L'Information, Séminaire Poincaré XXIII (2018) 43 – 77 **Séminaire Poincaré**

This is IT: A Primer on Shannon's Entropy and Information

Olivier Rioul
LTCI, Télécom ParisTech
Université Paris-Saclay
75013 Paris, France
CMAP
École Polytechnique
Université Paris-Saclay
91128 Palaiseau, France

I didn't like the term 'information theory'. Claude [Shannon] *didn't like it either. You see, the term 'information theory' suggests that it is a theory about information—but it's not. It's the transmission of information, not information. Lots of people just didn't understand this.*

Robert Fano, 2001

Abstract. What is Shannon's information theory (IT)? Despite its continued impact on our digital society, Claude Shannon's life and work is still unknown to numerous people. In this tutorial, we review many aspects of the concept of entropy and information from a historical and mathematical point of view. The text is structured into small, mostly independent sections, each covering a particular topic. For simplicity we restrict our attention to one-dimensional variables and use logarithm and exponential notations log and exp without specifying the base. We culminate with a simple exposition of a recent proof (2017) of the entropy power inequality (EPI), one of the most fascinating inequality in the theory.

1 Shannon's Life as a Child

Claude Elwood Shannon was born in 1916 in Michigan, U.S.A., and grew up in the small town of Gaylord. He was a curious, inventive, and playful child, and probably remained that way throughout his life. He built remote-controlled models and set up his own barbed-wire telegraph system to a friend's house [48]. He played horn and clarinet, and was interested in jazz. He was especially passionate about intellectual puzzles, riddles, cryptograms, gadgets and juggling.

He entered the university of Michigan at age 16, where he studied both electrical engineering and mathematics. He would later describe his information theory as "the most mathematical of the engineering sciences"[46].

2 A Noble Prize Laureate

Shannon graduated in 1936. He found an internship position at MIT as an assistant programmer for the "differential analyzer"—an analog machine to solve differential equations up to the sixth order—under the supervision of Vannevar Bush, who would become his mentor. Relay switches control the machine, which brings Shannon to a systematic study of the relay circuits. Using his mathematical knowledge, he established the link between circuits and the symbolic formalism of the Boolean algebra. At only 21, his master thesis [40] revolutionized the use of logic circuits by founding digital circuit design theory. It was described as "possibly the most important, and also the most famous, master's thesis of the century"[22].

For his master's work, Shannon received the Alfred Noble prize in 1940. This prize is an award presented by the American Society of Civil Engineers, and has no connection to the better known Nobel Prize established by Alfred Nobel. But Shannon's masterpiece was yet to come: *Information theory*—for which he certainly would have deserved the genuine Nobel Prize.

3 Intelligence or Information?

Shannon's PhD thesis [41], defended at MIT in 1940, develops an algebra applied to genetics. But without much contact with practitioners of this discipline, his thesis was never published and remained relatively unknown. It must be noted that immediately after receiving his degree he went to work for the Bell telephone laboratories. At this time, Shannon's major concern was what he called "the transmission of intelligence"—what will become later the theory of information. In a letter to Vannevar Bush dated February 16, 1939, he wrote:

> *Off and on I have been working on an analysis of some of the fundamental properties of general systems for the transmission of intelligence, including telephony, radio, television, telegraphy, etc.* [...] *There are several other theorems at the foundation of communication engineering which have not been thoroughly investigated.* [24]

Shannon read the works of Harry Nyquist [34] and Ralph Hartley [25], published in the late 1920s in the *Bell System Technical Journal*, the specialized research journal of the Bell Laboratories. Nyquist had written about the "transmission of intelligence by telegraph" and Hartley's 1928 paper is entitled "transmission of information." Their works will have a decisive influence on Shannon's information theory.

4 Probabilistic, not Semantic

So what is information? Shannon spent ten years (1939–1948), most of it during wartime effort at Bell Laboratories, of intense reflexion about this notion. During this period, he did not publish a single article on the subject—except for a classified memorandum on cryptography in 1945 [42]. He actually used the term 'communication theory', not 'information theory' in most of his work, and first coined the term 'uncertainty' for what would later become Shannon's 'entropy'. The term 'information' or rather 'mutual information' as a mathematical notion in the theory appeared only in the early 1950s in Robert Fano's seminars at MIT [17].

Shannon first deliberately removed the *semantic* questions from the engineering task. A famous paragraph at the very beginning of his seminal 1948 paper reads:

> *"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design."* [43]

Thus, Shannon models the information source as a *probabilistic* device that chooses among possible messages. A message (a sequence of symbols) is a realization of a stochastic process, like a Markov process. In summary, for Shannon, information is probabilistic, not semantic.

Of course, Shannon never said that the semantic aspects are not important. The concept of human intelligence is certainly not purely computational or probabilistic. This perhaps explains why Shannon preferred the term 'communication theory' over 'information theory'.

5 The Celebrated 1948 Paper

Shannon eventually published *A Mathematical Theory of Communication*, in two parts in the July and October issues of Bell System technical journal [43]. As his Bell Labs colleague John Pierce once put it, this paper "came as a bomb—something of a delayed-action bomb"[35]. It is one of the most influential scientific works that was ever published. Few texts have had such an impact in our modern world.

The paper is at the border between engineering and mathematics. At the time, it was not immediately understood by all: On the one hand, most engineers did not have enough mathematical background to understand Shannon's theorems. On the other hand, some mathematicians had trouble grasping the context of communications engineering and found it "suggestive throughout, rather than mathematical," according to the probabilist Joe Doob [13].

The theory is presented in complete form in this single article. It entirely solves the problems of data compression and transmission, providing the fundamental limits of performance. For the first time, it is proved that reliable communications must be essentially digital.

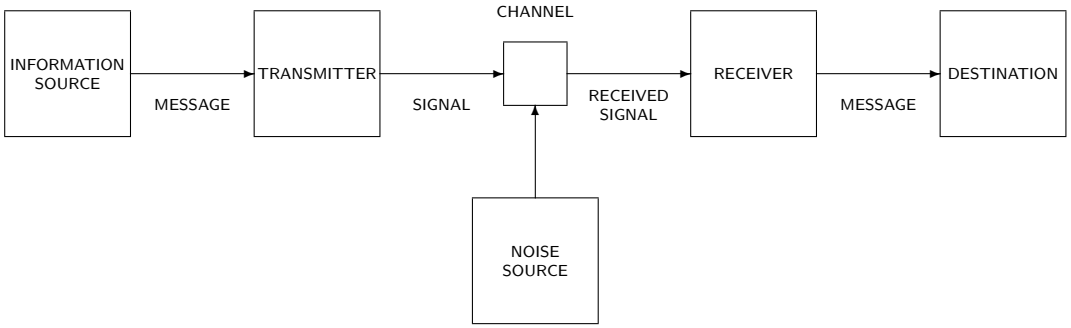

Figure 1: Shannon's paradigm, the mother of all models (redrawn from [43]).

Perhaps the most influential part of Shannon's work in all sciences is summarized in the first figure of his 1948 paper: A schematic diagram of a general communication system reproduced in Figure 1, which was called "the mother of all models" in some scientific circles [27]. In this figure, an information source is transmitted over a noisy channel and then received by a recipient. While this scheme seems quite natural today, it was revolutionary: For the first time, we clearly distinguish the roles of source, channel and recipient; transmitter and receiver; signal and noise.

6 Shannon, not Weaver

At the instigation of Shannon's employer Warren Weaver, the 1948 article is republished as a book [44] the following year, preceded by an introductory exposition of Weaver. On this occasion, Shannon's text receives some corrections and some references are updated. But the change that is both the most innocuous and the most important concerns the title: *A mathematical theory of communication* becomes *The mathematical theory of communication*.

