Information, Entropy and Their Geometric Structures

Edited by

Frédéric Barbaresco and

Ali Mohammad-Djafari

Printed Edition of the Special Issue Published in *Entropy*

www.mdpi.com/journal/entropy

Frédéric Barbaresco and Ali Mohammad-Djafari (Eds.)

Information, Entropy and Their Geometric Structures

MDPI

Chapter 1:
Origins of Entropy and Information Theory

On Shannon's Formula and Hartley's Rule: Beyond the Mathematical Coincidence

Olivier Rioul and José Carlos Magossi

Abstract: In the information theory community, the following "historical" statements are generally well accepted: (1) Hartley did put forth his rule twenty years before Shannon; (2) Shannon's formula as a fundamental tradeoff between transmission rate, bandwidth, and signal-to-noise ratio came out unexpected in 1948; (3) Hartley's rule is inexact while Shannon's formula is characteristic of the additive white Gaussian noise channel; (4) Hartley's rule is an imprecise relation that is not an appropriate formula for the capacity of a communication channel. We show that all these four statements are somewhat wrong. In fact, a careful calculation shows that "Hartley's rule" in fact coincides with Shannon's formula. We explain this mathematical coincidence by deriving the necessary and sufficient conditions on an additive noise channel such that its capacity is given by Shannon's formula and construct a sequence of such channels that makes the link between the uniform (Hartley) and Gaussian (Shannon) channels.

Reprinted from *Entropy*. Cite as: Rioul, O.; Magossi, J.C. On Shannon's Formula and Hartley's Rule: Beyond the Mathematical Coincidence. *Entropy* **2014**, *16*, 4892–4910.

1. Introduction

As researchers in information theory, we all know that the milestone event that founded our field is Shannon's publication of his seminal 1948 paper [1] that created a completely new branch of applied mathematics and called it to immediate worldwide attention. What has rapidly become the emblematic classical expression of the theory is *Shannon's formula* [1,2]

$$C = \tfrac{1}{2}\log_2\Big(1+\frac{P}{N}\Big) \tag{1}$$

for the information capacity of a communication channel with signal-to-noise ratio P/N.

Hereafter we shall always express information capacity in binary units (bits) per *sample*. Shannon's well-known original formulation was in bits per second:

$$C = W\log_2\Big(1+\frac{P}{N}\Big) \qquad \text{bits/s.}$$

The difference between this formula and (1) is essentially the content of the *sampling theorem*, often referred to as Shannon's theorem, that the number of independent samples that can be put through a channel of bandwidth W hertz is $2W$ samples per second. We shall not discuss here whether the sampling theorem should be attributed to Shannon or to other authors that predate him in this discovery; see e.g., [3] for a recent account and extensive study on this subject.

The classical derivation of (1) was done in [1] as an application of Shannon's coding theorem for a memoryless channel, which states that the best coding procedure for reliable transmission

achieves a maximal rate of $C = \max_X I(X;Y)$ bits per sample, where X is the channel input with average power $P = \mathbb{E}(X^2)$ and $Y = X + Z$ is the channel output. Here Z denotes the additive Gaussian random variable (independent of X) that models the communication noise with power $N = \mathbb{E}(Z^2)$. By expanding mutual information $I(X;Y) = h(Y) - h(Y|X)$ as a difference of differential entropies, noting that $h(Y|X) = h(Z) = \log_2 \sqrt{2\pi eN}$ is constant, and choosing X Gaussian so as to maximize $h(Y)$, Shannon arrived at his formula $C = \max_X h(Y) - h(Z) = \log_2 \sqrt{2\pi e(P+N)} - \log_2 \sqrt{2\pi eN} = \frac{1}{2}\log_2(1 + P/N)$.

Formula (1) is also known as the *Shannon–Hartley formula*, and the channel coding theorem stating that (1) is the maximum rate at which information can be transmitted reliably over a noisy communication channel is often referred to as the *Shannon–Hartley theorem* (see, e.g., [4]). The reason for which Hartley's name is associated to the theorem is commonly justified by the so-called *Hartley's law*, which is described as follows:

> *During 1928, Hartley formulated a way to quantify information and its line rate (also known as data signalling rate R bits per second) [5]. This method, later known as Hartley's law, became an important precursor for Shannon's more sophisticated notion of channel capacity. (...)*
>
> *Hartley argued that the maximum number of distinguishable pulse levels that can be transmitted and received reliably over a communications channel is limited by the dynamic range of the signal amplitude and the precision with which the receiver can distinguish amplitude levels. Specifically, if the amplitude of the transmitted signal is restricted to the range of* $[-A, +A]$ *volts, and the precision of the receiver is* $\pm\Delta$ *volts, then the maximum number of distinct pulses* M *is given by* $M = 1 + \frac{A}{\Delta}$*. By taking information per pulse in bit/pulse to be the base-2-logarithm of the number of distinct messages* M *that could be sent, Hartley [5] constructed a measure of the line rate* R *as* $R = \log_2(M)$ *[bits per symbol].*
>
> —Wikipedia [4]

In other words, within a noise amplitude limited by Δ, by taking regularly spaced input symbol values in the range $[-A, A]$ with step 2Δ:

$$-A, -A + 2\Delta, \ldots, A - 2\Delta, A,$$

one can achieve a maximum total number of $M = A/\Delta + 1$ possible distinguishable values. This holds in the most favorable case where A/Δ is an integer, where the "+1" is due to the sample values at the boundaries—otherwise, M would be the integer part of $A/\Delta + 1$. Therefore, error-free communication is achieved with at most

$$C' = \log_2\Big(1 + \frac{A}{\Delta}\Big) \tag{2}$$

bits per sample. This equation strikingly resembles (1). Of course, the "signal-to-noise ratio" A/Δ is a ratio of amplitudes, not of powers, hence should not be confused with the usual definition P/N;

accordingly, the factor $1/2$ in Formula (1) is missing in (2). Also, (2) is only considered as an approximation of (1):

> *Hartley's rate result can be viewed as the capacity of an errorless M-ary channel (...). But such an errorless channel is an idealization, and if M is chosen small enough to make the noisy channel nearly errorless, the result is necessarily less than the Shannon capacity of the noisy channel (...), which is the Hartley–Shannon result that followed later [in 1948].*
>
> —Wikipedia [4]

In the information theory community, the following "historical" statements are generally well accepted:

(1) *Hartley did put forth his rule* (2) *twenty years before Shannon.*

(2) *The fundamental tradeoff* (1) *between transmission rate, bandwidth, and signal-to-noise ratio came out unexpected in 1948: the time was not even ripe for this breakthrough.*

(3) *Hartley's rule is inexact while Shannon's formula is characteristic of the additive white Gaussian noise (AWGN) channel ($C' \neq C$).*

(4) *Hartley's rule is an imprecise relation between signal magnitude, receiver accuracy and transmission rate that is not an appropriate formula for the capacity of a communication channel.*

In this article, we show that all these four statements are somewhat wrong. The organisation is as follows. Sections 2–5 will each defend the opposite view of statements (1)–(4) correspondingly. Section 6 concludes through a detailed mathematical analysis.

