Société de Calcul Mathématique SA

Tools for decision help

Entropy and Information

by Bernard Beauzamy

February 2015

Abstract

The link between entropy and information, as it is presented for instance in the book by Léon Brillouin [Brillouin] is not satisfactory at all. There is a misunderstanding about what is sent and what is received. In Wikipedia's article "Entropy (information theory)", the starting sentence "In information theory, entropy is a measure of the uncertainty in a random variable" is almost correct (not completely !), but on the French Wikipedia "Entropie de Shannon", the starting sentence " L'entropie de Shannon, due à Claude Shannon, est une fonction mathématique qui, intuitivement, correspond à la quantité d'information contenue ou délivrée par une source d'information" is not correct at all.

For a given sequence, one may define its variance, which is a way of describing the "analytic" dispersion of the values, whereas the entropy characterizes the "probabilistic" dispersion. Both vary in opposite directions: when the variance is large, the entropy is small, and conversely. We introduce a "corrected entropy", which is a simple modification, and now the variations are in the same sense. We give comparison estimates between the corrected entropy and the variance and these estimates are best possible.

We characterize sequences with extremal entropy, given variance, and, using Archimedes Weighing Method, sequences of extremal weight, and given entropy.

Acknowledgments: We thank Michel Bénézit for his contributions, which led to significant simplifications and strengthening of the results.

Let us start with a proper definition:

I. Entropy

Let $p_1,...,p_N$ be a discrete probability law ($p_i \geq 0$, $\sum_{i=1}^{N} p_i = 1$). We define the entropy of this law by the quantity :

$$I = -\sum_{i=1}^{N} p_i Log\left(p_i\right)$$

This quantity is always positive. It measures the "dispersion" of the law. Indeed, if the law is quite concentrated (all $p_i = 0$ except one equal to 1), then $I = 0$. Conversely, if the law has maximal dispersion (all $p_i = 1/N$), we have $I = Log\left(N\right)$, and this value is a maximum for I, as we will see later. So we always have the bounds:

$$0 \leq I \leq Log\left(N\right)$$

Remark

One may also define the continuous entropy: from a probability density f , this entropy is defined by the formula:

$$I_c = \int Log\left(f(x)\right) f\left(x\right) dx$$

The continuous entropy has similar properties, but does not reduce to the entropy defined above if the density is discretized. The link between both is studied in the book [PIT].

II. Information transfer

It may happen that the numbers p_i are of the form $p_i = \frac{n_i}{N}$, where n_i is an integer. This is the case for instance if we work on a finite universe (set of balls, of cards, and so on). In this case:

$$I = -\sum_{i=1}^{N} p_i Log\left(\frac{n_i}{N}\right) = -\sum_{i=1}^{N} p_i Log\left(n_i\right) + Log\left(N\right)$$

But the quantity :

$$I_2 = \sum_{i=1}^{N} p_i Log_2\left(n_i\right)$$

is the average (since we use the coefficients p_i) of the numbers $Log_2(n_i)$, which represent the number of necessary characters in order to write n_i in base 2. We have :

$$I_2 = \frac{I - Log(N)}{Log(2)}$$

So we may consider I, after substracting $Log(N)$ and dividing by $Log(2)$, as an average of the number of characters needed in order to write each n_i.

Let us now study the examples given in Wikipedia's article "Entropie de Shannon" (in French). They are very poorly presented.

Assume first that we have a box, containing four types of balls : red, blue, yellow, green. We do not know how many balls there are, nor in which proportion. We simply know that only four types appear. We draw a ball out of the box, and want to send an information about its colour to someone else.

Then, the simplest procedure is to build two binary counters ("binary" means that each counter indicates 0 or 1) and define for instance :

0,0 = red ; 0,1 = blue ; 1,0 = yellow ; 1,1 = green.

So, what do we send to our correspondent ? First, the above line, which is a description of the experiment, and then, each time a ball is drawn, a couple (x,y) made of 0 and 1 only. From this, he will deduce the color of the ball.

What is the entropy ? We do not know, because we do not know what the probability law will be, for the colour drawn. But, since we know nothing, we may accept the fact that this probability law will be the uniform law, meaning that all colours have equal probability. It will be defined by $P(x,y) = 1/4$, for all x,y.

The entropy associated to this probability law is:

$$I_0 = -Log\left(\frac{1}{4}\right) = 2Log(2)$$

and this entropy will be extreme (that is, maximal), among all situations of the following type : we have four types of balls, and we want to send the colour of the ball we extract.

Now, let us take the second example mentioned by Wikipedia : the set of balls contains twice as many red balls than blue balls, and twice as many blue than yellow or green, the proportion of yellow and green being the same ($N_R = 2N_B$, $N_B = 2N_Y = 2N_G$).

Then we will need three binary counters, and we can define:

(Def) 0 = red ; 10 = blue ; 110 = yellow ; 111 = green.

The entropy associated to this definition is:

$$I_1 = \frac{1}{2} Log(2) + \frac{1}{4} Log(4) + \frac{1}{8} Log(8) + \frac{1}{8} Log(8) = \frac{7}{4} Log(2)$$

which is less than previously. This is completely in accordance to the theory, because now we have a specific probability law on the colour which is drawn.

However, this is valid only if the line (Def) has been sent to the receiver ; he should understand that he might receive "0" only (and not 000, which he might expect), and if he receives the one-digit 0, he must interpret that as red ; the same, he might receive two digits only, namely 10, and should interpret that as blue. Also, he should know that signals such as 001 are impossible and should be treated as mistakes.

Finally, we observe that, perhaps the entropy is lower, but we need in this second case three binary counters instead of two in the previous case. Of course, the two-digits counter works perfectly well in the second case. So, the "economy" is in the total number of bites sent, but this economy is compensated (which is never clearly said) by a more complex device used in order to store the information (three counters instead of two) and a more complex definition sentence, allowing many possible impossibilities or mistakes.

In real life situations, and this contradicts completely what is said by Brillouin and by Wikipedia, one never has a complete knowledge of the initial probability law. For instance, if a message is transmitted, the transmission line should be able to send messages in English, but also in other languages. It should be also able to send images, files, and so on. Therefore, the usual discussion about the links between entropy and information have very little practical contents.

III. Quantitative study of the entropy

As we saw, the concept of entropy is associated to any sequence $p_1, \dots, p_N$ of positive real numbers with sum equal to 1. We are now going to investigate some basic properties.

A. General extrema

Quite clearly, from the definition, $I \geq 0$ and this value is attained for any sequence where one of the p_i is equal to 1, the others to 0. We are now going to investigate the maximum of I.

Proposition 1. – *The maximum of I is attained for the constant sequence $p_i = \frac{1}{N}$ for all i, and the value of this maximum is $Log(N)$.*

Proof of Proposition 1

We define the Lagrange multipliers:

$$F = -\sum_{i=1}^{N} p_i Log(p_i) + \lambda\left(1 - \sum_{i=1}^{N} p_i\right)$$

Then:

$$\frac{\partial F}{\partial p_i} = -Log(p_i) - 1 - \lambda$$

The extrema are either solutions of the system $\frac{\partial F}{\partial p_i} = 0$ or on the boundaries of the definition domain.

