

# Illustrations of the notion of entropy in the two theorems of Claude Shannon in information theory

Joël Le Roux, leroux@essi.fr, April 2002

October 21, 2002

## Contents

<b>1</b>	<b>Abstract</b>	<b>1</b>
<b>2</b>	<b>Introduction</b>	<b>2</b>
<b>3</b>	<b>Illustration of Shannon's first theorem in a simple case</b>	<b>2</b>
3.1	Interpretation of the first theorem . . . . .	3
3.2	The number of probable messages approximated by Stirling formula . . . . .	3
3.2.1	Approximation based on Stirling formula . . . . .	3
3.2.2	Gaussian approximation . . . . .	4
3.2.3	Inequality of Bienaymé Chebyshev: . . . . .	5
3.3	What is the number of messages having exactly or more or less $pL$ ones and $(1-p)L$ zeros ? . . . . .	6
3.3.1	Approximation using Stirling formula . . . . .	6
3.3.2	The number of bits given by the first theorem is sufficient to code all messages for which the probability of 'ones' is less than $p$ . . . . .	7
3.4	The principle of compression techniques . . . . .	9
3.4.1	Approximation of the entropy for small values of $p$ . . . . .	9
<b>4</b>	<b>Illustration of Shannon's second theorem in a simple case</b>	<b>9</b>
4.1	Interpretation of the Second theorem . . . . .	10
4.2	Formulation of the problem in a simple case . . . . .	11
4.2.1	Random coding . . . . .	11
4.2.2	Decoding criterion . . . . .	11
4.3	An upper bound on the number of possible messages of length $L$ . . . . .	11
4.4	Interpretation based on Stirling formula . . . . .	12
<b>5</b>	<b>Conclusion</b>	<b>13</b>
<b>6</b>	<b>Bibliography</b>	<b>14</b>
6.1	Some historical references . . . . .	14
6.2	Printed references . . . . .	14
6.3	Web sites in April 2002 . . . . .	14

## 1 Abstract

The development of Shannon's theorems is illustrated in the case of a binary random signal and of a binary noise modifying a binary message in a binary symmetric memoryless channel. The development does not start from the notion of entropy (of course it appears in the computations) and avoids the use of joint entropy. The main tool used in the development is the approximation of the binomial law based on Stirling formula.

## 2 Introduction

Proofs of Shannon theorems are rather abstract and difficult to understand for many students in digital transmission who would like to have an intuitive idea of these theorems without becoming expert in information theory. It may be helpful to present an illustration of these proofs avoiding the direct use of difficult notions as mutual information and conditional or joint entropy.

The purpose of this short course is an attempt to illustrate in the simplest cases (binary messages and binary memoryless channel), the concept of entropy, and more specifically its use in the two theorems of C. Shannon on source (in section 3) and channel (in section 4) coding. The main tool that I shall use in this presentation is the approximation of probabilities densities of the binomial law based on Stirling formula which was used by L. Boltzmann in his statistical interpretation of entropy. This approach was also used by D. MacKay in his very good presentation.

It is perhaps interesting to recall the importance of the notion of entropy in the development of science ; I do not list the numerous domains where entropy is a useful tool :

Rudolf Clausius (1865) has invented the notion of entropy in the domain of thermodynamics. This word was derived by him from the greek “ $\eta\tau\rho\omega\pi\eta$ ” meaning “changing”. Ludwig Boltzmann (1877) gave an interpretation of this concept in terms of probability theory, relating temperature and random motion of gaz molecules.

Max Plank (1901) has used this statistical interpretation in order to modelize black body radiation, which led him to the discovery of quantum mechanics, discovery which was enriched by Albert Einstein (1905) who also based his development on the works of Boltzmann.

Claude Shannon (1948) also found inspiration in the works of L. Boltzmann in his founding of information theory, and in establishing the basic theorems on the lower bound concerning messages compression (source coding), and the upper bound giving the maximum number of errors that one can accept in a message transmission so that the original can be reconstructed in full (channel coding).

The lower bound in source coding is reached in using the arithmetic coding of J. Rissanen and G. Langdon (1978) ; and the performances of turbocodes (C. Berrou *et al.*, 1993) are close to the upper bound of channel coding.

## 3 Illustration of Shannon’s first theorem in a simple case

We consider the emission of a message  $B(\ell)$  of length  $L$  composed only of binary independant random data : ‘ones’ with probability  $p$  less than  $1/2$  and ‘zeros’ with probability  $(1 - p)$  (fig. 1).

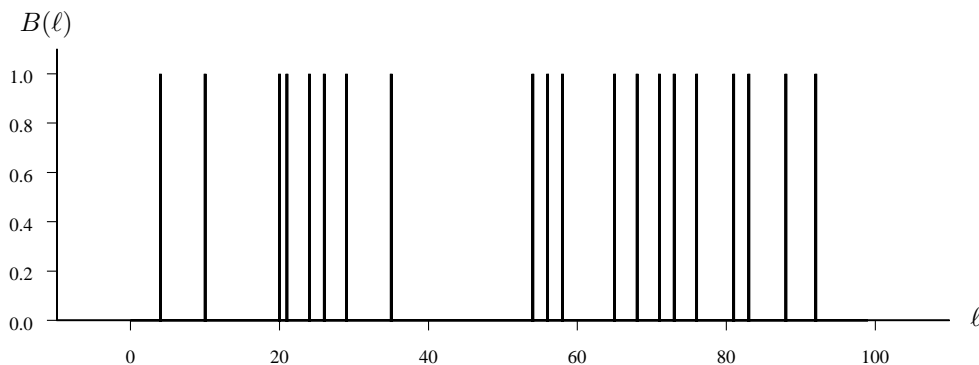


Figure 1: Example of message made of ‘1’ with probability  $p = 0.2$  and of ‘0’ with probability  $(1 - p) = 0.8$ .