Weaver's text, "Recent contributions to the theory of communication," is one of the many contributions to the diffusion of the theory to the general public. A condensed form was published the same year in the popular journal *Scientific American* [57]. Driven by great enthusiasm, Weaver attempts to explain how Shannon's ideas could extend well beyond his initial goals, to all sciences that address communication problems in the broad sense—such as linguistics and social sciences. Weaver's ideas, precisely because they precede Shannon's text in the book, had a tremendous impact: It is likely that many readers came up with the theory while reading Weaver and stopped at Shannon's first mathematical statements. Even today, the theory is sometimes attributed to Weaver as much as to Shannon, especially in the social sciences. Weaver is often cited as the first author, if not the only author of information theory. It is of course a misinterpretation to attribute the theory to Weaver as well. As Weaver himself declared,

> *"No one could realize more keenly than I do that my own contribution to this book is infinitesimal as compared with Shannon's."* [31]

7 Shannon, not Wiener

Norbert Wiener, the father of cybernetics, has somewhat influenced Shannon. Shannon took Wiener's course in Fourier analysis at MIT [47] and read his wartime classified report "The interpolation, extrapolation and smoothing of stationary time series"[58]. The report is primarily concerned with the linear prediction and filtering problems (the celebrated Wiener filter) but also has some formulation of communication theory as a statistical problem on time series. It was later known to generations of students as the *yellow peril* after its yellow wrappers and the fact that it full of mathematical equations that were difficult to read. Shannon was kind enough to acknowledge that "communication theory is heavily indebted to Wiener for much of its basic philosophy" and that his "elegant solution of the problems of filtering and prediction of stationary ensembles has considerably influenced the writer's thinking in this field."

However, it should be noted that never in Wiener's writings does any precise communication problem appear, and that his use of the term 'information' remained quite loose and not driven by any practical consideration. In his book *Cybernetics* [59], also published in 1948, Wiener deals with the general problems of communication and control. In the course of one paragraph, he considers "the information gained by fixing one or more variables in a problem" and concludes that "the excess of information concerning X when we know Y" is given by a formula identical in form to Shannon's best known formula $\frac{1}{2}\log(1+P/N)$ (see § 33). However, his definition of information is not based on any precise communication problem.

Wiener's prolix triumphalism contrasts with Shannon's discretion. It is likely that the importance of Shannon's formula $\frac{1}{2}\log(1 + P/N)$ for which he had made an independent derivation led 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.* [60]

John Pierce comments:

> *Wiener's head was full of his own work and an independent derivation of* $[\frac{1}{2}\log(1 + P/N)]$. *Competent people have told me that Wiener, under the misapprehension that he already knew what Shannon had done, never actually found out.* [35]

8 Shannon's Bandwagon

In the 1950s, Shannon-Weaver's book made an extraordinary publicity. As a result, information theory has quickly become a fashionable field like cybernetics or automation. But, as Shannon himself reckoned, this popularity "carries at the same time an element of danger". While its hard core is essentially a branch of mathematics, the use of exciting words like information, entropy, communication, had led many scientists to apply it indiscriminately to diverse areas such as fundamental physics, biology, linguistics, psychology, economics and other social sciences. So much so that Shannon, in a 1956 editorial entitled "The Bandwagon"[45], warns against the excesses of such popularity:

> [Information theory] *has perhaps been ballooned to an importance beyond its actual accomplishments.* [...] *The subject of information theory has certainly been sold, if not oversold. We should now turn our attention to the business of research and development at the highest scientific plane we can maintain.* [45]

So let us now turn our attention to mathematics.

9 An Axiomatic Approach to Entropy

Entropy is perhaps the most emblematic mathematical concept brought by Shannon's theory. A well-known derivation of Shannon's entropy [43] follows an axiomatic approach where one first enunciates a few desirable properties and then derives the corresponding mathematical formulation. This offers some intuition about a "measure of information". Several variants are possible based on the following argument.

Consider any event with probability p. How should behave the corresponding amount of information $i(p)$ as a function of p? First, the event should bring all the more information as it is unlikely to occur; second, independent events should not interfere, the corresponding amounts of information simply add up. Therefore, two desirables properties are:

(a) $i(p) \geq 0$ is a decreasing function of p;

(b) for any two independent events with probabilities p and q, $i(pq) = i(p) + i(q)$.

Here $i(p)$ can also be interpreted as a measure of "surprise", "unexpectedness", or "uncertainty" depending on whether the event has or has not yet occurred.

Let n be a positive integer and r be the rank of the first significant digit of p^n so that $10^{-r} \geq p^n \geq 10^{-(r+1)}$. Applying (a) and (b) several times we obtain $r \cdot i(1/10) \leq n \cdot i(p) \leq (r+1)\, i(1/10)$, that is,

$$\frac{r}{n} \leq c \cdot i(p) \leq \frac{r}{n} + \frac{1}{n}, \tag{1}$$

where c is constant independent of r and n. Now since the function $\log(1/p)$ satisfies the same properties (a), (b) above, it also satifies

$$\frac{r}{n} \leq c' \cdot \log\frac{1}{p} \leq \frac{r}{n} + \frac{1}{n}, \tag{2}$$

where c' is another constant independent of r and n. It follows from (1) and (2) that

$$\left| c \cdot i(p) - c' \cdot \log\frac{1}{p} \right| \leq \frac{1}{n}. \tag{3}$$

Letting $n \to +\infty$ we obtain that $i(p)$ is proportional to $\log(1/p)$, where the constant of proportionality can be arbitrary. Since the choice of the constant amounts to specifying the base of the logarithm (see § 10), we can simply write

$$i(p) = \log\frac{1}{p}. \tag{4}$$

Now consider the case of a random variable X with probability distribution $p(x)$. The amount of information of an elementary event $X = x$ is then $\log\frac{1}{p(x)}$. Therefore, the *average amount of information* about X is given by the expected value:

$$H(X) = \sum_x p(x) \log\frac{1}{p(x)}. \tag{5}$$

This is Shannon's *entropy* $H(X)$ of the random variable X having distribution $p(x)$.

The notation $H(X)$ may be a little confusing at first: This is not a function of X but rather of its probability distribution $p(x)$. Some authors write $H(p)$ in place of $H(X)$ to stress the dependence on the probability distribution $p(x)$.

One often sees the equivalent formula

$$H(X) = -\sum_x p(x) \log p(x) \tag{6}$$

which is essentially a matter of taste. Note, however, that since probabilities $p(x)$ lie between 0 and 1, the above expression is *minus* the sum of *negative* quantities, whereas (5) is simply the sum of positive quantities.

10 Units of Information

The base of the logarithm in (5) can be chosen freely. Since a change of base amounts to a multiplication by a constant, it specifies a certain *unit* of information.

Suppose, for example, that X takes M equiprobable values $x = 0, 1, 2, \ldots, M-1$ so that $p(x) = 1/M$ in (5). Then Shannon's entropy is simply the logarithm of the number of possible values:

$$H(X) = \log M. \tag{7}$$

If the values x are expressed in base 10, a randomly chosen m-digit number between 0 and $10^n - 1$ corresponds to $M = 10^n$. With a logarithm to base 10, the entropy is simply the number $m = \log_{10} M$ of decimal digits. Similarly, a randomly chosen m-digit number in base 2 (between 0 and $2^n - 1$) gives an entropy of m binary digits. This generalizes to any base.

With the emergence of computers, the base 2 is by far the most used in today's technology. Accordingly, the entropy is often expressed with a logarithm to base 2. The corresponding unit is the *bit*, a contraction of binary digit. Thus M possible values correspond to $\log_2 M$ bits. It was Shannon's 1948 paper [43] that introduced the word *bit* for the very first time—a word widely used today.

While a bit (a binary digit) is either 0 or 1, the entropy $H(X)$, expressed in bits, can take any positive value. For example, $\log_2 3 = 1.58496\ldots$ bits. Here the word bit (as a unit of information) can be thought of as the contraction of "binary unit" rather than of "binary digit". Similarly, with base 10, the unit of information is the *dit* (decimal unit). For natural logarithms to base e, the unit of information is the *nat* (natural unit).

To illustrate the difference between binary digit and binary unit, consider one random bit $X \in \{0, 1\}$. This random variable X follows a Bernoulli distribution with some parameter p. Its entropy, expressed in bits, is then

$$H(X) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} \quad \text{(in bits)} \tag{8}$$

which can take any value between 0 bit and 1 bit. The maximum value 1 bit is attained in the equiprobable case $p = 1/2$. Otherwise, the entropy of one bit is actually less than one bit.

The *Système International d'unités* [61] recommends the use of the *shannon* (Sh) as the information unit in place of the *bit* to distinguish the amount of information from the quantity of data that may be used to represent this information. Thus according to the SI standard, $H(X)$ should actually be expressed in shannons. The entropy of one bit lies between 0 and 1 Sh.

11 H or Êta?

In information theory, following Shannon, the entropy is always denoted by the letter H. Where does this letter come from?

Ralph Hartley was perhaps Shannon's greatest influence, and he had already used the letter H—arguably his last name initial—as early as 1928 to denote the "amount of information" [25] with a formula identical to (7). Therefore, since Shannon generalized Hartley's measure of information, it seems logical that he would have adopted Hartley's letter H. In fact, Shannon did not at first use the name "entropy" for H but rather "uncertainty"[42]. All this seems to have nothing to do with the notion of entropy in physics.