2. Hartley's Rule is not Hartley's

Hartley [5] was the first researcher to try to formulate a theory of the transmission of information. Apart from stating explicitly that the amount of transmitted information is proportional to the transmission bandwidth, he showed that the number M of possible alternatives from a message source over given a time interval grows exponentially with the duration, suggesting a definition of information as the logarithm $\log M$. However, as Shannon recalled in 1984:

> *I started with information theory, inspired by Hartley's paper, which was a good paper, but it did not take account of things like noise and best encoding and probabilistic aspects.*
>
> —Claude Elwood Shannon [6]

Indeed, no mention of signal vs. noise, or of amplitude limitation A or Δ was ever made in Hartley's paper [5]. One may then wonder how (2) was coined as Hartley's law.

The oldest reference that attributes (2) to Hartley—and incidentally cited in the Wikipedia page [4]—seems to be the classical 1965 textbook of Wozencraft and Jacobs, most notably its introduction chapter:

(...) in 1928, Hartley [5] reasoned that Nyquist's result, when coupled with a limitation on the accuracy of signal reception, implied a restriction on the amount of data that can be communicated reliably over a physical channel. Hartley's argument may be summarized as follows. If we assume that (1) the amplitude of a transmitted pulse is confined to the voltage range $[-A, A]$ *and (2) the receiver can estimate a transmitted amplitude reliably only to an accuracy of* $\pm\Delta$ *volts, then, as illustrated in [the] Figure (...), the maximum number of pulse amplitudes distinguishable at the receiver is* $(1 + A/\Delta)$*. (...)*

[in the Figure's legend:*] Hartley considered received pulse amplitudes to be distinguishable only if they lie in different zones of width* 2Δ *(...)*

Hartley's formulation exhibits a simple but somewhat inexact interrelation among (...) the maximum signal magnitude A*, the receiver accuracy* Δ*, and the allowable number of message alternatives. Communication theory is intimately concerned with the determination of more precise interrelations of this sort.*

—John M. Wozencraft; Irwin Mark Jacobs [7]

The textbook was highly regarded and still widely used today. Its introductive text has become famous to many researchers in the field of communication theory and has had a tremendous impact. This would explain why (2) is now widely known as Hartley's capacity law.

One may then wonder whether Wozencraft and Jacobs have found such a result themselves while attributing it to Hartley or whether it was inspired from other researchers. We found that the answer is probably in very first tutorial article in information theory that was ever published by E. Colin Cherry in 1951:

Although not explicitly stated in this form in his paper, Hartley [5] has implied that the quantity of information which can be transmitted in a frequency band of width B *and time* T *is proportional to the product:* $2BT \log M$*, where* M *is the number of "distinguishable amplitude levels." [...] He approximates the waveform by a series of steps, each one representing a selection of an amplitude level. [...] For example, consider a waveform to be traced out on a rectangular grid [...], the horizontal mesh-width representing units of time (equal to* $1/2B$ *in order to give the necessary* $2BT$ *data in a time* T*), and the vertical the "smallest distinguishable" amplitude change; in practice this smallest step may be taken to equal the noise level* n*. Then the quantity of information transmitted may be shown to be proportional to* $BT \log(1 + a/n)$ *where* a *is the maximum signal amplitude, an expression given by Tuller [8], being based upon Hartley's definition of information.*

—E. Colin Cherry [9]

Cherry attributes (2) to an *implicit* derivation of Hartley but cites the explicit derivation of Tuller [8]. The next section investigates the contribution of Tuller and others.

3. Independent 1948 Derivations of Shannon's Formula

In the introduction to his classic textbook, Robert McEliece wrote:

> *With many profound scientific discoveries (for example Einstein's discovery in 1905 of the special theory of relativity) it is possible with the aid of hindsight to see that the times were ripe for a breakthrough. Not so with information theory. While of course Shannon was not working in the vacuum in the 1940's, his results were so breathtakingly original that even the communication specialists of the day were at a loss to understand their significance.*
>
> —Robert McEliece [10]

One can hardly disagree with this statement when one sees the power and generality of Shannon's results. Just to mention a few examples: the introduction of the formal architecture of communication systems (Shannon's paradigm) with explicit distinction between source, channel and destination; the emphasis on digital representation to make the chance of error as small as desired; the consideration of codes in high dimensions; and the use of probabilistic models for the signal as well as for the noise, *via* information theoretic tools like entropy and mutual information. Shannon's ideas were revolutionary, in keeping with J.R. Pierce's famous quote:

> *In the end, [1] and the book based on it came as a bomb, and something of a delayed-action bomb.*
>
> —John R. Pierce [11]

Indeed, [1] being so deep and profound, did not have an immediate impact. As Robert Gallager recalls:

> *(...) two important papers (...) were almost concurrent to [1].*
>
> *The first subsequent paper was [12], whose coauthors were B. R. Oliver and J. R. Pierce. This is a very simple paper compared to [1], but it had a tremendous impact by clarifying a major advantage of digital communication. (...) It is probable that this paper had a greater impact on actual communication practice at the time than [1].*
>
> *The second major paper written at about the same time as [1] is [2]. This is a more tutorial amplification of the AWGN channel results of [1]. (...) This was the paper that introduced many communication researchers to the ideas of information theory.*
>
> —Robert Gallager [13]

In [12], Shannon's Formula (1) was used without explicit reference to the Gaussian nature of the added white noise, as the capacity of an "ideal system". On the other hand, [2] was devoted to a geometric proof of Formula (1).

It appears, therefore, that Shannon's Formula (1) was *the* emblematic result that impacted communication specialists at the time, as expressing the correct tradeoff between transmission

rate, bandwidth, and signal-to-noise ratio. It is one Shannon's result that is the best known and understood among communications engineers. As Verdú has noticed in [14], only a few months after the publication of [2], M. Golay [15] referred to (1) as "the now classical expression for the information reception capacity of a channel." In the following years, finding "codes to reach the promised land (1)" [16] became the "holy grail of information theory" [14].

As far as (1) is concerned, Shannon, after the completion of [1], acknowledged other works:

> *Formulas similar to* (1) *for the white noise case have been developed independently by several other writers, although with somewhat different interpretations. We may mention the work of N. Wiener [17], W. G. Tuller [8], and H. Sullivan in this connection.*
>
> —Claude Elwood Shannon [1]

Unfortunately, Shannon gave no specific reference to H. Sullivan. S. Verdú cited many more contributions during the same year of 1948:

> *By 1948 the need for a theory of communication encompassing the fundamental tradeoffs of transmission rate, reliability, bandwidth, and signal-to-noise ratio was recognized by various researchers. Several theories and principles were put forth in the space of a few months by A. Clavier [18], C. Earp [19], S. Goldman [20], J. Laplume [21], C. Shannon [1], W. Tuller [8], and N. Wiener [17]. One of those theories would prove to be everlasting.*
>
> —Sergio Verdú [14]

Lundheim reviewed some of these independent discoveries and concludes:

> *(...) this result [Shannon's formula] was discovered independently by several researchers, and serves as an illustration of a scientific concept whose time had come.*
>
> —Lars Lundheim [22]

This can be contrasted to the above citation of R. McEliece.