The emitted messages have the following property (according to the law of large numbers): in a message of length  $L$ , there are about  $pL$  ‘1’s and  $(1 - p)L$  ‘0’s (fig. 2). The main idea in the first theorem is that it is sufficient to code those messages with  $pL$  ‘1’s and  $(1 - p)L$  ‘0’s since the others almost never occur.

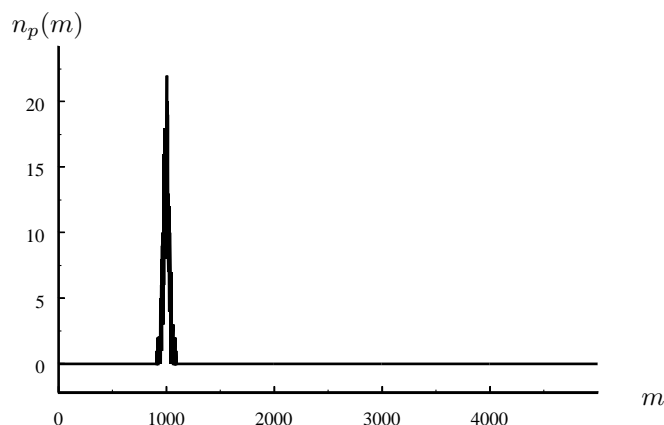


Figure 2: Histogram of the number of "1"s in a message of length  $L = 5000$  when the probability  $p$  of emission of a "1" is 0.2 ; this number is almost always between 900 and 1100 ; according to the law of large numbers, when the length  $L$  is large , the histogram of this number  $m$  divided by  $L$  tends to a Dirac distribution in  $m/L = p$ .

### 3.1 Interpretation of the first theorem

The first theorem states that it is possible to code the sequence  $B(\ell)$  with only

$$L \left( (1-p) \log_2 \frac{1}{1-p} + p \log_2 \frac{1}{p} \right) \text{ bits}$$

instead of the  $L$  bits necessary to code  $2^L$  different binary sequences. The quantity

$$H_B(p) = (1-p) \log_2 \frac{1}{1-p} + p \log_2 \frac{1}{p}, \quad (1)$$

is the entropy of the sequence  $B(\ell)$ . C. Shannon follows in his development the idea of L. Boltzmann who proceeded in a similar way in order to count the number of molecules in motion in a given volume. We shall illustrate the main steps of the development :

1. We show the distribution of probable messages and compute an approximation (section 3.2);
2. We estimate their number and compare this number to the number of all possible messages (section 3.3) ;
3. We deduce from this number the basis of compression methods (section 3.4).

### 3.2 The number of probable messages approximated by Stirling formula

According to the binomial law, the number of messages with  $m$  ones and  $L - m$  zeros is given by (fig 3)

$$n_p(m) = \frac{L!}{m!(L-m)!} (1-p)^{L-m} p^m, \quad (2)$$

when the probability of occurrence of a 'one' is  $p$ .

#### 3.2.1 Approximation based on Stirling formula

Using Stirling formula,

$$m! \simeq \sqrt{2\pi m} \left( \frac{m}{e} \right)^m, \quad (3)$$

$n_p(m)$  becomes

$$n_p(m) \simeq \sqrt{\frac{L}{2\pi m(L-m)}} \left( \frac{L}{e} \right)^L \left( \frac{e(1-p)}{L-m} \right)^{L-m} \left( \frac{ep}{m} \right)^m, \quad (4)$$

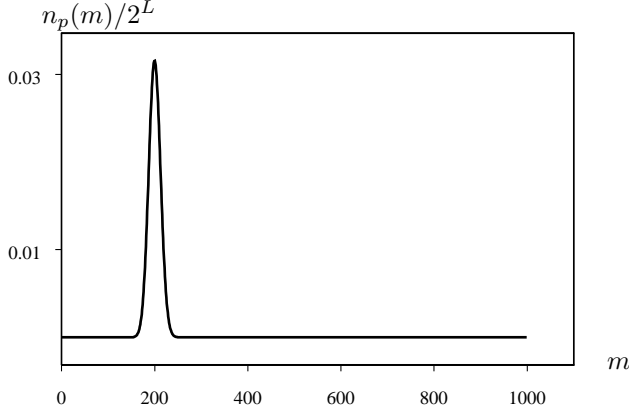


Figure 3: Probability density of the number of "1"s in messages of length  $L = 1000$  when the probability  $p$  of emitting a "1" is 0.2 ;  $m$  is almost always between 180 and 220 ; According to the law of large numbers, when the number of data,  $L$  is large , this density tends to the one of a Gaussian law with average  $p \times L$  and variance  $Lp(1-p)$  , the density of  $m/L$  tends to a Dirac distribution in  $\frac{m}{L} = p$  .

or

$$n_p(m) \simeq \sqrt{\frac{1}{2\pi L \frac{m}{L} (1 - \frac{m}{L})}} \left(\frac{1-p}{1-\frac{m}{L}}\right)^{L(1-\frac{m}{L})} \left(\frac{p}{\frac{m}{L}}\right)^{L\frac{m}{L}} . \quad (5)$$