Later Shannon adopted the term "entropy" [43] and mentioned that (5) is formally identical with Boltzmann's entropy in statistical mechanics, where $p(x)$ is the

probability of a system being in a given cell x of its phase space. In fact, the very same letter H is used in Boltzmann's H-theorem to denote the *negative* continuous entropy (see § 15):

$$H = \int f \ln f \, \mathrm{d}^3 v, \tag{9}$$

where f denotes a distribution of particle velocities v. Boltzmann himself used the letter E at first [2], and it has been suggested that the first occurrence of the letter H in a paper by Burbury [4] was for "Heat"[30]. There is some indirect evidence, however, that in this context, H is in fact the capital greek letter H (Êta), the uppercase version of η, but the reason for which this choice was made is mysterious [28]. It does not seem to relate to the etymology of entropy, a term coined by Clausius [8] from the Greek $\varepsilon\nu\tau\rho o\pi\acute{\eta}$ ("inside transformation").

12 No One Knows What Entropy Really Is

Since the information-theoretic measure of information H is named *entropy* with reference to Boltzmann's entropy in statistical thermodynamics, the big question is: Is there a deep-lying connection between information theory and thermodynamics?

It is clear that the Shannon entropy is identical in form with previous expressions for entropy in statistical mechanics. The celebrated Boltzmann's entropy formula $S = k \log W$, where log denotes the natural logarithm and k is Boltzmann's constant equal to $1.3806485\ldots 10^{-23}$ joules per kelvin, can be identified with (7) where $M = W$ is the number of microstates of a system in thermodynamic equilibrium. The integral version of entropy (with a minus sign) also appears in Boltzmann's first entropy formula $S = -\int \rho \ln \rho \, \mathrm{d}x$ where the probability distribution ρ represents the fraction of time spent by the system around a given point x of its space phase. Von Neumann's 1932 entropy formula $S = -\mathrm{Tr}(\hat{D} \log \hat{D})$ in quantum statistical mechanics [56] is also formally identical with (5) where $p(x)$ represent the eigenvalues of the density operator $\hat{D}$.

It is quite striking that such a strong formal analogy holds. Thus, although Shannon's information theory is certainly more mathematical than physical, any mathematical result derived in information theory could be useful when applied to physics with the appropriate interpretation.

Beyond the formal analogy, many physicists soon believed that a proper understanding of the second law of thermodynamics requires the notion of information. This idea can be traced back to Leó Szilárd [51] who attempted in the 1929 to solve Maxwell's demon problem by showing that an entropy decrease of $k \log 2$ per molecule is created by *intelligence* (the exactly informed Maxwell's demon). This was later recognized as the measure of *information* acquired by the demon, the term $k \log 2$ being identified with one "bit" of information. Szilárd was a personal friend of John von Neumann who derived his entropy formula a few years later. It is plausible that when von Neumann discovered Shannon's "information" formula, he immediately made the link with his entropy. In 1961, Shannon told Myron Tribus that von Neumann was the one who told him to call his new formula by the name 'entropy' in the early 1940s. According to Tribus, Shannon recalled:

> *"My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'.*

> *When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name. In the second place, and more importantly, no one knows what entropy really is, so in a debate you will always have the advantage.'"* [53]

When asked twenty years later about this anecdote, however, Shannon did not remember von Neumann giving him such advice [47].

Norbert Wiener was also influenced by von Neumann who suggested to him the entropy formula $\int f(x) \log f(x)\, \mathrm{d}x$ as a "reasonable" measure of the amount of information associated with the curve $f(x)$ [59]. Shannon may have first come across the notion of entropy from Wiener, which was one of Shannon's teachers at MIT. Robert Fano, one of Shannon's colleagues at Bell Labs who worked on information theory in the early years, reported that when he was a PhD student at MIT, Wiener would at times enter his office, puffing at a cigar, saying "You know, information is entropy" [18]. Later the French-American physicist Léon Brillouin, building on Szilárd's and Shannon's works, coined the concept of *negentropy* to demonstrate the similarity between entropy and information [3].

Despite many attempts in the literature, it is still not clear why information theoretic principles should be necessary to understand statistical mechanics. Is there any physical evidence of a fundamental thermodynamic cost for the physical implementation of a informational computation, purely because of its logical properties? This is still a debated topic today [52]. As in the above Neumann-Shannon anecdote, no one knows what entropy really is.

13 How Does Entropy Arise Naturally?

Going back to mathematics, Shannon's entropy as a mathematical quantity arises in a fairly natural way. Shannon proposed the following line of reasoning [43]: Consider a long sequence of independent and identically distributed (i.i.d.) outcomes

$$\underline{x} = (x_1, x_2, \ldots, x_n). \tag{10}$$

To simplify, assume that the symbols x_i take a finite number of possibles values. Let $p(x)$ denote the probability that an outcome equals x. Thus each outcome follows the same probability distribution $p(x)$, of some random variable X.

By independence, the probability $p(\underline{x})$ of the long sequence $\underline{x}$ is given by the product

$$p(\underline{x}) = p(x_1)p(x_2)\cdots p(x_n). \tag{11}$$

Re-arrange factors according to the number $n(x)$ of x_i equal to x to obtain

$$p(\underline{x}) = \prod_x p(x)^{n(x)}, \tag{12}$$

where the product is over all possible outcome values x. Since n is taken very large, according to the law of large numbers, the empirical frequency of x can be identified to its probability:

$$\frac{n(x)}{n} \approx p(x). \tag{13}$$

Pugging this expression into (12) gives

$$p(\underline{x}) \approx \Big(\prod_x p(x)^{p(x)}\Big)^n = \exp\big(-nH(X)\big) \tag{14}$$

which exponentially decreases as $n \to +\infty$. The exponential decay $H(X) \geq 0$ is precisely given by Shannon's entropy (5). It is impressive to observe how this entropy arises "out of nowhere" in such a simple derivation.

Shannon's equation (14) is a fundamental result known as the *asymptotic equipartition property*: Essentially, this means that for very large (but fixed) n, the value of the probability of a given "typical" sequence $\underline{x} = (x_1, x_2, \ldots, x_n)$ is likely to be close to the constant $\exp(-nH(X))$. Moreover, any randomly chosen sequence is very likely to be "typical" (with probability arbitrarily close to one). As we have seen in the above derivation, this is essentially a consequence of the law of large numbers.

The fact that $p(\underline{x}) \approx \exp(-nH(X))$ for any typical sequence turns out to be very useful to solve the problem of information compression and other types of coding problems. This is perhaps the main mathematical justification of the usefulness of Shannon's entropy in science.

14 Shannon's Source Coding Theorem

The asymptotic equipartition property (14) for typical sequences is used to solve the *information compression* problem: How can we reliably encode a source of information with the smallest possible rate? This corresponds to the model of Figure 1 where the channel is noiseless and the information source is to be transmitted reliably at the destination while achieving the maximum possible compression.

Since non-typical sequences have a arbitrarily small probability, an arbitrary reliable compression is obtained by encoding typical sequences only. The resulting coding rate is computed from the number N of such typical sequences. Summing (14) over all the N typical sequences gives the total probability that a randomly chosen sequence is typical, which we know is arbitrarily close to one if the length n is taken sufficiently large:

$$1 \approx N \exp(-nH(X)). \tag{15}$$

This gives $N \approx \exp(nH(X))$ typical sequences. The resulting coding rate R is its logarithm per element in the sequence of length n:

$$R = \frac{\log N}{n} \approx H(X). \tag{16}$$

This is the celebrated *Shannon's first coding theorem* [43]: The minimal rate at which a source X can be encoded reliably is given by its entropy $H(X)$.

This important theorem provides the best possible performance of any data compression algorithm. In this context, Shannon's entropy receives a striking operational significance: It is the minimum rate of informational bits in a source of information.

15 Continuous Entropy

So far Shannon's entropy (5) was defined for discrete random variables. How is the concept generalized to continuous variables? An obvious way is to proceed by analogy. Definition (5) can be written as an expectation

$$H(X) = \mathbb{E} \log \frac{1}{p(X)}, \tag{17}$$

where $p(x)$ is the discrete distribution of X. When X follows a continuous distribution (pdf) $p(x)$, we may define its *continuous entropy* with formally the same formula:

$$h(X) = \mathbb{E} \log \frac{1}{p(X)} \tag{18}$$

which is

$$h(X) = \int p(x) \log \frac{1}{p(x)} \, \mathrm{d}x. \tag{19}$$

The discrete sum in (5) is simply replaced by an integral (a continuous sum).