The solution of the system $\frac{\partial F}{\partial p_i} = 0$ satisfies $Log(p_i) = -1 - \lambda$, which means that p_i is independent of i. The condition $\sum_{i=1}^{N} p_i = 1$ then implies that $p_i = \frac{1}{N}$ for all i. The value of the entropy is $I = Log(N)$.

Since the function $-pLog(p)$ is concave (see the graph below), the solutions of the system above are maxima, and the points at the boundary are minima. So, with no additional constraint, the absolute maximum of the entropy is obtained by the only sequence $p_i = \frac{1}{N}$ for all i. This proves Proposition 1.

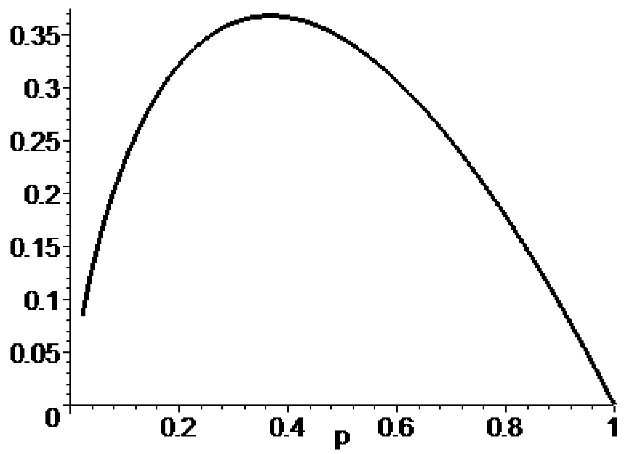


Figure 1 : Graph of the function $f(p) = -pLog(p)$

Remark. – Proposition 1 follows also immediately from Gibbs Inequality :
If $q_1, ..., q_N$ is a sequence of positive numbers, with sum 1,

$$-\sum_{i=1}^{N} p_i Log(p_i) \leq -\sum_{i=1}^{N} p_i Log(q_i)$$

Indeed, just take $q_i = \frac{1}{N}$ for all i.

B. Extrema with constraints

We now look at the extrema of $I = -\sum_{i=1}^{N} p_i Log(p_i)$, assuming $\sum_{i=1}^{N} p_i = s$ and the additional constraints $p_i \geq \delta$ or $p_i \leq \delta$.

We observe that, in all cases, we may assume $s = 1$. Indeed:

$$-\sum_{i=1}^{N} \frac{p_i}{s} Log\left(\frac{p_i}{s}\right) = -\frac{1}{s}\sum_{i=1}^{N} p_i Log(p_i) + Log(s)$$

By the above equality, if we set $q_i = \frac{p_i}{s}$, we reduce ourselves to a problem with $\sum_{i=1}^{N} q_i = 1$. So we always assume $s = 1$ in the sequel.

1. Looking for a maximum of I

Assume first we have the additional constraint $p_i \leq \delta$. This implies:

$N\delta \geq \sum_{i=1}^{N} p_i = 1$, that is $\delta \geq \frac{1}{N}$

But then the constant sequence $p_i = \frac{1}{N}$ for all i automatically satisfies the constraint. We see, in this case, that the extra constraint $p_i \leq \delta$ is either impossible (if $\delta < \frac{1}{N}$) or vacuous (if $\delta \geq \frac{1}{N}$).

Assume now that we have the additional constraint $p_i \geq \delta$; the same applies: the solution will be the constant sequence $\frac{1}{N}$ if $\frac{1}{N} \geq \delta$ or there will be no solution otherwise.

2. Looking for a minimum of I

Assume first that we have the additional constraint $p_i \leq \delta$. This implies as above:

$N\delta \geq \sum_{i=1}^{N} p_i = 1$, that is $\delta \geq \frac{1}{N}$.

The boundaries of the domain are now sequences made only of 0 and $\delta,$ except perhaps for one term, in order to satisfy the condition $\sum_{i=1}^{N} p_i = 1$. More precisely, let $k = \left[\frac{1}{\delta}\right]$. The minimal sequence will have k terms equal to $\delta,$ one term equal to $1-k\delta$ and the rest equal to 0.

Let us give an example : $N = 100,\ \delta = \frac{2}{17}$, so $k = 8.$ The sequence giving minimal entropy will be : $\frac{2}{17}$ repeated 8 times, $\frac{1}{17}$, and 0 repeated 91 times.

Assume now that we have the additional constraint $p_i \geq \delta.$ This implies:

$N\delta \leq \sum_{i=1}^{N} p_i = 1,$ that is $\delta \leq \frac{1}{N}.$ The boundaries of the domain are sequences made only of 1 and $\delta,$ except perhaps for one term, in order to satisfy the condition $\sum_{i=1}^{N} p_i = 1$.

So the minimal sequence will be δ repeated $N-1$ times and the last term will be $1-(N-1)\delta$. Indeed, $1-(N-1)\delta \geq \delta$ since $\delta \leq \frac{1}{N}$.

C. Entropy and variance

For a given sequence p_i, $i = 1,...,N$ satisfying $\sum_{i=1}^{N} p_i = 1,$ we can define the variance of the sequence by the formula :

$$V = \mathrm{var}(p_i) = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - \frac{1}{N}\right)^2$$

This definition is consistent with the usual definition ; here the average of the p_i's is $\frac{1}{N}$. The variance is also a way to measure the concentration of the sequence, but of completely different nature. We insist that this variance is connected to a sequence, not to a random variable.

The minimum value of the variance is 0 , it is attained when $p_i = \frac{1}{N}$ for all $i,$ and we saw that in this case the entropy is maximal.

On the other hand,

$$V = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - \frac{1}{N}\right)^2 = \frac{1}{N}\left(\sum p_i^2 - \frac{1}{N}\right) \leq \frac{1}{N}\left(\sum p_i - \frac{1}{N}\right) = \frac{N-1}{N^2}$$

and the maximum value of the variance is attained when one p_i is equal to 1, all others 0.

In this case, the entropy is minimal.

In fact,

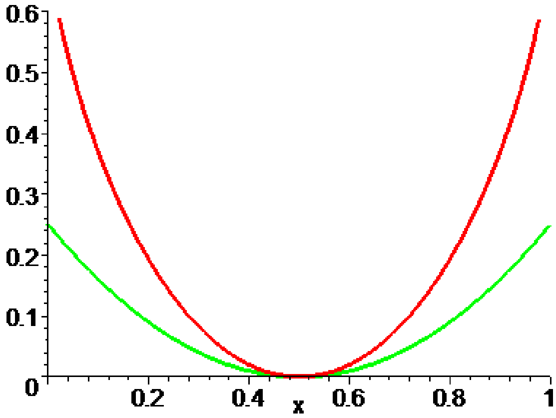
- The variance measures the "geometric" dispersion of the values (considered for instance as points on a axis) ;
- The entropy measures the "probabilistic" dispersion of the values, that is the concentration of a probability law.

Their behavior is opposite : the variance is minimal when the entropy is maximal, and conversely.

We observe that a better definition of the entropy might be the "corrected entropy", defined by:

$$I_c = \sum_{i=1}^{N} p_i Log\left(Np_i\right)$$

Indeed, $I_c = Log\left(N\right) - I,$ so we have $0 \le I_c \le Log\left(N\right)$; the corrected entropy has the same range of variation. But this time the variation goes in the same sense as the variance: they are both extreme at the same places.