Naming

$$\frac{m}{L} = q, \quad (6)$$

eq. (5) becomes

$$n'_p(q) = n_p(m) \simeq \frac{1}{\sqrt{2\pi Lq(1-q)}} \left(\frac{1-p}{1-q}\right)^{L(1-q)} \left(\frac{p}{q}\right)^{Lq}, \quad (7)$$

or in terms of logarithms

$$\begin{aligned} \log_e n'_p(q) &\simeq \log_e n_p(m) = \log_e \frac{1}{\sqrt{2\pi Lq(1-q)}} \\ &+ L[(1-q)(\log_e(1-p) - \log_e(1-q)) + q(\log_e(p) - \log_e(q))]. \end{aligned} \quad (8)$$

### 3.2.2 Gaussian approximation

According to the law of large numbers, this law is close to the Gaussian law of mean value  $Lp$  and variance  $Lp(1-p)$ . This is verified in replacing  $q$  by  $p + \varepsilon$  :

$$\begin{aligned} n''_p(\varepsilon) &= \log_e n'_p(q) \simeq -\frac{1}{2} \log_e [2\pi L(p + \varepsilon)(1 - p - \varepsilon)] \\ &+ L[(1 - p - \varepsilon)(\log_e(1 - p) - \log_e(1 - p - \varepsilon)) + (p + \varepsilon)(\log_e(p) - \log_e(p + \varepsilon))], \end{aligned} \quad (9)$$

$$\begin{aligned} n''_p(\varepsilon) &\simeq -\frac{1}{2} \log_e (2\pi Lp(1-p)) + \left(-\frac{\varepsilon}{2(1-p)} + \frac{\varepsilon}{2p}\right) \\ &+ L\left((1-p-\varepsilon)\frac{\varepsilon}{1-p} - (p+\varepsilon)\frac{\varepsilon}{p}\right). \end{aligned} \quad (10)$$

The second term in eq. (10) can be neglected when  $L$  is large

$$n''_p(\varepsilon) \simeq -\frac{1}{2} \log_e (2\pi Lp(1-p)) - L\frac{\varepsilon^2}{2p(1-p)}. \quad (11)$$

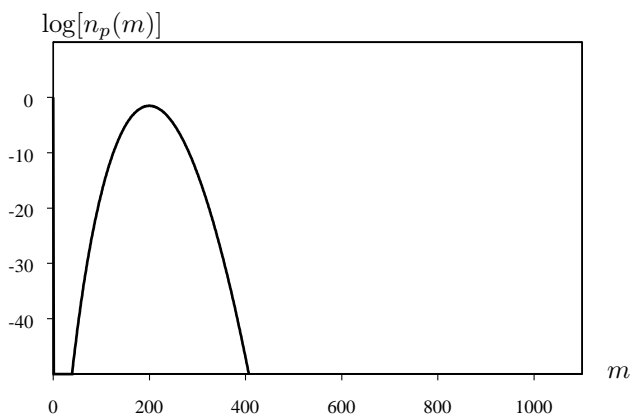


Figure 4: Probability density of the number of "1"s in messages of length  $L = 1000$  when the probability  $p$  of emitting a "1" is 0.2 ; same data as previously presented on a logarithmic scale. The parabolic shape of the function characterises the Gaussian law approximation.

The logarithm of the density of  $\frac{m}{L}$  is (fig 4)

$$\boxed{\frac{1}{2} \log_e \frac{L}{2\pi p(1-p)} - \frac{L \left(\frac{m}{L} - p\right)^2}{2p(1-p)}}. \quad (12)$$

When  $L$  is large the first term of the sum (11) can also be neglected. Most of the sequences have a number of ones between  $L(p - \delta)$  and  $L(p + \delta)$  where  $\delta$  can be as small as desired. When  $L$  is large the probability that  $m$  is outside this range decreases to zero (see fig. 5 and 6).

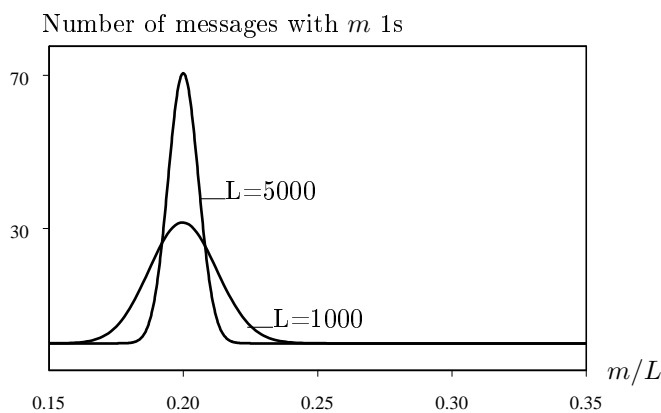


Figure 5: Probability density of the number of "1"s in messages of length  $L = 1000$  and  $L = 5000$  when the probability  $p$  of emitting a "1" is 0.2.

### 3.2.3 Inequality of Bienaymé Chebyshev:

It is perhaps useful to recall this inequality formulated in the case of the present approximation :

$$\text{probability that } \left( \left| \frac{m}{L} - p \right| > s \right) < \frac{p(1-p)}{s^2 L}. \quad (13)$$

When  $L$  is large, the probability that  $m$  lies outside this range decreases at least as fast as  $1/L$ . In the present particular case, it decreases much faster, like  $e^{-2(1-2p)L}$  (see section 3.3.2).