Notice, however, that $p(x)$ does not refer to a probability anymore in (19), but to a probability *density*, which is not the same thing. For example, when the continuous random variable U is uniformly distributed over the interval (a, b), one has $p(u) = 1/(b-a)$ so that (18) becomes

$$h(U) = \mathbb{E} \log \frac{1}{p(U)} = \log(b-a). \tag{20}$$

While the discrete entropy is always nonnegative, the above continuous entropy expression becomes negative when the interval length is < 1. Moreover, taking the limit as the interval length tends to zero, we have

$$h(c) = -\infty \tag{21}$$

for any *deterministic* (constant) random variable $X = c$. This contrasts with the corresponding discrete entropy which is simply $H(c) = 0$.

Therefore, contrary to Shannon's entropy (5) for discrete variables, one cannot assign an "amount of information" to the continuous entropy $h(X)$ since it could be negative. Even though Shannon himself used the letter H for both discrete and continuous entropies [43], the capital H was soon degraded by information theorists to the lowercase letter h to indicate that the continuous entropy does not deserve the status of the genuine entropy H.

16 Change of Variable in the Entropy

In order to better understand why discrete and continuous entropies behave differently, consider $Y = T(X)$ with some invertible transformation T. If X is a discrete random variable, so is Y; the variables have different values but share the same probability distribution. Therefore, their discrete entropies coincide: $H(X) = H(Y)$. It is obvious, in this case, that X and Y should carry the same amount of information.

When X and $Y = T(X)$ are continuous random variables, however, their continuous entropies do *not* coincide. In fact, assuming T satisfies the requirements for

an invertible change of variable (i.e., a diffeomorphism) with $\frac{\mathrm{d}y}{\mathrm{d}x} = T'(x) > 0$, the relation $p(x)\,\mathrm{d}x = \widetilde{p}(y)\,\mathrm{d}y$ gives

$$h(T(X) = h(Y) = \int \widetilde{p}(y) \log \frac{1}{\widetilde{p}(y)}\,\mathrm{d}y \tag{22}$$

$$= \int p(x) \log \frac{\mathrm{d}y/\,\mathrm{d}x}{p(x)}\,\mathrm{d}x \tag{23}$$

$$= \int p(x) \log \frac{1}{p(x)}\,\mathrm{d}x + \int p(x)T'(x)\,\mathrm{d}x, \tag{24}$$

hence the change of variable formula [43]:

$$h(T(X)) = h(X) + \mathbb{E} \log T'(X). \tag{25}$$

The difference $h(T(X)) - h(X)$ depends on the transformation T. For $T(x) = x + c$ where c is constant, we obtain

$$h(X + c) = h(X), \tag{26}$$

so the continuous entropy is invariant under shifts. For a linear transformation $T(x) = sx$ $(s > 0)$, however, we obtain the following *scaling property*:

$$h(sX) = h(X) + \log s. \tag{27}$$

Since $s > 0$ is arbitrary, the continuous entropy can take arbitrarily large positive or negative values, depending on the choice of s. For sufficiently small s (or sufficiently small variance), $h(X)$ becomes negative as we already have seen in the case of the uniform distribution (20).

17 Discrete vs. Continuous Entropy

Beyond the analogy between the two formulas, what is the precise relation between *discrete* entropy $H(X) = \sum p(x) \log \frac{1}{p(x)}$ and *continuous* entropy $h(X) = \int p(x) \log \frac{1}{p(x)}\,\mathrm{d}x$? To understand this, let us consider a *continuous* variable X with continuous density $p(x)$ and the corresponding *discrete* variable $[X]$ obtained by *quantizing* X with small quantization step δ. This means that we have a relation of the form

$$[X] = \delta \Big\lfloor \frac{X}{\delta} \Big\rfloor, \tag{28}$$

where $\lfloor \cdot \rfloor$ denotes the integer part. How $h(X)$ can be written in terms of $H([X])$? The integral (19) defining $h(X)$ can be approximated as a Riemann sum:

$$h(X) \approx \sum_k p(x_k) \log\Big(\frac{1}{p(x_k)}\Big) \delta x_k, \tag{29}$$

where $x_k = k\delta$, $\delta x_k = \delta$ and the approximation holds for small values of δ. Since the probability of a quantized value $[X] = k$ is

$$p(k) = \int_{k\delta}^{(k+1)\delta} p(x)\,\mathrm{d}x \approx p(x_k)\delta, \tag{30}$$

we obtain

$$h(X) \approx \sum_k p(k) \log\Big(\frac{\delta}{p(k)}\Big) = H([X]) - \log \frac{1}{\delta}. \tag{31}$$

This gives the desired relation between discrete and continuous entropies:

$$h(X) \approx H([X]) - \log \frac{1}{\delta}. \tag{32}$$

As $\delta \to 0$, $[X]$ converges to X but $H([X])$ does *not* converge to $h(X)$: In fact, $H([X]) - h(X) = \log(1/\delta)$ goes to $+\infty$. This confirms that discrete and continuous entropies behave very differently.

From (32), the continuous entropy $h(X)$, also known as *differential* entropy, is obtained as the limit of a *difference*. In particular, when X is deterministic, $H([X]) = 0$ and we recover that $h(X) = -\infty$ in this particular case. When X is a continuous random variable with finite differential entropy $h(X)$, since the limit is finite as $\delta \to 0$ and $\log(1/\delta) \to +\infty$, it follows that the discrete entropy $H([X])$ should actually diverge to $+\infty$:

$$\Big\{h(X) \text{ is finite}\Big\} \implies \Big\{H([X]) \to +\infty\Big\}. \tag{33}$$

This is not surprising in light of Shannon's first coding theorem (16): An arbitrarily fine quantization of a continuous random variable requires a arbitrarily high precision and therefore, an infinite coding rate.

18 Most Beautiful Equation

It has long been said that the most beautiful mathematical equation is Euler's identity $e^{i\pi} + 1 = 0$, because it combines the most important constants in mathematics like π and e together. Here's another one that should perhaps be considered equally beautiful. Let $X^* \sim \mathcal{N}(0, 1)$ by the standard normal, with density

$$p(x^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^{*2}}{2}}. \tag{34}$$

This of course is a fundamental distribution in mathematics and in physics, the limit of the well-known central limit theorem. Its entropy is easily computed from (18) as

$$h(X^*) = \mathbb{E} \log\big(\sqrt{2\pi} e^{X^{*2}/2}\big) \tag{35}$$

$$= \log \sqrt{2\pi} + \log e \cdot \mathbb{E}(X^{*2})/2. \tag{36}$$

Here $\mathbb{E}(X^{*2}) = 1$ since the standard normal has zero mean and unit variance. We obtain

$$\boxed{h(X^*) = \log \sqrt{2\pi e}}, \tag{37}$$

a lovely formula that combines the three most important real constants in mathematics: $\sqrt{2}$ (diagonal of a unit square), π (circumference of a circle with unit diameter), and e (base of natural logarithms).

The more general case where $X^* \sim \mathcal{N}(\mu, \sigma^2)$ follows a Gaussian distribution with mean μ and variance σ^2 is obtained by multiplying the standard variable by σ and adding μ. From (26) and (27) we obtain

$$h(X^*) = \log \sqrt{2\pi e} + \log \sigma = \frac{1}{2} \log(2\pi e \sigma^2). \tag{38}$$

19 Entropy Power

Shannon advocated the use of the *entropy power* rather than the entropy in the continuous case [43]. Loosely speaking, the entropy power is defined as the power of the noise having the same entropy. Here the noise considered is the most common type of noise encountered in engineering, sometimes known as "thermal noise", and modeled mathematically as a zero-mean Gaussian random variable X^*. The (average) noise power N^* is the mean squared value $\mathbb{E}(X^{*2})$ which equals the variance of X^*. By (38) its entropy is

$$h(X^*) = \frac{1}{2}\log(2\pi e N^*) \tag{39}$$

so that

$$N^* = \frac{\exp\big(2h(X^*)\big)}{2\pi e}. \tag{40}$$

The entropy power $N(X)$ of a continuous random variable X is, therefore, the power N^* of the noise X^* having the same entropy $h(X^*) = h(X)$. This gives

$$N(X) = \frac{\exp\big(2h(X)\big)}{2\pi e}. \tag{41}$$

Interestingly, it turns out that the "entropy *power*" is essentially a constant raised to the *power* of the (continuous) entropy. Thus, the physicist's view of entropy power uses the notion of power in physics while the mathematician's view refers to the notion of power in the mathematical operation of exponentiation.