On the picture above, we see both the corrected entropy (in red) and the variance (in green), in the one dimensional case (that is $p_1 = x,\ p_2 = 1 - x.$

We now investigate the links between corrected entropy and variance.

Theorem 1. – *For any sequence* $\left(p_i\right)$ *of length* $N,$ *one has:*

$$I_c \ge \alpha V$$

with:

$$\alpha = \frac{N\left(N-1\right)}{N-2} Log\left(N-1\right)$$

This estimate is best possible.

Proof of Theorem 1

We need a Lemma, which will be fundamental in the sequel.

Lemma 2. – *Let* (p_i) *be a sequence (with* $\sum p_i = 1$*) which realizes the minimum of* $I_c - \alpha V$*. Then this sequence takes at most two different values.*

Proof of Lemma 2

We may of course assume that the $p_i's$ are written in decreasing order, that is :

$$p_1 \geq p_2 \geq \cdots \geq 0$$

Assume that $p_1 > p_2$. Take $\varepsilon > 0$ small enough and consider the sequence:

$$p_1 - \varepsilon, p_2 + \varepsilon, p_3, \ldots, p_N \tag{1}$$

Set :

$$f(\varepsilon) = (p_1 - \varepsilon) Log\left(N(p_1 - \varepsilon)\right) + (p_2 + \varepsilon) Log\left(N(p_2 + \varepsilon)\right) + \sum_3^N p_i Log(Np_i)$$
$$- \frac{\alpha}{N}\left(\left(p_1 - \varepsilon - \frac{1}{N}\right)^2 + \left(p_2 + \varepsilon - \frac{1}{N}\right)^2 + \sum_3^N \left(p_i - \frac{1}{N}\right)^2\right)$$

which represents the quantity $I_c - \alpha V$ evaluated at the sequence (1). Then, by definition of the sequence (p_i), $f(\varepsilon)$ is minimal for $\varepsilon = 0$.

We have :

$$f'(\varepsilon) = Log\frac{p_2 + \varepsilon}{p_1 - \varepsilon} - \frac{2\alpha}{N}(p_2 - p_1 + 2\varepsilon)$$

which leads to the condition:

$$Log\frac{p_2}{p_1} = \frac{2\alpha}{N}(p_2 - p_1) \tag{2}$$

Set $\beta = \dfrac{2\alpha}{N}$ and write $x = p_1 - p_2$. We deduce from (2) :

$$Log\left(1 + \frac{x}{p_2}\right) = \beta x \tag{3}$$

A solution in x to this equation is obtained the following way : one takes the intersection of the curve $y = Log\left(1 + \dfrac{x}{p_2}\right)$ with the straight line $y = \beta x$.

We observe that this intersection exists only if the slope of the tangent at 0 to the first curve is larger than β, which leads to the condition $p_2 < \frac{1}{\beta}$.

Assume this condition on p_2 to be satisfied ; then, for given p_2, equation (3) has one and only one solution in x.

Assume now that the minimal sequence (p_i) contains at least 3 different non-zero terms:

$$p_1 > p_2 > p_3 \geq p_4 \geq \cdots \geq p_N$$

Apply the above reasoning to the couples (p_1, p_3) and (p_2, p_3) : the smallest one is the same in both cases, so we deduce that the differences $p_1 - p_3$ and $p_2 - p_3$ must be the same. This shows that $p_1 = p_2$ and proves Lemma 2.

We now turn to the proof of the Theorem. We have :

$$I_c = \sum_i p_i Log(Np_i)$$

Assume, using Lemma 2, that n of the p_i's take the value p and $N-n$ take the value q. Then $np + (N-n)q = 1$, that is $q = \frac{1-np}{N-n}$ and $0 \leq p \leq \frac{1}{n}$.

Then, in this case:

$$I_c = npLog(Np) + (1-np)Log\left(N\left(\frac{1-np}{N-n}\right)\right)$$

and:

$$V = \mathrm{var}(p_i) = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - \frac{1}{N}\right)^2 = \frac{n}{N-n}\left(p - \frac{1}{N}\right)^2$$

The proof of the Theorem will be decomposed into several parts. A first remark is that the quotient $\frac{I_c}{V}$ is quite hard to study directly. Therefore, we find (by empirical means) what the lower bound α of the difference $I_c - \alpha V$ may be, and then we prove that this guess is correct.

Part I : a guess for the bound

We first investigate the case $n = 1$, which will allow us to obtain a "guess" for the value of α. With the above defintions for I_c and V, we set:

$$D = I_c - \alpha V$$

and we will compute α so that $\frac{\partial D}{\partial p}$ vanishes for $p=1-\frac{1}{N}$.

We have :

$$D'\left(1-\frac{1}{N}\right)=2Log\left(N-1\right)-\frac{2\alpha}{N-1}\left(1-\frac{2}{N}\right)$$

and so the equation $D'\left(1-\frac{1}{N}\right)=0$ is equivalent to:

$$\alpha=\frac{N\left(N-1\right)}{N-2}Log\left(N-1\right) \tag{1}$$

We now check that $D=0$ for $p=1-\frac{1}{N}$:

$$\begin{aligned} D\left(1-\frac{1}{N}\right)&=\left(1-\frac{1}{N}\right)Log\left(N-1\right)-\frac{1}{N}Log\left(N-1\right)-\frac{N}{N-2}Log\left(N-1\right)\left(1-\frac{2}{N}\right)^2 \\ &=Log\left(N-1\right)\left(\frac{N-1}{N}-\frac{1}{N}-\frac{N-2}{N}\right)=0 \end{aligned}$$

and we will see later that this leads to a minimum for D. If this is proved, then $I_c \geq \alpha V$, where α is given by (1).

Part II : Proof of the Theorem

Now, α has been chosen. We want to prove that for all n and p, we have $D \geq 0$. This will be done in several steps.

II. 1. – A change of variables

We will make a change of variables, as follows:

We set $t=\frac{n}{N}$ and $x=np$, so $n=tN$, $p=\frac{x}{tN}$, and we have the intervals of variation:

$$\frac{1}{N}\leq t\leq\frac{N-1}{N},\ 0\leq x\leq 1 \tag{2}$$

The value $x=0$ corresponds to $p=0$; since $np+\left(N-n\right)q=1$, this means $q=\frac{1}{N-n}$. So we have 0 repeated n times and $\frac{1}{N-n}$ repeated $N-n$ times.

The value $x=1$ corresponds to $np=1$, so $q=0$. So we have $\frac{1}{n}$ repeated n times and 0 repeated $N-n$ times. So we observe that the extreme values $x=0$ and $x=1$ are possible.