Wiener's independent derivation [17] of Shannon's formula is certainly the one that is closest to Shannon's. He also used probabilistic arguments, logarithmic measures (in base 2) and differential entropy, the latter choice being done "mak[ing] use of a personal communication of J. von Neumann". Wiener considers "the information gained by fixing one or more variables in a problem", e.g., fixing $Y = X + Z$ where X and Z are independent Gaussian. By computing the difference $h(X) - h(X|Y)$, he concludes that "the excess of information concerning X when we know Y is" (1). Unlike Shannon, however, his definition of information is not based on any precise communication problem. There is also no relation to Hartley's argument leading to (2).

Concerning the idea of information theory, Wiener wrote in his book Cybernetics:

> *This idea occurred 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.*
>
> —Norbert Wiener[17]

It is likely that it is the importance of Shannon's formula for which he has made an independent derivation that lead him to declare:

> *Information theory has been identified in the public mind to denote the theory of information by bits, as developed by C. E. Shannon and myself.*
>
> —Norbert Wiener[23]

J.R. Pierce comments:

> *Wiener's head was full of his own work and an independent derivation of* (1) *(...) Competent people have told me that Wiener, under the misapprehension that he already knew what Shannon had done, never actually found out.*
>
> —John R. Pierce [11]

All other independent discoveries in the year of 1948 were in fact essentially what is now referred to Hartley's rule leading to (2). Among these, the first published work in April 1948 was by the French engineer Jacques Laplume [21] from Thompson-Houston. He essentially gives the usual derivation that gives (2) for a signal amplitude range $[0, A]$. C. Earp's publication [19] in June 1948 also makes a similar derivation of (2) where the signal-to-noise amplitude ratio is expressed as a "root-mean-square ratio" for the "step modulation", which is essentially pulse-code modulation. In a footnote, Earp claims that his paper "was written in original form in October, 1946". In an another footnote at the first page, he mentions that

> *A symposium on "Recent Advances in the Theory of Communication" was presented at the November 12, 1947, meeting of the New York section of the Institute of Radio Engineers. Four papers were presented by A. G. Clavier (...); B.D. Loughlin (...); and J. R. Pierce and C. E. Shannon, both of Bell Telephone Laboratories.*
>
> —C.W. Earp [19]

André Clavier is another French engineer from LMT laboratories (subsidiary of ITT Corporation), who published "Evaluation of transmission efficiency according to Hartley's expression of information content" [18] in December 1948. He again makes a similar derivation of (2) as Earp's, expressed with root-mean-square values. As Lundheim notes [22],"it is, perhaps, strange that neither Shannon nor

Clavier have mutual references in their works, since both [2] and [18] were orally presented at the same meeting (...) and printed more than a year afterwards."

In May 1948, Stanford Goldman again re-derived (2), acknowledging that the equation "has been derived independently by many people, among them W. G. Tuller, from whom the writer first learned about it" [20]. William G. Tuller's thesis was defended in June 1948 and printed as an article in May 1949 [8]. His derivation uses again root-mean-square (rms) ratios.

> *Let S be the rms amplitude of the maximum signal that may delivered by the communication system. Let us assume, a fact very close to the truth, that a signal amplitude change less than noise amplitude cannot be recognized, but a signal amplitude change equal to noise is instantly recognizable. Then, if N is the rms amplitude of the noise mixed with the signal, there are $1 + S/N$ significant values of signal that may be determined. (...) the quantity of information available at the output of the system [is $= \log(1 + S/N)$].*
>
> —William G. Tuller [8]

In the 1949 article [8] he explains that

> *The existence of [Shannon's] work was learned by the author in the spring of 1946, when the basic work underlying this paper had just been completed. Details were not known by the author until the summer of 1948, at which time the work reported here had been complete for about eight months.*
>
> —William G. Tuller [8]

In view of this note it is perhaps not completely fair so say, following J.R. Pierce [11] (Shannon's co-author of [12]), that

> *(...) much of the early reaction to Shannon's work was either uninformed or a diversion from his aim and accomplishment. (...) In 1949, William G. Tuller published a paper giving his justification of* (1) *[8].*
>
> —John R. Pierce [11]

Considering that Tuller's work is—apart from Wiener's—the only work referenced by Shannon in [1], and that the oldest reference known (1946) is Tuller's, it should be certainly appropriate to refer to (2) as *Tuller's formula* or to (1) as the *Tuller–Shannon formula*.

There is perhaps no better conclusion for this section than to cite Shannon's 1949 article [2] where he explicitly mentioned (and criticized) Hartley's Law as the property that the maximum amount of information per second is proportional to the bandwidth (without reference to noise limitation), and where he proposed his own interpretation of (2) making the link with his formula (1):

How many different signals can be distinguished at the receiving point in spite of the perturbations due to noise? A crude estimate can be obtained as follows. If the signal has a power P, then the perturbed signal will have a power $P+N$. The number of amplitudes that can be reasonably well distinguished is $K\sqrt{\frac{P+N}{N}}$ where K is a small constant in the neighborhood of unity depending on how the phrase "reasonably well" is interpreted. (...) The number of bits that can be sent in this time is $\log_2 M$ [$= \frac{1}{2}\log_2 K^2\Big(1+\frac{P}{N}\Big)$].

—Claude Elwood Shannon [2]

It may be puzzling to notice, as Hodges did in his historical book on A. Turing [24], that Shannon's article [2] mentioned a manuscript with a received date of 23 July, 1940! But this was later corrected by Shannon himself in 1984 (cited in [6], Reference 10):

(...) Hodges cites a Shannon manuscript date 1940, which is, in fact, a typographical error.
(...) First submission of this work for formal publication occurred soon after World War ll.

—Claude Elwood Shannon [6]

This would mean in particular that Shannon's work leading to his formula was completed in 1946, at about the same time as Tuller's.

4. Hartley's Rule yields Shannon's Formula: $C' = C$

Let us consider again the argument leading to (2). The channel input X is taking $M = 1 + A/\Delta$ values in the set $\{-A, -A+2\Delta, \ldots, A-2\Delta, A\}$, which is the set of values $(M-1-2k)\Delta$ for $k = 0, \ldots, M-1$. A maximum amount of information will be conveyed through the channel if the input values are equiprobable. Then, using the well-known formula for the sum of squares of consecutive integers, one finds:

$$P = \mathbb{E}(X^2) = \frac{1}{M}\sum_{k=0}^{n}(M-1-2k)^2 = \Delta^2\frac{M^2-1}{3}$$

Interestingly, this is the classical formula for the average power of a M-state pulse-code modulation or pulse-amplitude modulation signal, as was derived by Oliver, Pierce and Shannon in [12].