When X is itself zero-mean Gaussian, its entropy power equals its actual power $\mathbb{E}(X^2)$. In general, X is not necessarily Gaussian, but the entropy power still satisfies some properties that one would expect for a power: It is a positive quantity, with the following scaling property:

$$N(aX) = a^2 N(X) \tag{42}$$

which is an immediate consequence of (27).

20 A Fundamental Information Inequality

A fundamental inequality, first derived by Gibbs in the 19th century [23], is sometimes known as the *information inequality* [9, Thm. 2.6.3]: For any random variable X with distribution $p(x)$,

$$\mathbb{E}\log\frac{1}{p(X)} \le \mathbb{E}\log\frac{1}{q(X)}, \tag{43}$$

where the expectation is taken with respect to p, and where $q(x)$ is any other probability distribution. Equality holds if and only if distributions p and q coincide. The left-hand side of (43) is the discrete or continuous entropy, depending on whether the variable X is discrete or continuous. Thus

$$H(X) = \sum_x p(x)\log\frac{1}{p(x)} \le \sum_x p(x)\log\frac{1}{q(x)} \tag{44}$$

when X is discrete with probability distribution $p(x)$ and

$$h(X) = \int p(x) \log \frac{1}{p(x)} \, \mathrm{d}x \leq \int p(x) \log \frac{1}{q(x)} \, \mathrm{d}x \tag{45}$$

when X is continuous with probability density $p(x)$. Notice that the right-hand side is always identical to the left-hand side except for the distribution inside de logarithm.

Gibbs' inequality (43) is an easy consequence of the concavity of the logarithm. By Jensen's inequality, the difference between the two sides of (43) is

$$\mathbb{E} \log \frac{q(X)}{p(X)} \leq \log \mathbb{E} \frac{q(X)}{p(X)} = \log 1 = 0. \tag{46}$$

Indeed, $\mathbb{E} \frac{q(X)}{p(X)} = \sum_x \frac{q(x)}{p(x)} p(x) = \sum_x q(x) = 1$ in the discrete case, and $\mathbb{E} \frac{q(X)}{p(X)} = \int \frac{q(x)}{p(x)} p(x) \, \mathrm{d}x = \int q(x) \, \mathrm{d}x = 1$ in the continuous case. Because the logarithm is strictly concave, equality in Jensen's inequality holds if and only if $q(x)/p(x)$ is constant, which implies that the two distributions $p(x)$ and $q(x)$ coincide.

The fundamental information inequality (43) is perhaps the most important inequality in information theory because, as seen below, every classical information-theoretic inequality can be easily derived from it.

21 The MaxEnt Principle

The maximum entropy (MaxEnt) principle first arose in statistical mechanics, where it was shown that the maximum entropy distribution of velocities in a gas under the temperature constraint is the Maxwell-Boltzmann distribution. The principle has been later advocated by Edwin Jaynes for use in a general context as an attempt to base the laws of thermodynamics on information theory [29]. His "MaxEnt school" uses Bayesian methods and has been sharply criticized by the orthodox "frequentist" school [14]. In an attack against the MaxEnt interpretation, French mathematician Benoît Mandelbrot once said: "Everyone knows that Shannon's derivation is in error"[54]. Of course, as we now show, Shannon's mathematical derivation is mathematically correct. Only physical misinterpretations of his calculations could perhaps be questionable.

Consider the following general maximum entropy problem: Maximize the (discrete or continuous) entropy over all random variables satisfying a constraint of the form $\mathbb{E}\{w(X)\} = \alpha$, where $w(x)$ is a some given weight function. A classical approach to solving the problem would use the Lagrangian method, but a much simpler derivation is based on Gibbs' inequality (43) as follows.

Consider the "exponential" probability distribution

$$q(x) = \frac{e^{-\lambda w(x)}}{Z(\lambda)}, \tag{47}$$

where $Z(\lambda)$ is a normalizing factor, known in physics as the canonical partition function, and λ is chosen so as to meet the constraint $\mathbb{E}\{w(X)\} = \alpha$. Plugging (47)

into (43) gives an upper bound on the discrete or continuous entropy:

$$\begin{aligned} H(X) \text{ or } h(X) \quad &\leq \mathbb{E}\log\big(Z(\lambda)e^{\lambda w(X)}\big) && (48)\\ &= \log Z(\lambda) + (\log e)\lambda\,\mathbb{E}\{w(X)\} && (49)\\ &= \log Z(\lambda) + \alpha\lambda\log e. && (50)\end{aligned}$$

The entropy's upper bound has now become constant, independent of the probability distribution $p(x)$ of X. Since equality (43) holds if and only if $p(x)$ and $q(x)$ coincide, the upper bound (50) is attained precisely when $p(x)$ is given by (47). Therefore, $\log Z(\lambda) + \alpha\lambda\log e$ is in fact the desired value of the maximum entropy. The above method can be easily generalized in the same manner to more than one constraint.

This general result, sometimes known as the *Shannon bound*, can be applied to many important problems. First, what is the maximum entropy of a discrete random variable that can take at most M values? Set $w(x) = 0$, $\alpha = 0$ so that $Z(\lambda) = M$ where the actual value of λ is of no importance. Then

$$\max H(X) = \log M \tag{51}$$

attained for a uniform distribution. Thus (7) is the maximum uncertainty, when all outcomes are equally probable: One event cannot be expected in preference to another. This is the classical assumption in the absence of any prior knowledge.

Similarly, what is the maximum entropy of a continuous random variable having values in a finite-length interval $[a, b]$? Again set $w(x) = 0$, $\alpha = 0$ so that $Z(\lambda) = b-a$, the interval length. Then

$$\max_{X\in[a,b]} h(X) = \log(b-a). \tag{52}$$

Thus (20) is the maximum entropy, attained for a uniform distribution on $[a, b]$.

More interestingly, what is the maximum entropy of a continuous variable with fixed mean μ and variance σ^2? Set $w(x) = (x-\mu)^2$, then $\alpha = \sigma^2$, $\lambda = 1/2\sigma^2$, $Z(\lambda) = \sqrt{2\pi/\sigma^2}$, hence the maximum entropy is $\log\sqrt{2\pi\sigma^2} + (1/2)\log e$:

$$\max_{\mathsf{Var}\,X=\sigma^2} h(X) = \frac{1}{2}\log(2\pi e\sigma^2) \tag{53}$$

attained for a normal $\mathcal{N}(\mu, \sigma^2)$ distribution. In other words, (38) is the maximum entropy for fixed variance. When X is zero mean, $\sigma^2 = \mathbb{E}(X^2)$ is its power, hence the entropy power (41) cannot exceed the actual power, which is attained if and only if X is a Gaussian random variable.

The fact that the Gaussian (normal) distribution maximizes the entropy for fixed first and second moments is of paramount importance in many engineering methods, such as Burg's spectral estimation method [5].

22 Relative Entropy or Divergence

The fundamental information inequality (43) gives rise to a new informational measure which is in many respects even more fundamental than the entropy itself. Let X be distributed according to the distribution $p(x)$ and let X^* be distributed according

to another distribution $q(x)$. Then the difference between the two sides of (43) is the *relative entropy*

$$D(X, X^*) = \mathbb{E} \log \frac{p(X)}{q(X)} \geq 0, \tag{54}$$

often noted $D(p, q)$ to stress the dependence on the two distributions[1]. This is also known as the Kullback-Leibler *divergence* $D(p, q)$ between the two distributions p and q [33].

In contrast to Shannon's entropy, the relative entropy is positive in both cases of discrete or continuous distributions:

$$D(X, X^*) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0 \tag{55}$$

for discrete probability distributions p, q and

$$D(X, X^*) = \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \geq 0 \tag{56}$$

for probability density functions p, q. In addition, the relative entropy $D(X, X^*) = D(p, q)$ vanishes if and only if equality holds in (43), that is, when p and q coincide. Therefore, $D(X, X^*) = D(p, q)$ can be seen as a measure of "informational distance" relative to the two distributions p and q. Notice, however, that the above expressions are *not* symmetrical in (p, q).