In order to prove the Theorem, we need to show that, for all x and t:

$$x\,Log\frac{x}{t}+(1-x)\,Log\frac{1-x}{1-t}-\frac{(N-1)\,Log(N-1)}{(N-2)\,N\,t(1-t)}(x-t)^2\geq 0$$

We set:

$$\beta=\frac{(N-1)\,Log(N-1)}{(N-2)\,N}\ .$$

and:

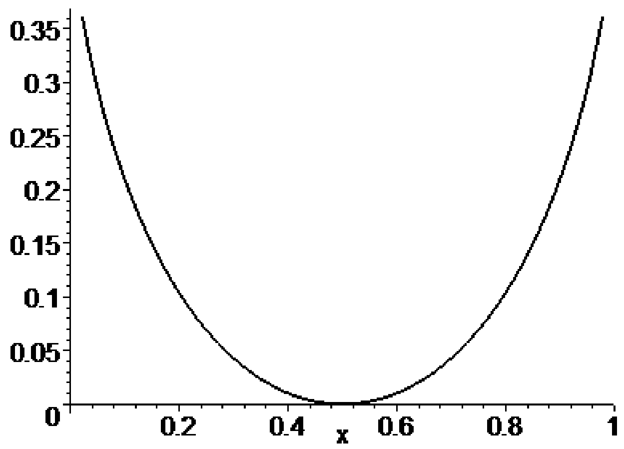
$$y=x\,Log\frac{x}{t}+(1-x)\,Log\frac{1-x}{1-t}-\beta\frac{(x-t)^2}{t(1-t)}$$

considered as a function of x. We have:

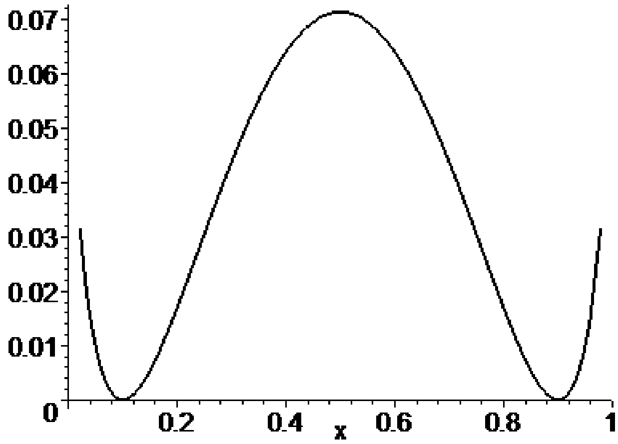
$$y'=Log\frac{x}{t}-Log\frac{1-x}{1-t}-2\beta\frac{x-t}{t(1-t)}$$

$$y''=\frac{1}{x(1-x)}-\frac{2\beta}{t(1-t)}$$

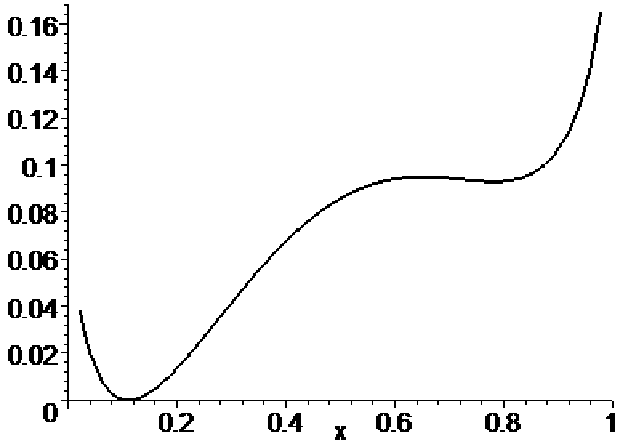
We first observe that if $x=t$, then $y=0$ and $y'=0$. We will see that the point $x=t$ is always a minimum for y; however, depending on the value of t, it may have another local minimum. In fact, three different shapes will be observed, as the following graphs show:



Graph of y, $N=10$, $t=\frac{1}{2}$



Graph of y, $N = 10$, $t = \frac{1}{10}$



Graph of y, $N = 10$, $t = \frac{1}{9}$

II.2. - Study of the sign of y''

Lemma 3.- Let t_N be the unique solution $< \frac{1}{2}$ of the equation:

$$t(1-t) = \frac{(N-1)Log(N-1)}{2N(N-2)} \tag{3}$$

then we have:

$$\frac{1}{N} < t_N \tag{4}$$

and, if $t_N \leq t \leq 1 - t_N$, we have $y'' > 0$ for all x.

Proof of Lemma 3

Let us first prove that $\frac{1}{N} < t_N$. Indeed, since t_N is solution of the equation (3), all we have to show is that:

$$\frac{1}{N}\left(1-\frac{1}{N}\right) < \frac{(N-1)Log(N-1)}{2N(N-2)}$$

which is equivalent to:

$$\frac{N-2}{N} < \frac{Log(N-1)}{2}$$

which is true for $N \geq 3$ and proves our claim.

Let us now prove the second statement. The condition $y'' > 0$ is equivalent to:

$$x(1-x) < \frac{t(1-t)}{2\beta} \tag{5}$$

The maximum value of $x(1-x)$ is $1/4$. Therefore, (5) will be satisfied for all x if $\frac{t(1-t)}{2\beta} \geq \frac{1}{4}$, or:

$$t(1-t) \geq \frac{(N-1)Log(N-1)}{2N(N-2)}$$

This is equivalent to $t_N \leq t \leq 1-t_N$, where t_N is the unique solution $< \frac{1}{2}$ of the equation(3) and our Lemma is proved.

A precise value of t_N is given by the expression:

$$t_N = \frac{1}{2} - \frac{1}{2}\sqrt{1-2\beta}$$

II. 3. – Proof of the Theorem in the case $t_N \leq t \leq 1-t_N$

In this case, we saw that $y'' \geq 0$ for all x. So y' is increasing. But:

$y' \to -\infty$ when $x \to 0$

$y' = 0$ when $x = t$

$y' \to +\infty$ when $x \to 1$

Therefore, the unique solution of $y'=0$ is obtained for $x=t$; $y'<0$ if $x<t$ and $y'>0$ if $x>t$. The minimum of y is obtained for $x=t$ and this minimum is 0. So $y\geq 0$ for all x and the Theorem is proved in this case. This is a simple case where the function y has only one minimum (namely $x=t$).

We now assume $\frac{1}{N}\leq t\leq t_N$; the discussion would be the same in the case $1-t_N\leq t\leq 1-\frac{1}{N}$, since all quantities are invariant under the transformations $t\rightarrow 1-t$ and $x\rightarrow 1-x$. But first we have to study the extreme case $t=\frac{1}{N}$.

II.4. – Study of the case $t=\frac{1}{N}$

We have in this case:

$$y=x\,Log\left(Nx\right)+\left(1-x\right)Log\left(\frac{N\left(1-x\right)}{N-1}\right)-\frac{NLog\left(N-1\right)}{N-2}\left(x-\frac{1}{N}\right)^2$$

Lemma 4. – *The function y vanishes at the points $x=\frac{1}{N}$ and $x=1-\frac{1}{N}$.*

Proof of Lemma 4

For $x=\frac{1}{N}$, this is obvious. Let us take $x=1-\frac{1}{N}$. Then:

$$\begin{aligned}
y&=\left(1-\frac{1}{N}\right)Log\left(N\left(1-\frac{1}{N}\right)\right)+\frac{1}{N}\,Log\,\frac{1}{N-1}+-\frac{NLog\left(N-1\right)}{N-2}\left(1-\frac{2}{N}\right)^2\\
&=\frac{N-1}{N}Log\left(N-1\right)-\frac{1}{N}Log\left(N-1\right)-\frac{N-2}{N}Log\left(N-1\right)\\
&=Log\left(N-1\right)\left(\frac{N-1}{N}-\frac{1}{N}-\frac{N-2}{N}\right)=0
\end{aligned}$$

which proves our claim.