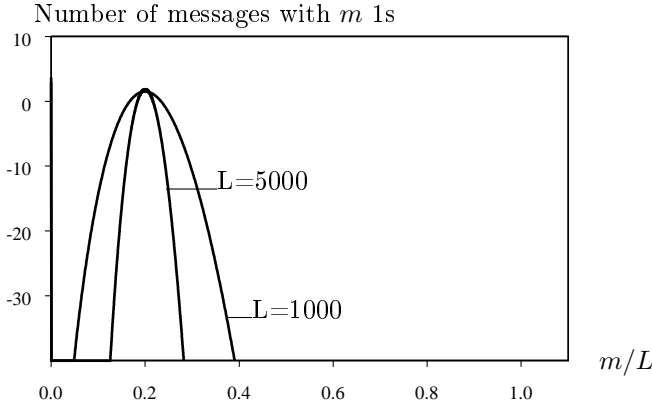


Figure 6: Probability density of the number of "1"s in messages of length  $L = 1000$  and  $L = 5000$  when the probability  $p$  of emitting a "1" is 0.2 ; same data as previously presented on a logarithmic scale. The second derivative of the parabola is  $-\frac{L}{2p(1-p)}$ .

### 3.3 What is the number of messages having exactly or more or less $pL$ ones and $(1-p)L$ zeros ?

According to the binomial law,  $pL$  being an integer, the number of messages having  $pL$  '1's and  $(1-p)L$  '0's is :

$$n_{1/2}(p) = \frac{L!}{(pL)![(1-p)L]!}, \quad (14)$$

while the total amount of possible messages is  $2^L$ . We have

$$\sum_{pL=0}^L n_{1/2}(p) = \sum_{pL=0}^L \frac{L!}{(pL)![(1-p)L]!} = 2^L. \quad (15)$$

#### 3.3.1 Approximation using Stirling formula

According to Stirling formula, this number (14) is approximated by

$$n_{1/2}(p) = \frac{\sqrt{2\pi L} \left(\frac{L}{e}\right)^L}{\sqrt{2\pi Lp} \left(\frac{Lp}{e}\right)^{Lp} \sqrt{2\pi L(1-p)} \left(\frac{L(1-p)}{e}\right)^{L(1-p)}}. \quad (16)$$

It can be written

$$n_{1/2}(p) = \frac{1}{\sqrt{2\pi Lp(1-p)} p^{Lp} (1-p)^{L(1-p)}}, \quad (17)$$

or

$$n_{1/2}(p) = \frac{1}{\sqrt{2\pi Lp(1-p)} 2^{-L(p \log_2 p + (1-p) \log_2 (1-p))}}, \quad (18)$$

or also

$$n_{1/2}(p) = 2^{-L(p \log_2 p + (1-p) \log_2 (1-p)) - \frac{1}{2} \log_2 (2\pi Lp(1-p))}. \quad (19)$$

When  $L$  is large, the prevailing term in the exponent is proportional to the entropy (fig. 7 and 8)

$$\boxed{LH_B(p) = L(-p \log_2 p - (1-p) \log_2 (1-p))}, \quad (20)$$

and so

$$\boxed{n_{1/2}(p) = 2^{LH_B(p)}}. \quad (21)$$

The number of messages with  $Lp$  ones is less than  $2^{L(H_B(p)+\delta)}$  where  $\delta$  can be as small as desired.  $L(H_B(p) + \delta)$  bits are sufficient to code them.

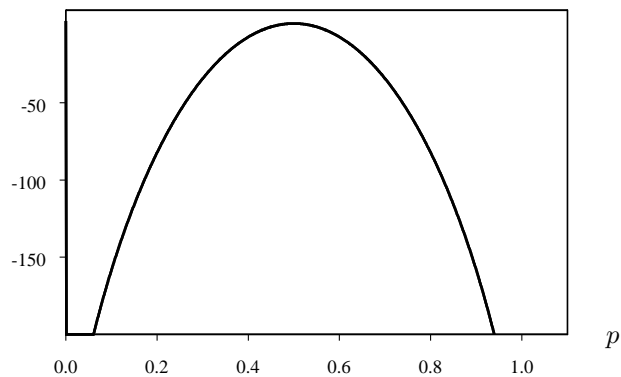


Figure 7: Proportion of messages of length  $L$  having exactly  $m$  ones, (logarithmic scale); this proportion is  $2^{L(H_B-1)}$ .

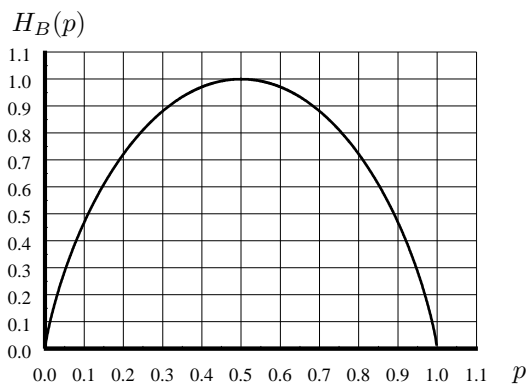


Figure 8: Entropy function.

**3.3.2 The number of bits given by the first theorem is sufficient to code all messages for which the probability of 'ones' is less than  $p$**

We consider messages for which the probability of 'ones' is

$$\sum_{Lx=0}^{Lp} n_{1/2}(x) = \sum_{Lx=0}^{Lp} \frac{L!}{Lx!L(1-x)!}, \quad (22)$$

where  $x$  lies in the range  $[0, 1]$  and  $Lx$  is an integer.