Furthermore, if we consider continuous random variables X, X^* and quantize them as in (28) with small quantization step δ to obtain discrete random variables $[X], [X^*]$, then the $\log \frac{1}{\delta}$ term present in § 17 cancels out on both sides of (45) and we obtain

$$D([X], [X^*]) \to D(X, X^*) \qquad \text{as } \delta \to 0. \tag{57}$$

Thus, as the discrete random variables $[X], [X^*]$ converge to the continuous ones X, X^*, their relative entropy $D([X], [X^*])$ similarly converge to $D(X, X^*)$. This important feature of divergence allows one to deduce properties for continuous variables from similar properties derived for discrete variables.

Finally, in the MaxEnt principle described in § 21, letting $q(x)$ (the distribution of X^*) be the entropy-maximizing distribution (47), we see that the right-hand side of Gibbs' inequality (43) equals the maximum entropy of X^*. Therefore, in this case, the divergence is simply the difference between the entropy and its maximum value:

$$D(X, X^*) = \begin{cases} H(X^*) - H(X) & \text{in the discrete case,} \\ h\,(X^*) \;-\; h(X) & \text{in the continuous case.} \end{cases} \tag{58}$$

For example, for M-ary variables, $D(X, X^*) = H(X^*) - H(X) = \log M - H(X)$ can be seen as a measure of redundancy: It is the amount of rate reduction performed by an optimal coding scheme according to Shannon's source coding theorem (§ 14). When X and X^* are continuous variables with the same variance σ^2, X^* is normally distributed and $D(X, X^*) = h(X^*) - h(X)$ represents the "non-Gaussianity" of the random variable X, which vanishes if and only if X is Gaussian.

[1] It has now become common practice for information theorists to adopt the notation $D(p\|q)$ with a double vertical bar. The origin of such an unusual notation seems obscure.

23 Generalized Entropies and Divergences

There exist numerous generalizations of Shannon's entropy and relative entropy. In 1960, Alfréd Rényi looked for the most general definition of information measures that would preserve the additivity of independent events [37]. The Rényi entropy is defined for discrete random variables as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_x p(x)^\alpha \tag{59}$$

and the continuous version is accordingly $h_\alpha(X) = \frac{1}{1-\alpha} \log \int p(x)^\alpha \, \mathrm{d}x$. One recovers Shannon's entropy by letting $\alpha \to 1$. The most interesting special cases are $\alpha = 0$ (the max-entropy), $\alpha = \infty$ (the min-entropy) and $\alpha = 2$ (the collision entropy). There is also a Rényi α-divergence $D_\alpha(p,q) = \frac{1}{\alpha-1} \log \sum_x p(x)^\alpha / q(x)^{\alpha-1}$. Rényi entropies have found many applications such as source coding, hypothesis testing, channel coding, guessing, quantum information theory, and computer science.

The Tsallis entropy, first introduced by Havrda and Charvát [26], was proposed as a basis for generalizing the standard Boltzmann-Gibbs statistical mechanics [55]. Since then its physical relevance has been debated [7]. It is defined as

$$\frac{1}{\alpha-1}\big(1 - \sum_x p(x)^\alpha\big) = \frac{1 - \exp\big((1-\alpha)H_\alpha(X)\big)}{\alpha - 1} \tag{60}$$

with the continuous version $\frac{1}{\alpha-1}\big(1 - \int p(x)^\alpha \, \mathrm{d}x\big)$. Again Shannon's entropy is recovered by letting $\alpha \to 1$.

All these entropies and relative entropies have been further extensively generalized as *f-divergences* [11] for some convex function f. Instances of f-divergences are: relative entropy (Kullback-Leibler divergence), Rényi and Tsallis divergences, the Hellinger distance, the Jensen-Shannon divergence, Vajda divergences including the total variation distance and the Pearson (χ^2) divergence, etc. There is abundant literature on such generalized concepts and their applications in signal processing, statistics and information theory.

24 How Does Relative Entropy Arise Naturally?

Similarly as in § 13, the relative entropy receives a useful operational justification. Going back to the expression (12) for the probability of a sequence $\underline{x}$ of n independent outcomes, and letting

$$q(x) = \frac{n(x)}{n} \tag{61}$$

be the empirical probability of the sequence $\underline{x}$ (also referred to as its *type*), we can rewrite (14) as an *exact* expression

$$p(\underline{x}) = \Big(\prod_x p(x)^{q(x)}\Big)^n = \exp\big(-nH(X,X^*)\big), \tag{62}$$

where $H(X,X^*)$ is the so-called *cross-entropy*

$$H(X,X^*) = \sum_x q(x) \log \frac{1}{p(x)}. \tag{63}$$

Since a typical sequence is characterized by its type q, the probability that a randomly chosen sequence is typical is exactly $N\exp(-nH(X,X^*))$, where N is the number of typical sequences. Thus in particular $N\exp(-nH(X,X^*)) \leq 1$ for any choice of p, and in particular for $p=q$ we have $N\exp(-nH(X^*,X^*)) \leq 1$, that is, $N \leq \exp(-nH(X^*))$. Since

$$H(X^*,X^*) - H(X^*) = \sum_x q(x)\log\frac{1}{p(x)} - \sum_x q(x)\log\frac{1}{q(x)} = D(q,p), \tag{64}$$

the probability that a randomly chosen sequence is typical (according to the actual probability distribution p) is bounded by

$$N\exp(-nH(X,X^*)) \leq \exp(-nD(q,p)), \tag{65}$$

where $D(q,p) = D(X^*,X) \geq 0$ is the relative entropy or divergence. Therefore, if $q(x)$ diverges from $p(x)$, the exponent $D(q,p)$ is strictly positive and the probability (65) can be made exponentially small.

Juste like the asymptotic equipartition property (14) is used to solve the *information compression* problem (Shannon's source coding theorem in § 14), the above asymptotic "large deviation" bound (65) will be used in § 32 to solve the *information transmission* problem (Shannon's channel coding theorem).

25 Chernoff Information

Derivations similar to the above in the preceding section (§ 24) form the basis of the *method of types*, a powerful technique in large deviations theory. More generally, there is a strong relationship between information theory and statistics, and the Kullback-Leibler divergence (54) has become a fundamental tool for solving many problems in statistics.

For example, in the problem of testing hypotheses, the Kullback-Leibler divergence is used to derive the best possible error exponents for tests to decide between two alternative i.i.d. distributions p and q. In a Bayesian approach where we assign prior probabilities to both hypotheses, the exponent of the overall probability error is given by

$$D(X_\lambda, X) = D(X_\lambda, X^*), \tag{66}$$

where X follows p, X^* follows q, and X_λ follows a distribution r_λ proportional to $p^\lambda q^{1-\lambda}$. Here $\lambda \in [0,1]$ is chosen such that equality (66) holds, which gives the maximum error exponent.

The common value (66) is known as the *Chernoff information* $C(X,X^*)$. Just as for the Kullback-Leibler divergence, it was derived in the early 1950s [6]. An easy calculation shows that

$$C(X,X^*) = \max_\lambda\big(\lambda D(X_\lambda,X) + (1-\lambda)D(X_\lambda,X^*)\big) \tag{67}$$

$$= \max_\lambda \mathbb{E}\log\frac{r_\lambda(X_\lambda)}{p^\lambda(X_\lambda)q^{1-\lambda}(X_\lambda)} \tag{68}$$

$$= -\min_\lambda \log\sum_x p^\lambda(x)q^{1-\lambda}(x). \tag{69}$$

Such an information measure is symmetric in (p, q), positive and vanishes if and only if the two distributions p and q coincide. Today, Chernoff information plays an important role as a statistical distance for various data processing applications.

26 Fisher Information

In statistical parametric estimation, the concept of *information* was already explored by Ronald Fisher in the 1920s [20], following an early work of Edgeworth [15] forty years before Shannon. Loosely speaking, Fisher's information measures the amount of information about a parameter θ in an observed random variable X, where X is modeled by a probability density $p_\theta(x)$ that depends on θ.