Let us study y' when $t=\frac{1}{N}$. We have :

$$y'=Log\left(Nx\right)-Log\left(\frac{N\left(1-x\right)}{N-1}\right)-\frac{2NLog\left(N-1\right)}{N-2}\left(x-\frac{1}{N}\right)$$

Lemma 5. -*The function y' takes the value 0 at the points $x=\frac{1}{N}$, $x=\frac{1}{2}$ and $x=1-\frac{1}{N}$.*

Proof of Lemma 5

For $\frac{1}{N}$, $1-\frac{1}{N}$, the computation is the same as above. For $x=\frac{1}{2}$:

$$y' = Log\frac{N/2}{N-1} - Log(N/2) - 2\frac{NLog(N-1)}{N-2}\left(-\frac{1}{2}+\frac{1}{N}\right) = 0$$

which proves Lemma 5.

We now study the variations of y'.

We know that $y'' \geq 0$ if and only if $x(1-x) \leq \frac{t(1-t)}{2\beta} = \frac{\frac{1}{N}\left(1-\frac{1}{N}\right)}{2\beta}$

This happens if and only if $x \leq x_N$ or $x \geq 1-x_N$, where x_N is the unique solution $<\frac{1}{2}$ of the equation :

$$x_N(1-x_N) = \frac{\frac{1}{N}\left(1-\frac{1}{N}\right)}{2\beta} \tag{6}$$

Since:

$$\frac{1}{N}\left(1-\frac{1}{N}\right) < \frac{\frac{1}{N}\left(1-\frac{1}{N}\right)}{2\beta}$$

we see that $x_N > \frac{1}{N}$. So we have the ordering:

$$0 < \frac{1}{N} < x_N < \frac{1}{2} < 1-x_N < 1-\frac{1}{N} < 1 \tag{7}$$

The function y' is increasing between 0 and x_N, decreasing between x_N and $1-x_N$, increasing between $1-x_N$ and 1.

Using Lemma 5, we see that the sign of y', is as follows:

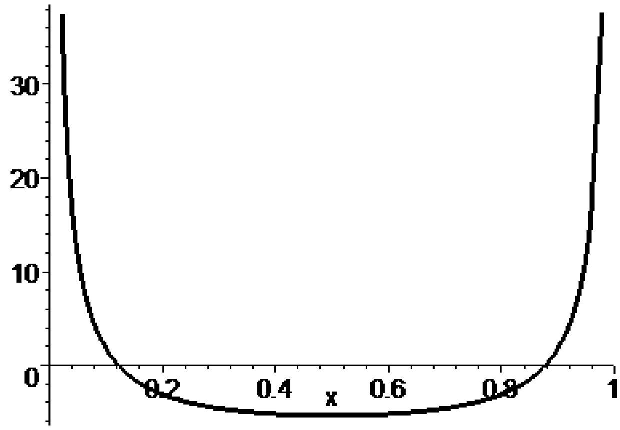
Between 0 and $\frac{1}{N}$: $y' < 0$

Between $\frac{1}{N}$ and $\frac{1}{2}$: $y' > 0$

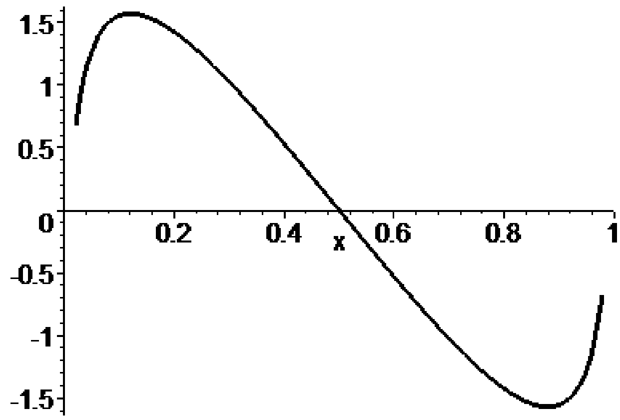
Between $\frac{1}{2}$ and $1-\frac{1}{N}$: $y' < 0$

Between $1-\frac{1}{N}$ and 1 : $y'>0$

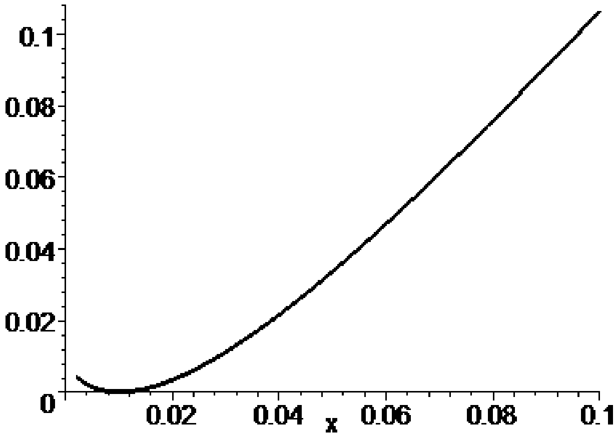
Therefore, y is decreasing between 0 and $\frac{1}{N}$, increasing between $\frac{1}{N}$ and $\frac{1}{2}$, decreasing between $\frac{1}{2}$ and $1-\frac{1}{N}$ and increasing between $1-\frac{1}{N}$ and 1. So the minimum is reached at the points $\frac{1}{N}$ and $1-\frac{1}{N}$ and the value of this minimum is 0. This proves that $y \geq 0$ for all x, when $t=1-\frac{1}{N}$.



Graph of y'', *case* $N=100$



Graph of y', *case* $N=100$

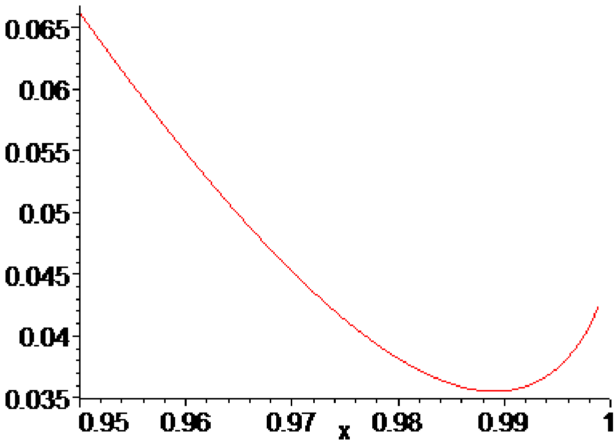


Graph of y *in a neighborhood of* $\frac{1}{N}$*, case* $N = 100$.

This finishes the proof of the Theorem in the case $t = 1 - \frac{1}{N}$.

II.5. – Case $\frac{1}{N} < t \le t_N$.

We observe that there is some difference with the case $t = \frac{1}{N}$. There will be a second local minimum (besides $x = t$), but the value at this second minimum will not be 0. For instance, let us draw the graph when $t = \frac{1}{N-1}$:



Graph of y*, case* $N = 100$

The function y does not vanish between $1 - \frac{1}{N}$ and 1, but still has a minimum in this interval.