We intend to show that this probability is bounded by

$$\sum_{Lx=0}^{Lp} n_{1/2}(x) \leq n_{1/2}(p)(1 + \delta), \quad (23)$$

where  $\delta$  can be made as small as desired in choosing  $L$  sufficiently large. When  $L$  is sufficiently large, we have the following Gaussian approximation

$$n_{1/2}(x) \simeq \sqrt{\frac{2L}{\pi}} e^{-2L(x-\frac{1}{2})^2}. \quad (24)$$

When  $x$  decreases starting from  $p$  ( $x < p$ ), this function decreases extremely fast (supposing  $p$  not too close to 0.5)

$$n_{1/2}(x) \simeq \sqrt{\frac{2L}{\pi}} e^{-2L(p-\frac{1}{2})^2} e^{-2L[(x-\frac{1}{2})^2 - (p-\frac{1}{2})^2]}, \quad (25)$$

or

$$n_{1/2}(x) \simeq n_{1/2}(p)e^{-2L[x-p](x-1+p)} < n_{1/2}(p)e^{2(1-2p)L[x-p]}. \quad (26)$$

Consequently,

$$\sum_{Lx=0}^{Lp} n_{1/2}(x) < \frac{1 - e^{-2Lp(1-2p)}}{1 - e^{-2(1-2p)L}} n_{1/2}(p). \quad (27)$$

or for  $L$  large :

$$\sum_{Lx=0}^{Lp} n_{1/2}(x) < (1 + e^{-2(1-2p)L})n_{1/2}(p). \quad (28)$$

So, the number of messages with probability less than  $p$  can be neglected since it is very small with respect to the number of those with probability close to  $p$  : the decrease of  $e^{-2L(x-p)^2}$  is extremely fast when  $(x-p)$  decreases (fig 9 and 10). This approximation will be useful in the illustration of the second theorem in section 4.

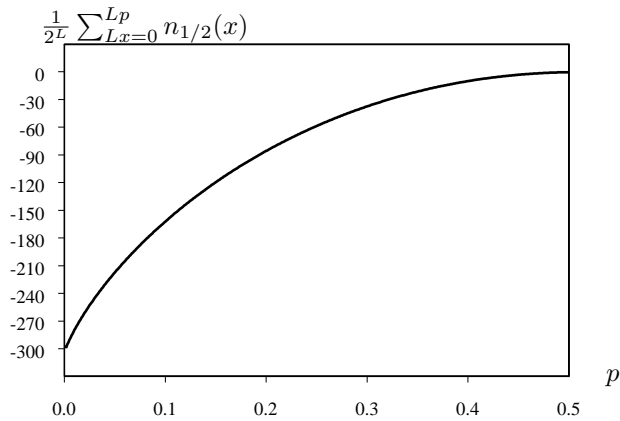


Figure 9: Proportion of messages where the probability of 'ones' is between zero and  $p$  (log scale) as a function of  $p$  for a given  $L = 1000$ . This number is comparable with the number of messages having exactly  $p$  'ones',  $2^{LH(p)}$ .

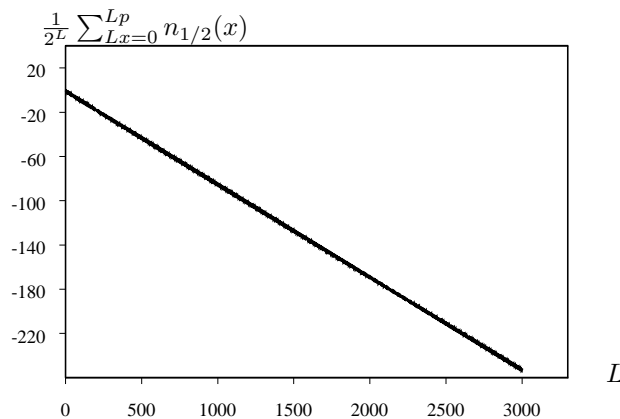


Figure 10: Proportion of messages where the probability of 'ones' is between zero and  $p = 0.2$  (log scale) as a function of  $L$ . This number is comparable to the number of messages having exactly  $p$  'ones'.

### 3.4 The principle of compression techniques

In order to compress the messages, the improbable messages (those having not around  $p$  ones, or if one prefers more than  $p + \delta$  ones as in section 3.3.2) are discarded (fig 11), and the remaining  $2^{LH_B(p)}$  messages are given a number, for instance in using the arithmetic coding of Rissanen and Langdon.

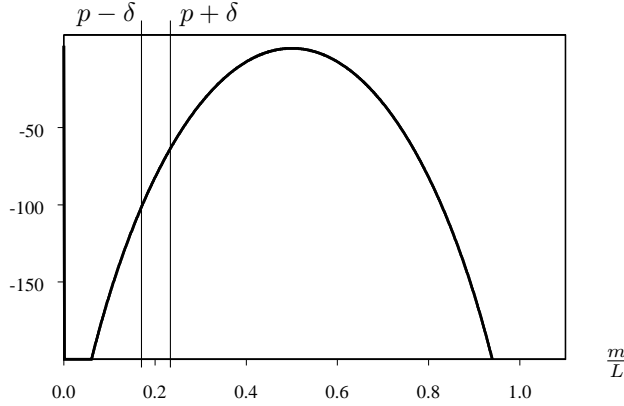


Figure 11: Proportion of messages of length  $L$  having exactly  $m$  ones, logarithmic scale ; it is not necessary to code the messages having more than  $p + \delta$  ones or less than  $p - \delta$  ones since they are very unlikely.

These acceptable messages are called “*typical*”. According to the number of different messages of this kind, for coding them in bits, a length slightly larger than  $LH_B(p)$  is sufficient.

#### 3.4.1 Approximation of the entropy for small values of $p$

It is perhaps useful to have this approximation in mind, as an order of magnitude. When  $p$  is small,

$$LH_B(p) = Lp \log_2 \frac{L}{Lp} - L(1-p) \log_2 e \log_e(1-p), \quad (29)$$

$$LH_B(p) \simeq Lp \left( \log_2 \frac{L}{Lp} + \log_2 e \right). \quad (30)$$

For each of the  $Lp$  ‘one’s of the message, the necessary number of bits is the number of bits necessary to code the average length between two ones, i.e.  $\frac{L}{Lp}$  plus  $\log_2 e$ .

