Property: for μ-almost every $\omega = \dots a_1 a_2 \dots a_k \dots \in \Omega$,*

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

In other words,

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

The following is a weak form of this statement, closer to the initial statements by Shannon.

Corollary 34 *Given any $\epsilon > 0$, one has*

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

Let us recall the interpretation of the corollary (see figure 4) : by definition of the limit, there exists, for any $\epsilon > 0$, an integer $n(\epsilon)$ with the following property: as soon as $n \geq n(\epsilon)$, the set A^n may be decomposed into two pieces: a "small" (precisely with a cardinal of the order of 2^{nH}, where $H = H_\mu(T)$) subset of sequences of length n almost equiprobable, and the complement whose probability is $\leq \epsilon$. In particular, *it is most of the time possible to treat the sequences of n symbols (n big) as if their total number was only 2^{nH}, each of these sequences having the probability 2^{-nH}.*

A companion assertion, also given by Shannon, follows: let us consider the minimum number of messages of length n whose union has a probability $\geq 1-\delta$. One could estimate this number by by choosing the messages of length n n by order of decreasing probability until the bound $1 - \delta$ is reached.

Definition 21 *The essential information $H_{\mu,\delta}^{<n>}$ of a discrete source is defined by the following formula, in which $|E|$ stands for the cardinal of E :*

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

Theorem 35 *Let (T, μ) be a discrete ergodic source. Given any $\epsilon > 0$ and $0 < \delta < 1$, there exists and integer N such that, for all $n \geq N$, we have*

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

Proof. 1) The real number ϵ and the integer n being given, let us note

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

One deduces from corollary 34 the existence of an integer $n_0 = n_0(\epsilon, \delta)$ such that,

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

As each of the elements $(a_1, a_2, \dots, a_n)$ of $E_1(n,\epsilon)$ satisfies

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

one deduces that $|E_1(\epsilon,\delta)| \leq 2^{n(H_\mu(T)+\epsilon)}$ and hance that

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

2) Conversely, let $E \subset A^n$ be such that $|E| \leq 2^{n(H_\mu(T)-\epsilon)}$. One has

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

Let us choose $\delta' > 0$ such that $\delta + \delta' < 1$. the first term of the right hand side is bounded by $2^{n(H_\mu(T)-\epsilon)} \times 2^{-n(H_\mu(T)-\frac{\epsilon}{2})} = 2^{-n\frac{\epsilon}{2}}$ and hence by $\frac{\delta'}{2}$ dès que $n \geq n_1(\epsilon,\delta')$; the second term is bounded by $\mu(E_1^c(n,\frac{\epsilon}{2}))$ and hence by $\frac{\delta'}{2}$ as soon as $n \geq n_0(\frac{\epsilon}{2}, \frac{\delta'}{2})$. One deduces that, for n large enough, $\mu(E) \leq \delta' < 1 - \delta$. Hence

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

which finishes the proof

The interpretation of this theorem is that, for n large enough, not only there is subset of approximately $2^{nH_\mu(T)}$ "typical" messages of length n among the $2^{n \log |A|}$ messages, with an arbitrarily small probability of encountering a non typical message, but this number cannot be substantially lowered even at the price of rising the error δ which is admitted.

References

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

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

[BB] P. Baudot & D. Bennequin, The Homologiccal Nature of Entropy, Entropy 17, 1–66 (2015)

[C] A. Chenciner, Discrete dynamical systems, Tsinghua University, march 2015), available at `http://www.imcce.fr/fr/presentation/equipes/ASD/person/chenciner/` (section "Polys")

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

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

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

[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)

[MK] D. MacKay, Information theory, inference, learning algorithms, Cambridge University press 2004 `http://www.inference.phy.cam.ac.uk/mackay/info-theory/course.html`

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

[Ru] D. Ruelle, Statistical mechanics, rigorous results, New York, Benjamin (1969)

[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)

[Tv] H. Tverberg, A new derivation of the information function, Math. Scand. 6, p. 297-298 (1958)

[W] N. Wiener, Cybernetics, Hermann 1948

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