We saw that the condition $y'' > 0$ is equivalent to $x(1-x) < \frac{t(1-t)}{2\beta}$ and we are in the case $\frac{t(1-t)}{2\beta} < \frac{1}{4}$. Therefore, $y'' > 0$ if $x < x_t$ and if $x > 1 - x_t$, where x_t is the unique solution $< 1/2$ of the equation :

$$x(1-x)=\frac{t(1-t)}{2\beta} \quad (8)$$

We observe that $t<x_t$. Indeed, x_t is solution of (8), whereas $t(1-t)<\frac{t(1-t)}{2\beta}$, since $2\beta<1$.

So we have the disposition:

$$0<t<x_t<\frac{1}{2}. \quad (9)$$

The proof will follow different patterns, depending on the position of x in this disposition.

We have $y''>0$ if $0\le x\le x_t$, $y''=0$ for $x=x_t$, and $y''<0$ if $x_t\le x\le 1-x_t$.

Therefore, y' is increasing if $0\le x\le x_t$, decreasing if $x_t\le x\le 1-x_t$ and increasing if $1-x_t\le x\le 1$.

At the point 0, y' has limit $-\infty$ and y' vanishes at $x=t$; therefore $y'<0$ if $0<x<t$. So y is decreasing on this interval.

Also, y' is increasing between t and x_t and therefore is positive, since y' vanishes at $x=t$. Therefore, y is increasing between t and x_t.

Case II.5.1. - $0\le x\le x_t$

So, if $0\le x\le x_t$, things are clear: the function y is first decreasing, then increasing ; it reaches its minimum at $x=t$ and this minimum is 0 ; so we have proved that $y\ge 0$ on this interval.

Case II.5.2. - $x_t\le x\le \frac{1}{2}$

On the interval $x_t\le x\le 1$, y' is first decreasing (if $x_t\le x\le 1-x_t$), reaches its minimum at $1-x_t$, then is increasing if $1-x_t\le x\le 1$.

We know that $y'(x_t)>0$ and that the limit of y' at $+\infty$ is $+\infty$, but this does not allow us to reach any conclusion on the sign of y' : if $y'(1-x_t)>0$, y' does not vanish, so it is always >0, y is always increasing and the problem is solved. However, if $y'(1-x_t)<0$, y' will vanish twice, will be positive first, then negative, then positive ; y will be first increasing, then decreasing, then increasing, and there will be a local minimum of y between x_t and 1.

We will first show that $y'>0$ when $x=\frac{1}{2}$.

We have:

$$y'(1/2) = Log\left(\frac{1}{2t}\right) - Log\left(\frac{1}{2(1-t)}\right) - 2\beta\frac{\frac{1}{2}-t}{t(1-t)} = Log\frac{1-t}{t} - \beta\frac{1-2t}{t(1-t)}$$

with $\frac{1}{N} < t < t_N$.

We set:

$$A = Log\frac{1-t}{t} - \beta\frac{1-2t}{t(1-t)}$$

and we have:

$$A' = -\frac{1}{t(1-t)} - \beta\frac{-2t^2+2t-1}{t^2(1-t)^2} = \frac{-t(1-t) - \beta\left(-2t^2+2t-1\right)}{t^2(1-t)^2}$$

which has the same sign as:

$$B = -t(1-t) - \beta\left(-2t^2+2t-1\right) = t^2(1+2\beta) - t(1+2\beta) + \beta$$

This is a quadratic function, which reaches its absolute minimum at $t = \frac{1}{2}$ and so it is decreasing for $t < \frac{1}{2}$. Since $t < t_N$, the minimum is obtained for $t = t_N$; its value is:

$$C = t_N^2(1+2\beta) - t_N(1+2\beta) + \beta = (1+2\beta)\left(t_N^2 - t_N\right) + \beta$$

But, by definition, $t_N(1-t_N) = \frac{\beta}{2}$. So:

$$C = -(1+2\beta)\frac{\beta}{2} + \beta = -\beta^2 + \frac{\beta}{2} = \beta\left(\frac{1}{2} - \beta\right) > 0$$

This shows that $B > 0$, so that $A' > 0$ for all t; therefore A is increasing. The smallest value is obtained at $t = \frac{1}{N}$:

$$A = Log(N-1) - \beta\frac{N(N-2)}{N-1} = 0$$

which proves that $A \geq 0$ if $\frac{1}{N} \leq t \leq t_N$, when $x = \frac{1}{2}$

We will now show that $y' > 0$ on the interval $t < x < 1/2$. Let x be in this interval. We have:

$$y' = Log\frac{x}{1-x}\frac{1-t}{t} - 2\beta\frac{x-t}{t(1-t)}$$

We set:

$$A = Log\frac{x}{1-x}\frac{1-t}{t} - 2\beta\frac{x-t}{t(1-t)}$$

considered as a function of t.

We have:

$$A' = -\frac{1}{t(1-t)} - 2\beta\frac{-t^2+2xt-x}{t^2(1-t)^2} = \frac{-t(1-t)-2\beta(-t^2+2xt-x)}{t^2(1-t)^2}$$

which has same sign as:

$$B = -t(1-t)-2\beta(-t^2+2xt-x) = t^2(1+2\beta)-t(1+4\beta x)+2\beta x$$

But, considered as a function of x, this last quantity is decreasing, therefore takes its minimum for $x = 1/2$; we are back to the previous case, and we have shown that $B > 0$, so that $A' > 0$, and therefore that A is increasing as a function of t, for fixed x.

Since $A = 0$ if $t = x$, we have $A > 0$ if $t < x \leq \frac{1}{2}$, which shows that $y' > 0$ on this interval. This implies that, as a function of x, y is increasing if $t < x < \frac{1}{2}$. Since $y = 0$ if $x = t$, we have $y > 0$ if $t < x < \frac{1}{2}$ and the Theorem is proved on the interval $0 < x < \frac{1}{2}$.

Case II.5.3. - $\frac{1}{2} < x \leq 1$

In the quantity:

$$y = xLog\left(\frac{x}{t}\right) + (1-x)Log\left(\frac{1-x}{1-t}\right) - \beta\frac{(x-t)^2}{t(x-t)}$$

we fix x and consider y as a function of t. We have:

$$\frac{\partial y}{\partial t} = \frac{-x}{t} + \frac{1-x}{1-t} - \beta\frac{(x-t)(2xt-t-x)}{t^2(1-t)^2} = \frac{(t-x)t(1-t)-\beta(x-t)(2xt-t-x)}{t^2(1-t)^2}$$

Since $x - t > 0$, this quantity has same sign as:

$$A = -t(1-t)-\beta(x-t)(2xt-t-x) = t^2 - 2\beta tx + \beta t - t + \beta x$$

We will show that $A \geq 0$ for all x and all t. We consider A as a function of x. We have:

$$\frac{\partial A}{\partial x} = -2\beta t + \beta = \beta(1-2t) > 0$$

So A is an increasing function of x and the minimum is obtained for x minimum, which is $x = \frac{1}{2}$. For this value, we have:

$$A_{\min} = t^2 - t + \frac{\beta}{2}$$

and:

$$\frac{\partial A_{\min}}{\partial t} = 2t - 1 < 0$$

So $A_{\min}$ is a decreasing function of t and the minimum of $A_{\min}$ is obtained for t maximal, that is for $t = t_N$. But t_N is defined by the equation $t_N(1-t_N) = \frac{\beta}{2}$ and therefore $A_{\min} = 0$, which proves that $A \geq 0$.