To understand the significance of the Fisher information, consider an *estimator* of θ, that is, some function $\hat{\theta}(X)$ of the observed random variable X that is used to estimate the value of θ. An optimal estimator would minimize the mean-squared error (MSE), given by

$$\mathrm{MSE} = \mathbb{E}\big((\hat{\theta}(X) - \theta)^2\big) = \int (\hat{\theta}(x) - \theta)^2 \, p_\theta(x) \, \mathrm{d}x. \tag{70}$$

Suppose, for simplicity, that the estimator is *unbiased*, i.e., its bias is zero for any value of θ:

$$\mathrm{Bias} = \mathbb{E}(\hat{\theta}(X) - \theta) = \int (\hat{\theta}(x) - \theta) \, p_\theta(x) \, \mathrm{d}x = 0. \tag{71}$$

Taking the derivative with respect to θ, we obtain

$$1 = \int (\hat{\theta}(x) - \theta) \frac{\partial p_\theta}{\partial \theta}(x) \, \mathrm{d}x \tag{72}$$

$$= \int (\hat{\theta}(x) - \theta) S_\theta(x) p_\theta(x) \, \mathrm{d}x \tag{73}$$

$$= \mathbb{E}\big((\hat{\theta}(X) - \theta) S_\theta(X)\big), \tag{74}$$

where

$$S_\theta(x) = \frac{\frac{\partial p_\theta}{\partial \theta}(x)}{p_\theta(x)} = \frac{\partial \log p_\theta}{\partial \theta}(x) \tag{75}$$

is known as the *score* or *informant*, the derivative of the log-likelihood with respect to θ. Now, by the Cauchy-Schwarz inequality,

$$1 = \big\{ \mathbb{E}\big((\hat{\theta}(X) - \theta) S_\theta(X)\big) \big\}^2 \le \mathbb{E}\big((\hat{\theta}(X) - \theta)^2\big) \cdot \mathbb{E}\big(S_\theta(X)^2\big), \tag{76}$$

where

$$J_\theta(X) = \mathbb{E}\big(S_\theta(X)^2\big) = \mathbb{E}\Big(\Big(\frac{\partial \log p_\theta}{\partial \theta}(X)\Big)^2\Big) \tag{77}$$

is the *Fisher information*. The above inequality now writes

$$\mathrm{MSE} \ge \frac{1}{J_\theta(X)}. \tag{78}$$

This is the celebrated *Cramér-Rao inequality* derived by Fréchet, Darmois, Rao and Cramér in the early 1940s [21, 12, 36, 10]. This inequality states that a universal

lower bound on the mean-squared error of any unbiased estimator is given by the reciprocal of the Fisher information. In other words, the larger amount of information $J_\theta(X)$ about θ, the more reliable its (unbiased) estimation can be.

Despite appearances, there is a strong relationship between Fisher's and Shannon's concepts of information. In fact, Fisher's information can be expressed in terms of the relative entropy (divergence)

$$D(p_\theta, p_{\theta'}) = \mathbb{E} \log \frac{p_\theta(X)}{p_{\theta'}(X)} = \int p_\theta(x) \log \frac{p_\theta(x)}{p_{\theta'}(x)} \, \mathrm{d}x. \tag{79}$$

By the fundamental information inequality (43), we know that $D(p_\theta, p_{\theta'})$ is positive and vanishes when $\theta' = \theta$ (since in this case the two distributions p_θ and $p_{\theta'}$ coincide). Therefore, the derivative with respect to θ' vanishes at $\theta' = \theta$, and the second derivative is positive and represents its curvature at $\theta' = \theta$.

For the first derivative, we have

$$\frac{\partial}{\partial \theta'} D(p_\theta, p_{\theta'})\Big|_{\theta'=\theta} = -\,\mathbb{E}\, \frac{\partial}{\partial \theta'} \log p_{\theta'}(X)\Big|_{\theta'=\theta} = -\,\mathbb{E}\big(S_\theta(X)\big) = 0 \tag{80}$$

which simply means that the score has zero mean—hence the Fisher information (77) also equals the *variance* of the score. That $\mathbb{E}\big(S_\theta(X)\big) = 0$ can easily be checked directly since $\int p_\theta(x) \frac{\frac{\partial p_\theta(x)}{\partial \theta}}{p_\theta(x)} \, \mathrm{d}x = \int \frac{\partial p_\theta(x)}{\partial \theta} \, \mathrm{d}x = \frac{\partial}{\partial \theta} \int p_\theta(x) \, \mathrm{d}x = \frac{\partial 1}{\partial \theta} = 0$.

For the second derivative, we have

$$\frac{\partial^2}{\partial \theta'^2} D(p_\theta, p_{\theta'})\Big|_{\theta'=\theta} = -\,\mathbb{E}\, \frac{\partial^2}{\partial \theta^2} \log p_\theta(X) \tag{81}$$

which is the expected value of $-\frac{\partial^2}{\partial \theta^2} \log p_\theta(X)$, sometimes referred to as the *observed information*. Expanding $\frac{\partial^2}{\partial \theta^2} \log p_\theta(x) = \frac{\partial}{\partial \theta} \frac{\frac{\partial p_\theta}{\partial \theta}(x)}{p_\theta(x)} = \frac{\frac{\partial^2 p_\theta}{\partial \theta^2}(x)}{p_\theta(x)} - \big(\frac{\frac{\partial p_\theta}{\partial \theta}(x)}{p_\theta(x)}\big)^2 = \frac{\frac{\partial^2 p_\theta}{\partial \theta^2}(x)}{p_\theta(x)} - S_\theta(x)^2$, one finds that $\mathbb{E}\, \frac{\frac{\partial^2 p_\theta}{\partial \theta^2}(X)}{p_\theta(X)} = \int p_\theta(x) \frac{\frac{\partial^2 p_\theta(x)}{\partial \theta^2}}{p_\theta(x)} \, \mathrm{d}x = \int \frac{\partial^2 p_\theta(x)}{\partial \theta^2} \, \mathrm{d}x = \frac{\partial^2}{\partial \theta^2} \int p_\theta(x) \, \mathrm{d}x = \frac{\partial^2 1}{\partial \theta^2} = 0$. Therefore,

$$\frac{\partial^2}{\partial \theta'^2} D(p_\theta, p_{\theta'})\Big|_{\theta'=\theta} = J_\theta(X). \tag{82}$$

Thus, the Fisher information also equals the expected "observed information" (81) and can be identified to the *curvature* of the relative entropy. In fact, the second-order Taylor expansion of relative entropy about θ is

$$D(p_\theta, p_{\theta'}) = \frac{1}{2} J_\theta(X) \cdot (\theta' - \theta)^2 + o(\theta' - \theta)^2. \tag{83}$$

Thus, the more information about θ, the more "sharply peaked" is the relative entropy about its minimum at θ. This means that θ is all more sharply localized as its Fisher information is large.

27 Kolmogorov Information

We have seen in § 9 that when X is a random variable with probability distribution $p(x)$, the amount of information associated to the event $X = x$ can be defined

as $\log \frac{1}{p(x)}$. From Shannon's source coding theorem (§ 14), $\log \frac{1}{p(x)}$ represents the minimal bit length required to describe x by an optimal code. Thus, the original approach of Shannon is to base information theory on *probability theory*, which incidentally was axiomatized as a rigorous mathematical theory by Andreï Kolmogorov in the 1930s.

Kolmogorov was an ardent supporter of Shannon's information theory in the 1950s and 1960s. He went further than Shannon by defining the algorithmic *complexity* of x as the length of the shortest binary computer program that describes x. Kolmogorov proved that not only his definition of complexity is essentially computer independent, but also that the average algorithmic complexity of a random variable is roughly equal to its entropy. In this way, the Kolmogorov complexity extends Shannon's entropy while dispensing with the notion of probability distribution. In a summary of his work on complexity theory, Kolmogorov wrote:

> *"Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory have a finite combinatorial character."* [32]

The concept of Kolmogorov's information or complexity is perhaps more philosophical than practical, closely related to Turing machines, Church's thesis, universal codes, the Occam's razor principle, and Chaitin's mystical number Ω—a well-known "philosopher's stone". The reader is referred to [9, Chap. 14] for a more detailed introduction.

28 Shannon's Mutual Information

As we already have noted, Claude Shannon used the term 'communication theory' in his seminal 1948 work, not 'information theory'. However, his most important results rely on the notion of transmitted information over a communication channel. This was soon formalized by Robert Fano who coined the term 'mutual information'. Fano recalled:

> *I didn't like the term 'information theory'. Claude didn't like it either. You see, the term 'information theory' suggests that it is a theory about information—but it's not. It's the transmission of information, not information. Lots of people just didn't understand this. I coined the term 'mutual information' to avoid such nonsense: making the point that information is always about something. It is information provided by something, about something.* [19]

Thanks to the notion of divergence $D(p,q)$ (§ 22), Shannon's mutual information can be easily defined as a measure of mutual dependence between two random variables X and Y. Let $p(x,y)$ be the joint distribution of X and Y. If X and Y were independent, the joint distribution would equal the product of marginals: $p(x,y) = p(x)p(y)$. In general, however, the two distributions $p(x,y)$ and $q(x,y) = p(x)p(y)$ do not coincide. The *mutual information* $I(X;Y)$ is simply the divergence $D(p,q) = \mathbb{E}\log\frac{p(X,Y)}{q(X,Y)}$:

$$I(X;Y) = \mathbb{E}\log\frac{p(X,Y)}{p(X)p(Y)}. \tag{84}$$

It is a measure of mutual dependence as expected: By the fundamental information inequality (43), $I(X;Y) \geq 0$ with equality $I(X;Y) = 0$ if and only if $p = q$, that is, X and Y are independent.