The input is mixed with additive noise Z with accuracy $\pm\Delta$. The least favorable case would be that Z follows a uniform distribution in $[-\Delta, \Delta]$. Then its average power is

$$N = \mathbb{E}(Z^2) = \frac{1}{2\Delta}\int_{-\Delta}^{\Delta} z^2\,\mathrm{d}z = \frac{\Delta^2}{3}$$

It follows that (2) takes the form of a striking identity!

$$C' = \log_2 M = \tfrac{1}{2}\log_2(1+M^2-1) = \tfrac{1}{2}\log_2\Big(1+\frac{3P}{\Delta^2}\Big) = \tfrac{1}{2}\log_2\Big(1+\frac{P}{N}\Big) = C.$$

A mathematical coincidence?

One may perhaps argue that if Tuller or others knew about such a coincidence, they would probably have followed Wiener's attitude in claiming paternity of information theory. In any case, such an identification of (1) and (2) calls for verification that Hartley's rule would in fact be "mathematically correct" as a capacity formula.

5. Hartley's Rule as a Capacity Formula

Consider the *uniform channel*, a memoryless channel with additive white noise Z with uniform density in the interval $[-\Delta, \Delta]$. If X is the channel input, the output will be $Y = X + Z$, where X and Z are independent. We assume that the input has the amplitude constraint $|X| \leqslant A$. The following calculation was proposed as a homework exercise in the excellent textbook by Cover and Thomas [25].

Theorem 1. *Assuming A/Δ is integral, the uniform channel has capacity C' given by* (2).

(If A/Δ is not integral, then the proof of the theorem shows that $C' \leqslant \log_2(1 + A/\Delta)$, yet C' cannot be obtained in closed form.)

Proof. From Shannon's coding theorem, the channel's capacity is $C = \max_X I(X;Y)$ bits per sample, where the maximum is taken over all distributions of X such that $|X| \leqslant A$, *i.e.*, with support $[-A, A]$. By expanding mutual information $I(X;Y) = h(Y) - h(Y|X)$ as a difference of differential entropies, and noting that $h(Y|X) = h(Z) = \log_2(2\Delta)$ is constant, the required capacity C' is obtained by maximizing $h(Y)$.

Now since $|X| \leqslant A$, by the triangular inequality, the output amplitude is limited by $|Y| \leqslant |X| + |Z| \leqslant A + \Delta$. Choosing $X = X^*$ to be discrete uniform taking $M = 1 + A/\Delta$ equiprobable values in the set $\{-A, -A + 2\Delta, \ldots, A - 2\Delta, A\}$, it is immediate to see that $Y = X^* + Z$ will have the uniform density over the interval $[-A - \Delta, A + \Delta]$, which is known to maximize $h(Y)$ under the constraint $|Y| \leqslant A + \Delta$. Therefore such an X^* achieve the capacity and we have $C' = \max_X h(Y) - h(Z) = \log_2(2(A + \Delta)) - \log_2(2\Delta) = \log_2(1 + A/\Delta)$. $\square$

Thus there is a sense in which the "Tuller–Shannon Formula" (2) is indeed *correct* as the capacity of a communication channel, except that the communication noise is *not* Gaussian, but uniform, and that the signal limitation is *not* on the power, but on the amplitude (as a side remark, it is interesting to mention that C' is in fact a zero-error capacity and that no coding is actually necessary to achieve it).

The analogy between the Gaussian and uniform channels can be pushed further. Both channels are memoryless and additive, with $Y = X + Z$ where X and Z are independent. Both have "additive" constraints on their inputs of the form $\Phi(X) \leqslant c$, where additivity means that $\Phi(X) \leqslant c$ and $\Phi(Z) \leqslant c'$ imply $\Phi(X + Z) \leqslant c + c'$. Specifically, in the Gaussian case, $\Phi(X) = \mathbb{E}(X^2)$ and additivity results from the fact that X and Z are uncorrelated; and in the uniform case, $\Phi(X) = |X|$ and additivity is simply a consequence of the inequality $|X + Z| \leqslant |X| + |Z|$. Also in both cases, the noise $Z = Z^*$ maximizes the differential entropy $h(Z)$ under the constraint $\Phi(Z) \leqslant c'$, and the input $X = X^*$ that maximizes mutual information $I(X;Y) = I(X; X + Z^*)$ is such that the

corresponding output $Y^* = X^* + Z^*$ also maximizes the differential entropy $h(Y)$ under the constraint $\Phi(Y) \leqslant c + c'$. When $\Phi(X) = \mathbb{E}(X^2)$ (power limitation), both Y^* and Z^* are Gaussian while for $\Phi(X) = |X|$ (amplitude limitation), both Y^* and Z^* have a uniform distribution.

Shannon used these properties for $\Phi(X) = \mathbb{E}(X^2)$ to show that under limited *power*, Gaussian noise is the *worst* possible noise that one can inflict in the channel (in terms of its capacity). To show this, he considered an arbitrary additive noise Z and defined $\tilde{Z}$ as a random variable of the same distribution type as Z^* but with the same differential entropy as Z. Thus for $\Phi(X) = \mathbb{E}(X^2)$, $\tilde{Z}$ is a zero-mean Gaussian variable of average power $\tilde{N} = 2^{2h(Z)}/2\pi e$, which is referred to as the *entropy power* [1] of Z. He then established that the capacity associated with the noise Z satisfies [1]

$$\tfrac{1}{2}\log_2\Big(1 + \alpha\frac{P}{N}\Big) \leqslant C \leqslant \tfrac{1}{2}\log_2\Big(1 + \frac{P}{N}\Big) + \tfrac{1}{2}\log_2\alpha, \tag{3}$$

where we have noted $\alpha = N/\tilde{N}$. The first inequality was in fact derived by Shannon as a consequence of the *entropy power inequality* (see, e.g., [26] for more details on this inequality). Since $h(\tilde{Z}) = h(Z) \leqslant h(Z^*)$, one has $\tilde{N} \leqslant N$ so that $\alpha \geqslant 1$ (with equality $\alpha = 1$ only in the case of Gaussian noise). It follows from the above inequality that the capacity has the lowest value for Gaussian noise.

The uniform channel enjoys a similar property: under limited *amplitude*, *uniform* noise is the worst possible noise that one can inflict in the channel. To show this, consider the following

Definition 2 (*Entropic Amplitude*)**.** *Given an arbitrary additive noise Z, let $\tilde{Z}$ be a random variable of the same distribution type as Z^* but with the same differential entropy as Z. Thus for $\Phi(X) = |X|$, $\tilde{Z}$ is a zero-mean uniformly distributed variable with amplitude $\tilde{\Delta}$. The entropic amplitude of Z is*

$$\tilde{\Delta} = 2^{h(Z)-1}.$$

The squared entropic amplitude is related to the entropy power by the relation $\tilde{\Delta}^2 = \tilde{N}\pi e/2$.