This implies that y is an increasing function of t, so the minimum of y is obtained for t as small as possible, that is for $t = t_N$. But, for this value of t, we know that $y \geq 0$ for all x and the Theorem is proved.

The fact that the estimate in Theorem 1 is best possible follows from the proof. Indeed, α was chosen by reference to the case $n = 1$, $p = 1 - \frac{1}{N}$, which corresponds to the sequence :

$$p_1 = 1 - \frac{1}{N}, p_2 = \ldots = p_N = \frac{1}{N(N-1)}$$

Direct computations show that, in this case:

$$I_c = \frac{N-2}{N} Log(N-1)$$

$$V = \frac{1}{N-1}\left(\frac{N-2}{N}\right)^2$$

and:

$$\frac{I_c}{V} = \alpha.$$

We now turn to converse estimates, relating I_c and the variance. We have:

Theorem 6. – *For any sequence* (p_i) *of length* N, *we have:*

$$I_c \le N^2 V$$

and this estimate is best possible.

Proof of Theorem 6

We know that, for any $x > 0$,

$$Log(x) \le x - 1$$

Therefore:

$$Log(Np_i) \le Np_i - 1 = N\left(p_i - \frac{1}{N}\right)$$

$$\begin{aligned}
\sum p_i Log(Np_i) &\le N \sum p_i \left(p_i - \frac{1}{N}\right) \\
&= N \sum \left(p_i - \frac{1}{N} + \frac{1}{N}\right)\left(p_i - \frac{1}{N}\right) \\
&= N \sum \left(p_i - \frac{1}{N}\right)^2 + \sum \left(p_i - \frac{1}{N}\right) \\
&= N \sum \left(p_i - \frac{1}{N}\right)^2
\end{aligned}$$

which proves the Theorem.

We observe that the constant N^2 in Theorem 6 is best possible, as the following example shows :

$n = 1$ $p_1 = \frac{1}{N^2}$, $p_2 = ... = p_N = \frac{1}{N} + \frac{1}{N^2}$. Then:

$$\frac{I_c}{V} \sim N^2 \quad \text{when } N \to +\infty.$$

We deduce from the previous Theorems :

Corollary 7. – *For the entropy* I, *we have the following bounds, which are best possible:*

$$Log(N) - N^2 V \le I \le Log(N) - \frac{N(N-1)Log(N-1)}{N-2} V$$

The following Corollary was communicated to us by Michel Bénézit :

Corollary 8. – *We have, for any sequence of length* N *:*

$$Log(N) \leq I + N^2 V \leq (N-1)(1+\varepsilon_N)$$

with :

$$\varepsilon_N = \frac{Log(N)}{N-1} - \frac{(N-1)Log(N-1)}{N(N-2)} \leq \frac{1}{N^2}$$

Proof of Corollary 8

The left hand-side inequality is clear. For the right one, we have, from Corollary 7:

$$I + N^2 V \leq Log(N) + \left(N^2 - \frac{N(N-1)Log(N-1)}{N-2} \right) V$$

and we use the upper estimate seen previously :

$$V \leq \frac{N-1}{N^2}$$

The result follows.

D. Sequences with fixed variance

We now return to the usual definition of entropy and investigate other quantitative connections with the variance. Let us fix $V = \mathrm{var}(p_i) = \delta$ for some $\delta > 0.$ We want to investigate the sequences satisfying this property, with largest entropy.

We now give a characterization of the extremal sequences. Since they take only two values, let us set :

$p_i = p,\ \ i = 1,\ldots,n$

$p_i = q,\ \ i = n+1,\ldots,N$

for some $n,\ \ 2 \leq n \leq N-1$. We may of course assume $p < q.$

Proposition 1. – *For the extremal sequences, the numbers* p, q *must satisfy:*

$$p = \frac{1}{N} - \sqrt{\frac{\delta(N-n)}{n}},\ q = \frac{1}{N} + \sqrt{\frac{\delta n}{N-n}}$$

Proof of Proposition 1

We have:

$$np + (N-n)q = 1$$

which gives:

$$q = \frac{1-np}{N-n} \tag{1}$$

By the definition of the variance, we have :

$$\frac{n}{N}\left(p - \frac{1}{N}\right)^2 + \frac{N-n}{N}\left(q - \frac{1}{N}\right)^2 = \delta \tag{2}$$

Replacing the value of q given by (1), we get:

$$\frac{n}{N}\left(p - \frac{1}{N}\right)^2 + \frac{N-n}{N}\left(\frac{1-np}{N-n} - \frac{1}{N}\right)^2 = \delta$$

which gives after simplification:

$$\left(p - \frac{1}{N}\right)^2 = \frac{\delta(N-n)}{n}$$

and the value of q follows from (1). This proves Proposition 1.

We observe that the value δ assigned to the variance is bound by some conditions. Indeed, we must have $p \geq 0$, which means:

$$\sqrt{\frac{\delta(N-n)}{n}} \leq \frac{1}{N}$$

or:

$$\delta \leq \frac{n}{N-n}\frac{1}{N^2}$$

This must hold for any $n = 1,\ldots,N-1$; the quantity $\frac{n}{N-n}$ is increasing with n, so the minimum is reached for $n = 1$. This gives the estimate:

$$\delta \leq \frac{1}{N^2(N-1)}$$

Finally, since an extremal sequence is made of n times p and $N-n$ times q, its entropy is:

$$I = npLog(p) + (N-n)qLog(q)$$

We may consider separately both terms, and set:

$$I_1 = npLog(p) = n\left(\frac{1}{N} - \sqrt{\frac{\delta(N-n)}{n}}\right)Log\left(\frac{1}{N} - \sqrt{\frac{\delta(N-n)}{n}}\right)$$

and:

$$I_2 = (N-n)qLog(q) = (N-n)\left(\frac{1}{N} + \sqrt{\frac{\delta n}{N-n}}\right)Log\left(\frac{1}{N} + \sqrt{\frac{\delta n}{N-n}}\right)$$

Then I_1 is obviously decreasing with k and I_2 is obviously increasing. The global behavior of $I = I_1 + I_2$ is not clear. On the numerical examples we treated, it was almost constant, decreasing slightly. This would mean that the sequences with lowest entropy might be of the type $(p,\dots,p,q)$ and the sequences with highest entropy might be of the type $(p,q,\dots,q)$.

E. Sequences with low entropy

We saw that, in general:

$$0 \le I \le Log(N)$$

We are going to investigate the sequences (p_i) such that :

$$I \le \alpha Log(N) \qquad (1)$$

for a given $\alpha < 1$. Such a sequence will be called "of $\alpha -$low entropy".