## 4 Illustration of Shannon’s second theorem in a simple case

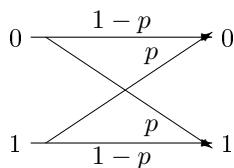
A message of entropy 1 ( $p_M = 1/2$ ) and of length  $M$  can be extended in appending a syndrome in order to build a message of length  $L$ . The syndrome is deduced from the original message by deterministic operations. It will be used in order to correct transmission errors.

The redundant message of length  $L$  does not necessarily contain explicitly the original message of length  $M$ . It may be obtained by any choice of  $2^M$  words among the  $2^L$  that can be transmitted. If it contains explicitly the original message, the code is called “systematic”, but this reduces the choice among the possible codes.

Then the entropy of the extended message is  $H_M = M/L$  : there are  $2^M$  messages and the entropy  $H_M$  of a subset of  $2^M$  elements taken among  $2^L$  elements is such that

$$2^M = 2^{LH_M}. \quad (31)$$

This message is transmitted to a receiver and modified by a binary noise independent of the original message (fig 12). This transmission is characterized by a probability of error  $p$ , or an entropy  $H_B = -p \log_2 p - (1-p) \log_2(1-p)$ .

Figure 12: binary symmetric channel with error probability  $p$ .

In section 4.1, we give the interpretation of the theorem in this simple case; in section 4.2 we give the corresponding formulation of the problem; in section 4.3 we deduce the maximal number of possible messages and in section 4.4 we finally show that we can have a number of different messages as close as desired to this bound.

#### 4.1 Interpretation of the Second theorem

Shannon's second theorem states that if the sum of the entropies of the message  $M$  and of the independant noise  $B$ , that is  $H_M + H_B$  satisfies

$$H_M + H_B < 1, \quad (32)$$

then it is possible to find a method for coding  $M$  so that it is almost always possible to reconstruct exactly the message  $M$  from the received message corrupted by noise. The length of the redundancy introduced in the message,  $L - M$  must be sufficient to describe the noise, and so to decode the original message (fig 13):

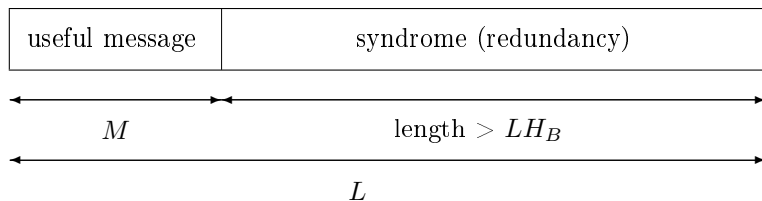


Figure 13: Complementation of a message by adjunction of a syndrome of sufficient length so that transmission errors can be corrected. The minimal length of the syndrome is proportional to the entropy of the noise  $B$  (we suppose that the entropy of the original message of length  $M$  is equal to 1, and consequently that it cannot be compressed).

$$L - M > LH_B, \quad (33)$$

or

$$\boxed{\frac{M}{L} < 1 - H_B.} \quad (34)$$

As a true mathematician, Claude Shannon shows that it certainly exists a method of coding allowing decoding without error, but he does not propose any track to find it ! He does not say if it is possible to find efficient coders for which decoding is not excessively complex... One can deduce from Shannon's proof that most of the codes are probably good codes, because on average they attain the object ; however the complexity of their decoding forbids to consider their use. To find a relevant coding technique, particularly a reasonably complex decoding technique, remains an open problem. One had to wait for 45 years a proposal of coding and decoding with performances close to Shannon's bound : turbocodes invented par Claude Berrou and his colleagues ... Still, if the experimental verification is convincing, it does not seem that there is, at the present moment, a rigorous theoretical justification of the performances of turbocodes.

Shannon's proof is very abstract, elegant and concise. The reader will find several textbooks (see the bibliography below). My purpose is only an attempt to illustrate in the simplest case the main points of the proof that are not intuitive ; I hope that this could help students and persons interested in this subject who are stopped by notions considered probably as obvious by experts in information theory and on which those experts do not insist. The central point in the theorem is the following:

## 4.2 Formulation of the problem in a simple case

Long messages (length  $L$ ) are sent. They are received with at most  $Lp$  erroneous bits. The  $2^M$  possible messages are coded by  $2^M$  words among the  $2^L$  possible ones ; if the fraction  $\frac{M}{L}$  is less than the bound (34) related to the entropy of the noise, then it is very unlikely that a message coded and corrupted by  $Lp$  errors can be mistaken for one of the other messages that could be sent; "*mistaken*" meaning that the Hamming between two messages is less than  $Lp$ .

### 4.2.1 Random coding

The number of possible coders is very large : a coder transforms two of the  $2^M$  messages in two different  $2^L$  codewords. The number of coders is

$$\frac{(2^L)!}{(2^{L-M})!} \text{ or } \sqrt{\frac{2^L}{2^{L-M}} \frac{2^{L2^L - (L-M)2^{L-M}}}{e^{2^L - 2^{L-M}}}}.$$

Shannon assumes that all these coders can be chosen with the same probability. Here we will choose one coder randomly. The probability that one of the  $2^L$  words is a codeword is

$$\frac{2^M}{2^L},$$

since there are  $2^L$  different words et  $2^M$  possible messages. When  $L$  increases for a fixed  $M$ , this proportion decreases quite fast. It will be possible to find codewords so that the Hamming distance between two codewords is larger than a given threshold and so to avoid confusion between these codewords. The question is then : for a given  $M$ , what is the lower possible value for  $L$  so that confusion is almost certainly avoided.