Theorem 3. *When $\Phi(X) = |X|$ (amplitude limitation) under the same conditions as Theorem 1, the capacity C' associated with an arbitrary additive noise Z satisfies*

$$\log_2\Big(1 + \frac{A}{\Delta}\Big) \leqslant C' \leqslant \log_2\Big(1 + \frac{A}{\Delta}\Big) + \log_2\alpha, \tag{4}$$

where $\alpha = \Delta/\tilde{\Delta} \geqslant 1$ (with equality $\alpha = 1$ only for uniform noise).

It follows as announced that the capacity has the lowest value for uniform noise.

Proof. One has $I(X; X+Z) = h(X+Z) - h(Z)$ where $h(Z) = \log_2(2\tilde{\Delta})$; since $|Y| \leqslant A + \Delta$, $h(Y) \leqslant \log_2(2(A+\Delta))$. Therefore, $I(X; X+Z) \leqslant \log_2(2(A+\Delta)) - \log_2(2\tilde{\Delta}) = \log_2\Big(1 + \frac{A}{\Delta}\Big) + \log_2\alpha$. Maximizing $I(X; X+Z)$ over the distribution de X in this inequality gives the second inequality in (4).

To prove the first inequality, notice that $C = \max_X I(X; X+Z) \geqslant I(X^*; X^*+Z) = h(X^*+Z) - h(Z)$ where, as above, X^* is discrete uniform in the M-ary set $\mathcal{X} = \{-A, -A+2\Delta, \ldots, A-2\Delta, A\}$

with $M = 1 + A/\Delta$. Now $Y = X^* + Z$ follows the density $p_Y(y) = \frac{1}{M}\sum_{x\in\mathcal{X}} p_Z(y-x)$ where $p_Z(z)$ is the density of Z. Since $|Z| \leqslant \Delta$ all terms in this sum have disjoint supports. Therefore,

$$h(X^*+Z) = -\sum_{x\in\mathcal{X}}\int_{-\Delta}^{\Delta}\Big(\frac{1}{M}p_Z(y-x)\Big)\log_2\Big(\frac{1}{M}p_Z(y-x)\Big)\,\mathrm{d}y = \log_2 M - \int p_Z(z)\log_2 p_Z(z)\,\mathrm{d}z$$

which reduces to the simple formula $h(X^* + Z) = \log_2 M + h(Z)$. Therefore, $C \geqslant h(X^* + Z) - h(Z) = \log_2 M = \log_2\Big(1 + \frac{A}{\Delta}\Big)$, which proves the first inequality in (4). □

6. A Mathematical Analysis

6.1. Conditions for Shannon's Formula to Hold

In this section, we consider a memoryless additive noise channel with zero-mean input X and output $Y = X + Z$. Such a channel is defined by:

- the probability density function (pdf) p_Z of the zero-mean noise Z, which is assumed independent of X;
- a constraint set $\mathcal{C}$ on the possible distributions of X. The channel capacity is computed under this constraint as

$$C = \max_{X\in\mathcal{C}} I(X;Y) = \max_{X\in\mathcal{C}} h(Y) - h(Z) = \big(\max_{X\in\mathcal{C}} h(X+Z)\big) - h(Z).$$

We let X^* be the input that attains this maximum and let $Y^* = X^* + Z$ be the corresponding output. Thus $C = h(Y^*) - h(Z) = h(X^* + Z) - h(Z)$. We also let $P = \mathbb{E}(X^{*2})$ and $N = \mathbb{E}(Z^2)$ so that P/N denotes the signal-to-noise ratio at the optimum.

Lemma 4. *If there exists a number $\alpha > 1$ such that αZ and Y^* share the same distribution, then the channel capacity C is given by Shannon's Formula* (1)*.*

Proof. One has $C = h(Y^*) - h(Z) = h(\alpha Z) - h(Z) = \log_2|\alpha| = \frac{1}{2}\log_2\alpha^2$. However, $P + N = \mathbb{E}(X^2) + \mathbb{E}(Z^2) = \mathbb{E}(Y^2) = \alpha^2\,\mathbb{E}(Z^2) = \alpha^2 N$ and so $\alpha^2 = 1 + P/N$. This gives (1). □

Example 1 (Gaussian channel)**.** *Here both Z and $Y^* = X^* + Z$ are zero-mean Gaussian so that the condition of the lemma is satisfied. We recover* (1) *as the classical expression for the channel capacity.*

Example 2 (uniform channel)**.** *Here both Z and $Y^* = X^* + Z$ are uniformly distributed over a centered interval so the condition of the lemma is also satisfied. This explains anew the coincidence found in the calculation of Section 4.*

In the following we note $\phi_X(\omega) = \mathbb{E}(e^{i\omega X})$, the characteristic function of any random variable X.

Lemma 5. *The condition of Lemma 4 is satisfied if and only if there exists $\alpha > 1$ such that*

$$\frac{\phi_Z(\alpha\omega)}{\phi_Z(\omega)} = \phi_{X^*}(\omega)$$

Proof. αZ and $Y^* = X^* + Z$ have the same distribution if and only if they share the same characteristic function, which is equal to $\phi_{\alpha Z}(\omega) = \phi_Z(\alpha\omega)$ and to $\phi_{Y^*}(\omega) = \phi_{X^*}(\omega)\phi_Z(\omega)$. □

In particular the above quotient must be a characteristic function of some random variable. This shows that the distribution of Z should be *divisible*.

Example 3 (Gaussian channel (continued))**.** *Here* $\alpha^2 = \frac{P+N}{N}$ *and*

$$\frac{\phi_Z(\alpha\omega)}{\phi_Z(\omega)} = \frac{e^{-\alpha^2\omega^2 N/2}}{e^{-\omega^2 N/2}} = e^{-\omega^2 P/2}$$

which the characteristic function of $X^* \sim \mathcal{N}(0, P)$.

Example 4 (uniform channel (continued))**.** *Here* $\alpha = \frac{A+\Delta}{\Delta} = M$ *is assumed integral and*

$$\frac{\phi_Z(M\omega)}{\phi_Z(\omega)} = \frac{\operatorname{sinc}(M\Delta\cdot\omega)}{\operatorname{sinc}(\Delta\cdot\omega)} = \frac{\sin(M\Delta\cdot\omega)}{M\sin(\Delta\cdot\omega)} = \frac{1}{M}(e^{-i(M-1)\omega\Delta} + e^{-i(M-3)\omega\Delta} + \cdots + e^{i(M-1)\omega\Delta})$$

where $\operatorname{sinc} x = \frac{\sin x}{x}$ *is the sine cardinal function and where the last equality is the well-known Dirichlet kernel expression. The result is the characteristic function of* X^* *which take* M *equiprobable values in the set* $\{-(M-1)\Delta, -(M-3)\Delta, \ldots, (M-3)\Delta, (M-1)\Delta\}$.