First, we observe that condition (1) can be written in an equivalent form:

$$-\sum_{i=1}^{N} p_i Log(p_i) \le \alpha Log(N) = Log(N^\alpha) = \sum_{i=1}^{N} p_i Log(N^\alpha)$$

that is:

$$\sum_{i=1}^{N} p_i Log(N^\alpha p_i) \ge 0 \qquad (2)$$

Condition (2) implies:

$$\max_i Log(N^\alpha p_i) \ge 0$$

that is:

$$\max_i p_i \ge \frac{1}{N^\alpha} \qquad (3)$$

So we see that condition (3) is necessary for the sequence to have low entropy. It means that one at least among the $p_i's$ must be large. In fact, a condition to have low entropy is precisely that the sequence should not be too concentrated near its average. The following Proposition characterizes this. It was communicated to us by Michel Bénézit, and is a strengthening of our original result:

Proposition 1. (Michel Bénézit) – Let α and δ, $0<\alpha<1$, $0<\delta<1$ be given. Let:

$$g(t)=\frac{(1+t)Log(1+t)-t}{t^2} \text{ for } t>-1$$

and $\gamma=g(N\delta-1)$.

A sequence (p_i) satisfying $p_i \le \delta$ for all i *and* $V \ge \frac{1-\alpha}{\gamma}\frac{Log(N)}{N^2}$ *has* $\alpha-$*low entropy.*

Proof of Proposition 1

First, we observe that the function g is decreasing. Therefore, for all i:

$$\gamma \le g(Np_i-1)=\frac{Np_iLog(Np_i)-Np_i+1}{(Np_i-1)^2}$$

From which we deduce, after simplification:

$$\gamma\sum(Np_i-1)^2 \le NLog(N)-NI$$

that is:

$$\gamma N^2V \le Log(N)-I$$

from which Proposition 1 follows immediately.

The next Proposition studies the reverse implication : a sequence with low entropy must have a large variance.

Proposition 3. – *If a sequence has* $\alpha-$*low entropy, its variance satisfies:*

$$V \ge \frac{(1-\alpha)Log(N)}{N^2}$$

Proof of Proposition 3

We will use the inequality, valid for all $t>-1$

$$(1+t)Log(1+t) \le t+t^2 \quad (1)$$

This inequality follows from $Log(1+t) \le t$, if $t>-1$.

Now, the condition :

$$\frac{1}{N}\sum_{i=1}^{N}\left(1+t_i\right) Log\left(1+t_i\right) \geq \left(1-\alpha\right) Log\left(N\right) \tag{2}$$

implies using (1):

$$\frac{1}{N}\sum_{i=1}^{N}\left(t_i^2+t_i\right) \geq \left(1-\alpha\right) Log\left(N\right)$$

and since $\sum_{i=1}^{N} t_i = 0$, we get:

$$\frac{1}{N}\sum_{i=1}^{N} t_i^2 \geq \left(1-\alpha\right) Log\left(N\right)$$

which is equivalent, by the same computations as before, to:

$$V \geq \frac{\left(1-\alpha\right) Log\left(N\right)}{N^2}$$

which proves Proposition 3.

F. Sequences with fixed entropy

In this section, we study the sequences $\left(p_i\right)$ with fixed entropy : let α $0<\alpha<1$, such that:

$$I = \alpha Log\left(N\right).$$

This means:

$$\sum_{i=1}^{N} p_i Log\left(N^{\alpha} p_i\right) = 0 \tag{1}$$

The arguments which follow are inspired by Archimedes' Weighing Method (see [AMW]).

Condition (1) may be written:

$$-\sum_{p_i<1/N^{\alpha}} p_i Log\left(N^{\alpha} p_i\right) = \sum_{p_i \geq 1/N^{\alpha}} p_i Log\left(N^{\alpha} p_i\right) \tag{2}$$

The quantity $-pLog\left(N^a p\right)$ may be viewed as the moment of a weight $Log\left(N^a p\right)$, put at the position $-p$, with respect to the origin, and similarly for the quantity $pLog\left(N^a p\right)$, put at the place p ; so condition (2) states that these weights must be in equilibrium with respect to the origin.

For this weighing procedure, on the left of O, the graph is $Log\left(-N^{\alpha}x\right)$ and on the right $Log\left(N^{\alpha}x\right)$; the black bar on the left, figure below, is at position $-p$ and its height is $Log\left(N^{\alpha}p\right)$ and the black bar on the right is as position p ; the collection of bars left of the origin should be in equilibrium with the collection of bars right of the origin.

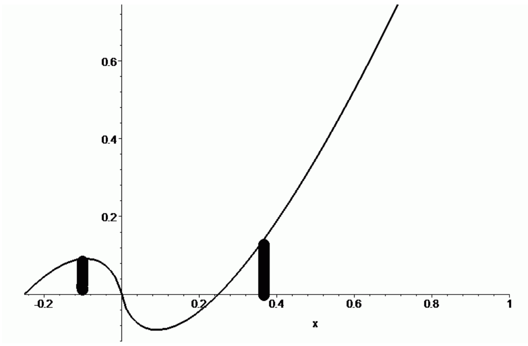


Figure 2: the weighing procedure

In condition (2) both terms are positive. We set:

$$I_1 = -\sum_{p_i<1/N^{\alpha}} p_i Log\left(N^{\alpha}p_i\right)$$

$$s_1 = \sum_{p_i<1/N^{\alpha}} p_i$$

and:

$$I_2 = \sum_{p_i\geq 1/N^{\alpha}} p_i Log\left(N^{\alpha}p_i\right)$$

$$s_2 = \sum_{p_i\geq 1/N^{\alpha}} p_i$$

We want to find the maximum value of I_1 for fixed s_1.

As we already saw, we can solve this problem assuming $s_1 = 1$. The maximum value of I_1 is obtained when all coefficients are equal.

Let k be the number of p_i such that $p_i < \dfrac{1}{N^{\alpha}}$. Then the maximum value of I_1 is reached for:

$$p_1 = \ldots = p_k = \frac{s_1}{k}.$$

We now turn to the term I_2.

$$I_2 = \sum_{p_i \geq 1/N^\alpha} p_i Log\left(N^\alpha p_i\right)$$

which is the opposite of an entropy. The same way, the maximum value of I_2 will be obtained for the minimum value of:

$$I_2' = -\sum_{p_i \geq 1/N^\alpha} p_i Log\left(p_i\right)$$

under the constraints $\sum p_i = s_2$ and $p_i \geq \frac{1}{N^\alpha}$. Then, as we saw earlier, the minimum of the entropy is obtained when we assign the value δ to $N-1$ of the p_i ; the last one satisfies:

$p_N = 1 - \delta\left(N-1\right) \geq \delta$ by (1).

So we see that the maximum value of I_2 is obtained in general for the following configuration: all $p_i = \frac{1}{N^\alpha}$ except perhaps one which is larger.

In short, the sequences with maximal terms I_1 and I_2 are made of 3 terms only: one common value for all $p_i < \frac{1}{N^\alpha}$, one common value, namely $\frac{1}{N^\alpha}$ for all terms $\geq \frac{1}{N^\alpha}$ and possibly one term $> \frac{1}{N^\alpha}$.

References

[AMW] Bernard Beauzamy : Archimedes' Modern Works. SCM SA, ISBN 978-2-9521458-7-9, ISSN 1767-1175. August 2012.

[Brilloin] Léon Brillouin La science et la théorie de l'information, 1959, réédité par les Editions Jacques Gabay, Paris.

[PIT] Olga Zeydina and Bernard Beauzamy : Probabilistic Information Transfer. SCM SA. ISBN: 978-2-9521458-6-2, ISSN : 1767-1175. 208 p., May 2013.