### 4.2.2 Decoding criterion

The probability of transmission error is  $p$ . A received codeword is recognized as one of the  $2^M$  messages if the Hamming distance between the word corresponding to this last message and the received word is less than  $Lp$ .

A received word corresponding to the emission of a codeword is certainly correctly recognized since there are at most  $Lp$  errors on a received message. However, we have to find the condition under which there cannot be recognition mistakes. We suppose that the receiver only receives words that are codewords corrupted by noise.

## 4.3 An upper bound on the number of possible messages of length $L$

If the probability of error is at most  $p$ , it is possible to find a code such that a message of length  $L$  assumed very large can be used to encode  $2^{L(1-H_B(p))}$  messages: each of the  $2^M$  messages can be transformed in one of the  $2^{LH_B}$  possible messages at reception : in order to avoid confusion,  $M$  must satisfy

$$2^M 2^{LH_B} \leq 2^L. \quad (35)$$

If there is no transmission noise ( $p = 0, H_B(p) = 0$ ), it is possible to code  $2^L$  words ; if  $p = \frac{1}{2}$  and  $H_B(p) = 1$ , it is not possible to transmit information on the channel.

There cannot be more than  $2^{L-LH_B(p)}$  different messages, since there are  $2^{LH_B(p)}$  configurations of noise and  $2^L$  codewords.

A larger number of codewords would necessarily generate recognition mistakes. We have to show that the number of messages can be as close as desired to this bound.

#### 4.4 Interpretation based on Stirling formula

When  $L$  is large, two different codewords among the  $2^M$  taken at random have on average  $\frac{L}{2}$  identical bits. The distribution of the number of identical bits between two of the  $2^L$  words follows the binomial law

$$n_{1/2}(x) = \frac{L!}{(xL)![(1-x)L!]}. \quad (36)$$

After transmission of a codeword, the erroneous corresponding codeword has on average  $\frac{L}{2}$  bits identical to the corresponding bits of any of the other codewords ; the number of these identical bits also follows the law (36).

There are  $2^M$  different codewords : we try to see if any of the  $(2^M - 1)$  among these codewords, when they are received after corruption by noise, can be confused with the last one, that is if the Hamming distance between one of the  $(2^M - 1)2^{LH_B}$  received words and the  $2^M$ -th is less than  $L(p + \varepsilon)$ .

The probability that this distance is less than  $Lp$  is

$$Q = \sum_{xL=0}^{Lp} \frac{1}{2^L} n_{1/2}(x) = \sum_{xL=0}^{Lp} \frac{L!}{xL!(L-xL)!} \left(\frac{1}{2^L}\right), \quad (37)$$

which is approximated by

$$Q = \sum_{xL=0}^{Lp} \frac{1}{\sqrt{2\pi L(1-x)x}} \frac{1}{(1-x)^{L(1-x)}} \frac{1}{(x)^{Lx}} \left(\frac{1}{2^L}\right). \quad (38)$$

We have seen in section 3.3.2, eq. (23 - 28), that the order of magnitude of this probability of error is given by (see fig. 14)

$$Q \simeq \frac{n_{1/2}(p)}{2^L}, \quad (39)$$

or, according to eq. (21):

$$Q \simeq \frac{2^{LH_B}}{2^L}. \quad (40)$$

This is the probability that one of the codewords can be mistaken for another one. We can deduce from this formula a bound on the probability that none of the  $2^M - 1$  words are closer than  $Lp$  from the  $2^M$ -th word :

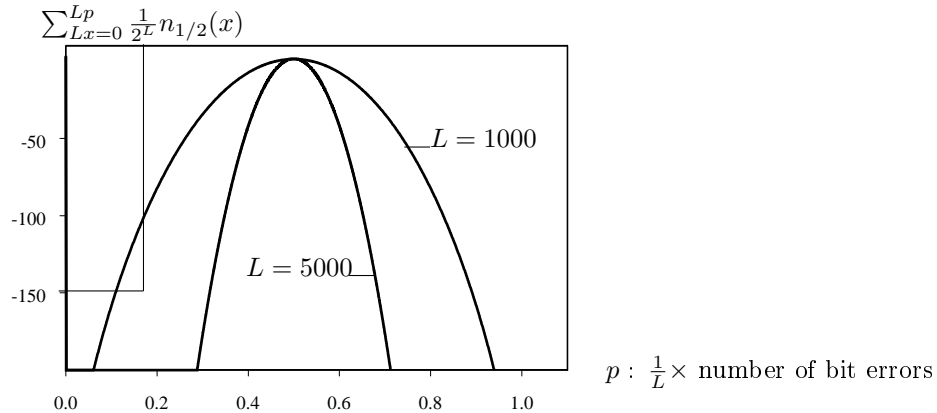


Figure 14: Probability of error between two words (log scale) : when  $L$  is large enough, the probability of mistaking one message for another can be made arbitrarily small. If this probability decreases faster than  $2^M$  when  $L$  increases,  $\frac{M}{L}$  remaining constant, it will be almost always possible to recognize the emitted codeword.