Example 5 (Cauchian channel)**.** *Let* Z *be Cauchy distributed with* $p_Z(z) = \frac{1}{\pi}\frac{a}{a^2+z^2}$*, where* $a > 0$*. Then for any* $\alpha > 0$*,*

$$\frac{\phi_Z(\alpha\omega)}{\phi_Z(\omega)} = \frac{e^{-a\alpha|\omega|}}{e^{-a|\omega|}} = e^{-a(\alpha-1)|\omega|}$$

is the characteristic function of X^**, which is Cauchy distributed with parameter* $(\alpha-1)a$*. However, in this particular case,* $P = \mathbb{E}(X^{*2}) = +\infty$ *and* $N = \mathbb{E}(Z^2) = +\infty$ *so that the signal-to-noise ratio is not defined.*

Lemma 6. *Let* p_Z *and* p_{Y^*} *be the pdf's of* Z *and* Y^**, respectively. Then* X^* *attains capacity subject to an average cost per channel use of the form* $\mathbb{E}(b(X)) \leqslant C$*, where*

$$b(x) = \mathbb{E}\Big(\log_2 \frac{p_Z(Z)}{p_{Y^*}(x+Z)}\Big). \tag{5}$$

Thus given the pdf of Y^*, (5) defines an adequate constraint set $\mathcal{C}$ so that $C = h(Y^*) - h(Z)$.

Proof. Let p_Y be the pdf of $Y = X + Z$. By the information inequality $D(p_Y \| p_{Y^*}) \geqslant 0$, we obtain

$$h(Y) \leqslant \mathbb{E}\log_2 \frac{1}{p_{Y^*}(Y)} = \mathbb{E}_X\Big(\mathbb{E}_Z \log \frac{1}{p_{Y^*}(X+Z)}\Big).$$

Therefore,

$$I(X;Y) = h(Y) - h(Z) \leqslant \mathbb{E}_X\Big(\mathbb{E}_Z \log \frac{p_Z(Z)}{p_{Y^*}(X+Z)}\Big) = \mathbb{E}(b(X))$$

Equality holds if and only if $p_Y = p_{Y^*}$, that is, when the channel capacity is attained. In this case $\max I(X;Y) = \mathbb{E}(b(X))$ should be equal to the capacity C. The assertion follows. □

Example 6 (Gaussian channel (continued))**.** *Here $Z \sim \mathcal{N}(0, N)$ and $Y^* \sim \mathcal{N}(0, P+N)$. Therefore,*

$$b(x) = \log_2 \sqrt{\frac{P+N}{N}} + \mathbb{E}\log_2 \exp\Big(\frac{(x+Z)^2}{2(P+N)} - \frac{Z^2}{2N}\Big) = C + \frac{\log_2 e}{2}\Big(\frac{x^2+N}{P+N} - 1\Big).$$

The constraint $\mathbb{E}(b(X)) \leqslant C$ is now equivalent to $\mathbb{E}(X^2) \leqslant P$ as expected.

Example 7 (uniform channel (continued))**.** *Here Z is uniformly distributed on the interval $[-\Delta, \Delta]$ and Y^* is uniformly distributed on $[-A-\Delta, A+\Delta]$ where $A = (\alpha - 1)\Delta > 0$. Therefore,*

$$b(x) = \log_2 \frac{A+\Delta}{\Delta} + \mathbb{E}\log \frac{1}{\mathbf{1}_{|x+Z| \leqslant A+\Delta}}$$

where $\mathbf{1}$ denotes the indicator function. The first term in the r.h.s. is equal to C. If $|x| \leqslant A$ then $|x+Z| \leqslant A+\Delta$ a.e. so that the second term equals $\log 1 = 0$. Otherwise, $\mathbf{1}_{|x+z| \leqslant A+\Delta}$ vanishes for z in some subinterval of $[-\Delta, \Delta]$ of positive length and the second term is infinite. Hence

$$b(x) = \begin{cases} C & \text{if } |x| \leqslant A \\ +\infty & \text{otherwise.} \end{cases}$$

The constraint $\mathbb{E}(b(X)) \leqslant C$ is equivalent to $|X| \leqslant A$ a.e. as expected.

Theorem 7. *Assume that there exists $\alpha > 1$ such that $\dfrac{\phi_Z(\alpha\omega)}{\phi_Z(\omega)}$ is a characteristic distribution and let $\mathcal{C}$ be defined by the condition $\mathbb{E}(b(X)) \leqslant C$ where*

$$b(x) = \mathbb{E}\log_2 \frac{\alpha p_Z(Z)}{p_Z((x+Z)/\alpha)}. \tag{6}$$

Then the channel capacity $C = \log_2 \alpha$ of the corresponding additive noise channel is given by Shannon's Formula (1)*.*

Proof. Apply the preceding lemmas, noting that $p_{Y^*}(y) = \frac{1}{\alpha} p_Z(\frac{y}{\alpha})$. □

6.2. B-Spline Channels of Any Degree

Equipped with Theorem 7 we can construct many additive noise channels whose capacities are given by Shannon's Formula (1).

Definition 8 (*B-spline Channel*)**.** *Let U_Δ be uniformly distributed over the interval $[-\Delta, \Delta]$ and let $d \in \mathbb{N}$. Define*

$$Z_d = U_{\Delta,0} + U_{\Delta,1} + \ldots + U_{\Delta,d}$$

where the $U_{\Delta,i}$ are independent copies of U_Δ. The (uniform) B-spline channel *of degree d is the associated additive noise channel $Y = X + Z_d$ with capacity C_d.*

For $d = 0$ one recovers the uniform channel. It is easily seen and well-known that the pdf of Z_d is the uniform B-spline function:

$$p_{Z_d}(z) = \frac{1}{2\Delta} \cdot \beta_d\Big(\frac{z}{2\Delta}\Big)$$

where β_d is the standard central B-spline [27] of order d, the $(d+1)$th convolution power of the indicator function of the interval $[-1/2, 1/2]$.

Theorem 9. *For all $d \in \mathbb{N}$ and any choice of a positive integer M, the capacity C_d of the B-spline channel of degree d under the input constraint $\mathbb{E}(b_d(X)) \leqslant C_d$ where*

$$b_d(x) = \mathbb{E}\log_2 \frac{M\beta_d\Big(\frac{Z}{2\Delta}\Big)}{\beta_d\Big(\frac{x+Z}{2M\Delta}\Big)}. \tag{7}$$

is given by Shannon's Formula (1).

Proof. Since $p_{Z_d}(z) = \frac{1}{2\Delta} \cdot \beta_d\big(\frac{z}{2\Delta}\big)$ is the $(d+1)$th convolution power of the rectangle function of the interval $[-\Delta, \Delta]$, the corresponding characteristic function is a $(d+1)$th power of a cardinal sine:

$$\phi_{Z_d}(\omega) = \mathrm{sinc}^{d+1}(\Delta \cdot \omega).$$

Let $M > 0$ be an integer. From Example 4, we have

$$\begin{aligned}\frac{\phi_{Z_d}(M\omega)}{\phi_{Z_d}(\omega)} &= \frac{\mathrm{sinc}^{d+1}(M\Delta \cdot \omega)}{\mathrm{sinc}^{d+1}(\Delta \cdot \omega)} = \Big(\frac{\sin(M\Delta \cdot \omega)}{M\sin(\Delta \cdot \omega)}\Big)^{d+1} \\ &= \Big(\frac{1}{M}\big(e^{-i(M-1)\omega\Delta} + e^{-i(M-3)\omega\Delta} + \cdots + e^{i(M-1)\omega\Delta}\big)\Big)^{d+1}.\end{aligned}$$

This is the characteristic function of the random variable

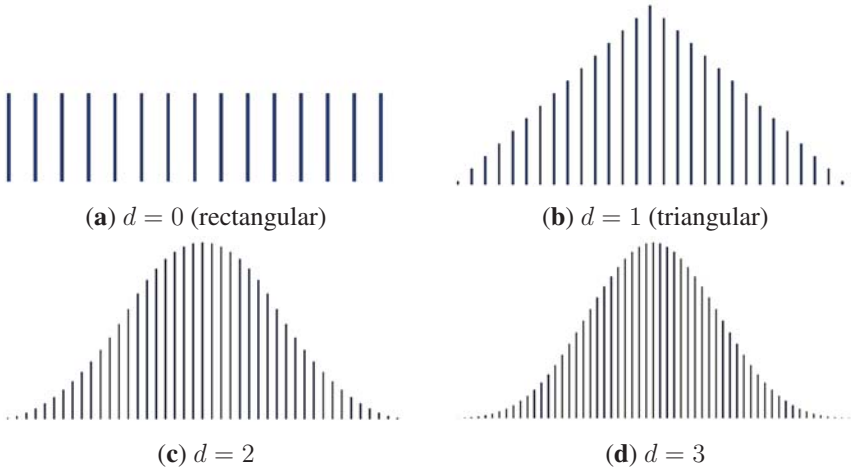
$$X_d = X_{M,0} + \cdots + X_{M,d},$$

where the $X_{M,i}$ are i.i.d. and take M equiprobable values in the set $\{-(M-1)\Delta, -(M-3)\Delta, \ldots, (M-3)\Delta, (M-1)\Delta\}$. Hence, Theorem 7 applies with $\alpha = M$ and cost function (7). □

Again for $d = 0$ one recovers the case of the uniform channel with input $X_0 = X_{M,0}$ taking M equiprobable values in the set $\{-(M-1)\Delta, -(M-3)\Delta, \ldots, (M-3)\Delta, (M-1)\Delta\}$ (Figure 1a). In general, the probability distribution of X_d is the $(d+1)$th discrete convolution power of the uniform distribution. For $d = 1$, the pdf of the noise has a triangular shape and the distribution of X_d is also triangular (Figure 1b). As d increases, it becomes closer to a Gaussian shape (Figure 1c,d).

6.3. Convergence as $d \to +\infty$

To determine the limit behavior as $d \to +\infty$, we need to apply some normalization on the probability distributions. Since the pdf of Z_d is obtained by successive convolutions of rectangles of length 2Δ, its support $[-(d+1)\Delta, (d+1)\Delta]$ as well as its average power (or variance) $N =$



(**a**) $d = 0$ (rectangular) (**b**) $d = 1$ (triangular)

(**c**) $d = 2$ (**d**) $d = 3$

Figure 1. Discrete plots of input probability distributions (of X_d) that attain capacity for $M = 15$ and different values of d.

$(d+1)\Delta^2/3$ increase linearly with $(d+1)$. Similarly, the support and average power P of X^* also increase linearly with $(d+1)$. Although this does not affect the ratio P/N, in order for average powers P and N to converge as $d \to +\infty$ we need to divide Z_d and X^*, hence their sum Y, by $\sqrt{d+1}$. The capacity will remain unaltered because

$$\begin{aligned} I\Big(\frac{X}{\sqrt{d+1}}; \frac{Y}{\sqrt{d+1}}\Big) &= h\Big(\frac{Y}{\sqrt{d+1}}\Big) - h\Big(\frac{Z}{\sqrt{d+1}}\Big) \\ &= h(Y) - \tfrac{1}{2}\log(d+1) - h(Z) + \tfrac{1}{2}\log(d+1) \\ &= h(Y) - h(Z) \\ &= I(X;Y). \end{aligned}$$

Therefore, in what follows, we assume that all random variables X, Y, Z have been normalized by the factor $\sqrt{d+1}$. We then say that the additive channel with input X_d, output Y_d, noise Z_d, and cost function $b_d(x)$ converges as $d \to +\infty$ to the additive channel with input X, output Y, noise Z, and cost function $b(x)$ if $X_d \to X$, $Y_d \to Y$, $Z_d \to Z$ in distribution, and $b_d(x) \to b(x)$.

Theorem 10. *The B-spline channel of degree d converges to the Gaussian channel as $d \to +\infty$.*

Proof. By the central limit theorem,

$$\frac{Z_d}{\sqrt{d+1}} = \frac{U_{\Delta,0} + U_{\Delta,1} + \ldots + U_{\Delta,d}}{\sqrt{d+1}}$$

converges in distribution to the Gaussian $Z \sim \mathcal{N}(0, N)$ (in fact, the B-spline pdf converges uniformly to the Gaussian pdf) [27]. Since Y_d has the same distribution as $M \cdot Z_d$, it also converges in distribution to the Gaussian $Y \sim \mathcal{N}(0, P+N)$. Again by the central limit theorem,

$$\frac{X^*}{\sqrt{d+1}} = \frac{X_0^* + \cdots + X_d^*}{\sqrt{d+1}}$$

converges in distribution to the Gaussian $\mathcal{N}(0,P)$. Finally, we can write

$$b_d(x) = \mathbb{E}\Big(\log_2 \frac{Mp_{Z_d}(Z_d)}{p_{Z_d}(\frac{x+Z_d}{M})}\Big) = \mathbb{E}\Big(\log_2 \frac{p_{Z_d}(Z_d)}{p_Z(Z_d)}\Big) - \mathbb{E}\Big(\log_2 \frac{p_{Z_d}(\frac{x+Z_d}{M})}{p_Z(\frac{x+Z_d}{M})}\Big) + \mathbb{E}\Big(\log_2 \frac{Mp_Z(Z_d)}{p_Z(\frac{x+Z_d}{M})}\Big)$$

The first term in the r.h.s. tends to zero by the strengthened central limit theorem of Barron [28] in relative entropy. The second term also tends to zero by a similar argument and change of variable. By a calculation identical to that of Example 6, the third term is equal to

$$\log_2 M + \mathbb{E}\log_2 \exp\Big(\frac{(x+Z_d)^2}{2(P+N)} - \frac{Z_d^2}{2N}\Big) = C + \frac{\log_2 e}{2}\Big(\frac{x^2+N}{P+N} - 1\Big) = b(x)$$

which shows that $b_d(x) \to b(x)$ as $d \to +\infty$. $\square$

Figure 2 shows the graphs of the cost functions $b_d(x)$ for different values of degree d. As the degree increases, the curves converge to the parabola that represents the quadratic cost function $b(x)$ for the Gaussian channel.