Since the probability of several events is less than the sum of the probabilities of each of these events

$$\text{proba } (a \text{ or } b) \leq \text{proba } (a) + \text{proba } (b); \quad (41)$$

Then, the probability that there is at least one error is bounded by

$$S = \sum_1^{2^M-1} Q \simeq 2^M Q \simeq 2^{M+LH_B-L}. \quad (42)$$

We suppose that the condition (34) is satisfied :

$$M + LH_B - L < 0. \quad (43)$$

If  $L$  is sufficiently large, the fraction  $\frac{M}{L}$  remaining constant, there is an  $\alpha$  positive such that

$$\frac{M}{L} < 1 - H_B - \alpha. \quad (44)$$

$$S \simeq 2^{M+LH_B-L} < 2^{-\alpha L}, \quad (45)$$

If  $L$  increases, the probability of error decreases to zero when condition (34) :

$$H_M = \frac{M}{L} < 1 - H_B, \quad (46)$$

is satisfied. In Shannon's terminology, this bound is the *capacity* of the channel. It is perhaps interesting to show the redundancy

$$\frac{L}{M} = \frac{1}{1 - H_B}, \quad (47)$$

that is necessary to satisfy Shannon's bound (fig 15).

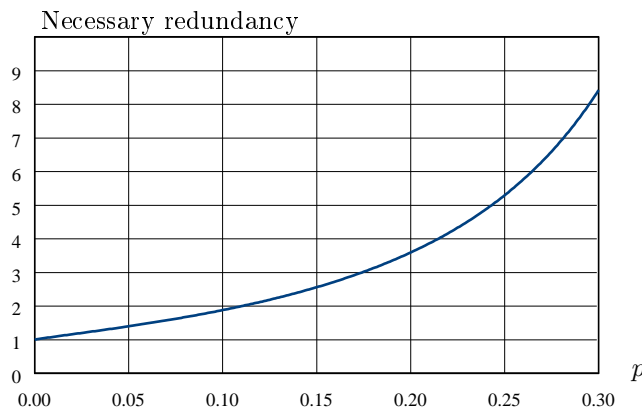


Figure 15: Redundancy corresponding to Shannon's bound as a function of the probability of transmission error.

## 5 Conclusion

We have proposed two simple illustrations of Shannon's theorems based on the use of Stirling formula. Although the results of C. Shannon are far more general, and in spite of the awkwardness of the computations, we hope that this presentation has been useful to help understanding the concrete aspects of the theorems. Improvements and corrections are welcome ; send mail to leroux@essi.fr.

## 6 Bibliography

### 6.1 Some historical references

R. Clausius, "Ueber verschiedene für die anwendung bequeme formen der Hauptgleichungen der mechanischen Wärmetheorie", (On different forms, convenient for application, of the main equations of the mechanical heat theory) *Annalen der physik und chemie*, band CXX5, no 7, 1865, pp 353-400.

L. Boltzmann, "Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht," (On the Relation Between the Second Law of the Mechanical Theory of Heat and the Probability Calculus with Respect to the Theorems on Thermal Equilibrium), *Sitzb. d. Kaiserlichen Akademie der Wissenschaften, mathematisch-naturwissen Cl. LXXVI, Abt II*, 1877, pp. 373-435.

M. Planck, "Über des Gesetz der Energieverteilung im Normalspectrum", "On the Law of Energy Distribution in Normal Spectra", *Annalen der Physik*, 4, 1901, pp 553-563. (french translation : A propos de la loi de distribution de l'énergie dans le spectre normal, *Sources et évolution de la physique quantique, textes fondateurs*, J. Leite-Lopes et B. Escoubès, eds, Masson, 1995. pp. 20-27.)

A. Einstein, "Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt," ("On a Heuristic Viewpoint Concerning the Production and Transformation of Light") *Annalen der Physik*, 17, 1905, pp. 132-148. (french translation : Un point de vue heuristique concernant la production et la transformation de la lumière, *Sources et évolution de la physique quantique, textes fondateurs*, J. Leite-Lopes et B. Escoubès, eds, Masson, 1995. pp. 28-40.)

C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

L. Brillouin, "Science and Information theory", Academic Press, 1962.

R. G. Gallager, "The work of Claude Shannon," *IEEE Trans. on IT*, nov. 2001.

### 6.2 Printed references

R.G. Gallager, "Information theory and reliable communication", Wiley, 1968.

T. M. Cover and J.A. Thomas, "Elements of information theory", Wiley, 1991.

G. Battail, "Théorie de l'information, application aux techniques de communications", Masson, 1997 (in french).

J. Rissanen and G.G. Langdon, "Arithmetic coding", *IBM J. Res. Develop.*, Vol. 23, No. 2, pp. 149-162, March 1979.

J. Rissanen and G.G. Langdon, "Universal modeling and coding", *IEEE Trans. on Information Theory*, Vol. 27, No. 1, pp. 12-23, January 1981.

C. Berrou, A. Glavieux and P. Thihimajshima, "Near Shannon limit error-correcting coding and decoding : turbo codes", *Proc. 1993, Int. Conf. Comm.*, pp 1064-1070.

C. Berrou and A. Glavieux, "Near Shannon limit error-correcting coding and decoding : turbo codes", *IEEE Trans. Comm.*, Oct. 1996, pp. 1261-1271.

### 6.3 Web sites in April 2002

Shannon's papers :

<http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html>

The courses of Marc Uro (Institute of telecommunications, Evry, France), in french :

<http://www-sim.int-evry.fr/~uro/old.htm>

David J.C. MacKay, "Information Theory, Inference and Learning Algorithms", Cavendish Laboratory, Cambridge, Great Britain, January 1995 :

<http://www.inference.phy.cam.ac.uk/mackay/info-theory/course.html>

<http://www.inference.phy.cam.ac.uk/mackay/itprnm/book.html#book>

Explanation of Stirling's formula on the page of B. Gourevitch about  $\pi$  (in french) :

<http://membres.lycos.fr/bgourevitch/mathematiciens/moivre/moivre.html>  
An english translation of one of Boltzmann's paper :  
<http://www.essi.fr/~leroux/boltztrad.ps>