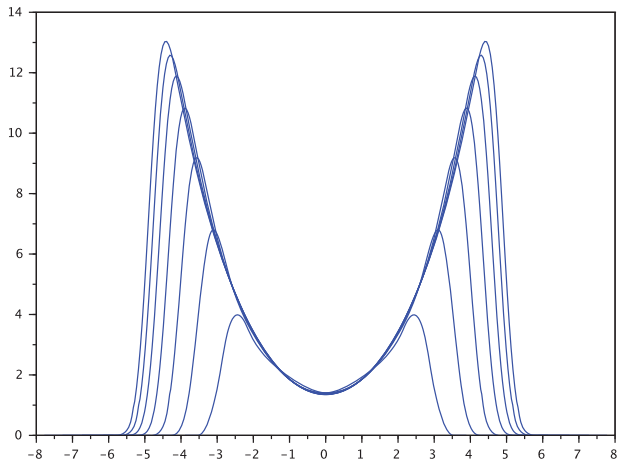


Figure 2. Cost functions $b_d(x)$ for $d = 1$ to 7 (with $M = 4$ and $\Delta = 1$). Convergence holds to the quadratic cost function $b(x)$.

Thus we have built a sequence of additive noise "B-spline" channels indexed by $d \in \mathbb{N}$ that makes the transition from the uniform ($d = 0$) to the Gaussian channel ($d \to \infty$). Shannon's Formula (1) holds for all these channels.

Acknowledgments

The authors wish to thank Max H. M. Costa for valuable discussions and suggestions. This work was partially supported by São Paulo Research Foundation (FAPESP) Grant # 2014/13835-6, under

the FAPESP thematic project *Segurança e Confiabilidade da Informação: Teoria e Prática*, Grant # 2013/25977-7.

Author Contributions

Both authors performed the historical research. Olivier Rioul wrote the paper and carried out the mathematical analysis. Both authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. Reprinted in C.E. Shannon and W. Weaver *The Mathematical Theory of Communication*; University Illinois Press: Champaign, IL, USA, 1949.
2. Shannon, C.E. Communication in the presence of noise. *Proc. Inst. Radio Eng.* **1949**, *37*, 10–21.
3. Butzer, P.; Dodson, M.; Ferreira, P.; Higgins, J.; Lange, O.; Seidler, P.; Stens, R. Multiplex signal transmission and the development of sampling techniques: The work of Herbert Raabe in contrast to that of Claude Shannon. *Appl. Anal.* **2011**, *90*, 643–688.
4. Wikipedia: Shannon–Hartley theorem. Available online: http://en.wikipedia.org/wiki/Shannon-Hartley_theorem (accessed on 28 August 2014).
5. Hartley, R.V.L. Transmission of information. *Bell Syst. Tech. J.* **1928**, *7*, 535–563.
6. Ellersick, F.W. A conversation with Claude Shannon. *IEEE Commun. Mag.* **1984**, *22*, 123–126.
7. Wozencraft, J.M.; Jacobs, I.M. *Principles of Communication Engineering*; John Wiley & Sons: New York, NY, USA, 1965; pp. 2–5.
8. Tuller, W.G. Theoretical limitations on the rate of transmission of information. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1948; Reprinted in *Proc. Inst. Radio Eng.* **1949**, *37*, 468–478.
9. Cherry, E.C. A history of information theory. *Proc. Inst. Elect. Eng.* **1951**, *98*, 383–393.
10. McEliece, R.J. *The Theory of Information and Coding*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2002.
11. Pierce, J.R. The early days of information theory. *IEEE Trans. Inf. Theory* **1973**, *19*, 3–8.
12. Oliver, B.; Pierce, J.; Shannon, C.E. The Philosophy of PCM. *Proc. Inst. Radio Eng.* **1948**, *36*, 1324–1331.
13. Gallager, R. Claude E. Shannon: A retrospective on his life, work, and impact. *IEEE Trans. Inf. Theory* **2001**, *47*, 2681–2695.
14. Verdú, S. Fifty years of Shannon theory. *IEEE Trans. Inf. Theory* **1998**, *44*, 2057–2078.
15. Golay, M.J.E. Note on the theoretical efficiency of information reception with PPM. *Proc. Inst. Radio Eng.* **1949**, *37*, 1031.

16. Slepian, D. Information theory in the fifties. *IEEE Trans. Inf. Theory* **1973**, *19*, 145–148.
17. Wiener, N. Time series, Information and Communication. In *Cybernetics*; John Wiley & Sons: New York, NY, USA, 1948; Chapter III, pp. 10–11.
18. Clavier, A.G. Evaluation of transmission efficiency according to Hartley's expression of information content. *Electron. Commun. ITT Tech. J.* **1948**, *25*, 414–420.
19. Earp, C.W. Relationship between rate of transmission of information, frequency bandwidth, and signal-to-noise ratio. *Electron. Commun. ITT Tech. J.* **1948**, *25*, 178–195.
20. Goldman, S. Some fundamental considerations concerning noise reduction and range in radar and communication. *Proc. Inst. Radio Eng.* **1948**, *36*, 584–594.
21. Laplume, J. Sur le nombre de signaux discernables en présence du bruit erratique dans un système de transmission à bande passante limitée. *Comptes rendus de l'Académie des Sciences de Paris* **1948**, *226*, 1348–1349. (In French)
22. Lundheim, L. On Shannon and "Shannon's Formula". *Telektronikk* **2002**, *98*, 20–29.
23. Wiener, N. What is information theory? *IRE Trans. Inf. Theory* **1956**, *2*, 48.
24. Hodges, A. *Alan Turing: The Enigma*; Simon and Schuster: New York, NY, USA, 1983; p. 552.
25. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2006.
26. Rioul, O. Information theoretic proofs of entropy power inequalities. *IEEE Trans. Inf. Theory* **2011**, *57*, 33–55.
27. Unser, M.; Aldroubi, A.; Eden, M. On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Trans. Inf. Theory* **1992**, *38*, 864–872.
28. Barron, A.R. Entropy and the central limit theorem. *Ann. Probab.* **1986**, *14*, 336–342.