The same definition and properties hold for both discrete or continuous random variables, thanks to the property (57) of divergence. Thus, if $[X]$ and $[Y]$ are quantized versions of continuous variables X, Y then $I([X];[Y]) \to I(X;Y)$ as the quantization step $\delta \to 0$.

Notice that although divergence $D(p,q)$ is not symmetric in (p,q), mutual information *is* symmetric in (X,Y): $I(X;Y) = I(Y;X)$, hence the term *mutual*. It was found convenient by information theorists to use the semi colon ';' as the argument separator in the mutual information with lower precedence over the comma ',' e.g., to make the distinction between mutual informations $I(X,Y;Z)$ (between (X,Y) and Z) and $I(X;Y,Z)$ (between X and (Y,Z)).

As usual for informational measures, mutual information $I(X;Y)$ does not actually depend on the real values taken by the variables X, Y, but only on their probability distributions. Thus, mutual information is well defined even for categorical variables. This can be seen as an advantage over other dependence measures like linear (or nonlinear) correlation.

As in Fano's quote above, mutual information $I(X;Y)$ can be interpreted as a measure of information *provided by* Y *about* X. To see this, rewrite (84) as

$$I(X;Y) = \mathbb{E} \log \frac{p(X|Y)}{p(X)}, \tag{85}$$

where $p(x|y) = p(x,y)/p(y)$ is the conditional distribution of X *knowing* $Y = y$. The (unconditional) distribution $p(x)$ of X (*not* knowing Y) is affected by the knowledge of Y and the corresponding average relative entropy is precisely the mutual information (85). By symmetry, $I(X;Y)$ is also a measure of information *provided by* X *about* Y.

The concept of mutual information has been generalized to more than two variables although the corresponding multivariate mutual information can sometimes be negative [17].

29 Conditional Entropy or Equivocation

As seen below, mutual information (85) is a central concept in Shannon's information theory. It can be easily related to the concept of entropy by simply rewriting (85) as

$$I(X;Y) = \mathbb{E} \log \frac{1}{p(X)} - \mathbb{E} \log \frac{1}{p(X|Y)}. \tag{86}$$

Thus mutual information is the difference between Shannon's entropy of X and a "conditional" entropy of X given Y:

$$I(X;Y) = \begin{cases} H(X) - H(X|Y) & \text{in the discrete case,} \\ h(X) - h(X|Y) & \text{in the continuous case,} \end{cases} \tag{87}$$

where the *conditional entropy*, also known as *equivocation*, is defined by

$$H(X|Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x|y)} \tag{88}$$

in the discrete case, and

$$h(X|Y) = \iint p(x,y) \log \frac{1}{p(x|y)} \, \mathrm{d}x \, \mathrm{d}y \tag{89}$$

in the continuous case. Notice that in contrast to the entropy of one variable, the above definitions of conditional entropy involve averaging over both variables X and Y. By symmetry of mutual information the variables X and Y can be interchanged in the above expressions. There is also a notion of *conditional mutual information*, e.g., $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$.

30 Knowledge Reduces Uncertainty — Mixing Increases Entropy

The conditional entropy can be written as an average value of entropies, e.g.,

$$H(X|Y) = \sum_y p(y) \sum_x p(x|y) \log \frac{1}{p(x|y)} = \sum_y p(y) H(X|Y=y). \tag{90}$$

Thus, for discrete variables, while $H(X)$ measures the uncertainty about X, $H(X|Y)$ is a measure of the average uncertainty about X *when* Y *is known*. By (87), $I(X;Y) = H(X) - H(X|Y) \geq 0$, hence *knowledge reduces uncertainty* (on average):

$$H(X|Y) \leq H(X). \tag{91}$$

The difference between the two uncertainties is precisely $I(X;Y)$, the amount of information provided by Y about X.

Similarly for continuous variables, since $I(X;Y) = h(X) - h(X|Y) \geq 0$, we still observe that *conditioning reduces entropy*:

$$h(X|Y) \leq h(X), \tag{92}$$

even though we have seen that these differential entropies cannot be interpreted as uncertainty measures.

As an interesting application of (91) consider two systems with probability distributions p_1 and p_2 and their linear mixture $p_\lambda = \lambda_1 p_1 + \lambda_2 p_2$ where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are such that $\lambda_1 + \lambda_2 = 1$. If X_1 follows p_1 and X_2 follows p_2, then $p_\lambda = \lambda_1 p_1 + \lambda_2 p_2$ can be seen as the distribution of the random variable X_λ, where $\lambda \in \{1,2\}$ is itself random with respective probabilities λ_1, λ_2. Then by (91), the entropy of the mixture satisfies

$$H(X_\lambda) \geq H(X_\lambda|\lambda) = \lambda_1 H(X_1) + \lambda_2 H(X_2). \tag{93}$$

In other words, *mixing increases entropy*: the entropy of the mixture is not less than the corresponding mixture of entropies. This can also be seen as a *concavity* property of entropy (with respect to the probability distribution), a classical statement in information theory.

It is quite fascinating to see how such exciting expressions such as "knowledge reduces uncertainty" or "mixing increases entropy" receive a rigorous treatment in information theory. This perhaps explains the extraordinary wave of popularity that Shannon's theory has experienced in the past. More suggestive results are derived in the next section.

31 A Suggestive Venn Diagram

The relationship between entropies, conditional entropies and mutual information is summarized for discrete variables in the Venn diagram of Figure 2. Using the

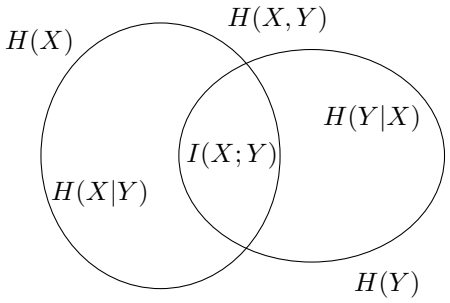

Figure 2: Venn diagram illustrating relationships among Shannon's measures of information. The mutual information $I(X;Y)$ corresponds to the intersection of the two "uncertainty sets", while the *joint entropy* $H(X,Y)$ corresponds to their union.

diagram, we recover the relations $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. In addition, many useful properties of information measures can be derived from it.

If, for example, $X = Y$ (or if X and Y are in bijection) then the two uncertainty sets coincide in Figure 2, $H(X|Y) = H(Y|X) = 0$, and

$$H(X) = I(X;X). \tag{94}$$

This means that *self-information is entropy*: $H(X)$ can be seen as the measure of information provided by X about X itself. In particular, we recover that $H(X) \geq 0$, with equality $H(X) = 0$ if and only if X is independent of itself (!) which simply means that $p(x) = 0$ or 1, i.e., X is deterministic or *certain*. This confirms the intuition that $H(X)$ is measure of randomness or uncertainty. For a continuous random variable, we would have $I(X;X) = H(X) = +\infty$ as explained in § 17.

For independent variables X and Y, $I(X;Y) = 0$ and the two sets in Figure 2 are disjoint. In this case the joint entropy is $H(X,Y) = H(X)+H(Y)$, the individual uncertainties simply add up. In the general case of dependent variables we would have $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$: the joint uncertainty is the sum of the uncertainty of one variable and the uncertainty of the other knowing the first.

At the other extreme, suppose Y is fully dependent on X so that $Y = f(X)$ where f is some deterministic function. Then $H(Y|X) = 0$ and the uncertainty Y set is contained inside the uncertainty X set in Figure 2. Therefore,

$$H\big(f(X)\big) \leq H(X). \tag{95}$$

This means that *processing reduces entropy*: any function of a random variable has the effect of decreasing its entropy.

Similarly, for any two random variables X and Y, both uncertainty sets in the Venn diagram become smaller for $f(X)$ and $g(Y)$ for any functions f and g. Therefore, we have

$$I\big(f(X);g(Y)\big) \leq I(X;Y). \tag{96}$$

In words, *data processing can only reduce information*. This is a particular instance of an important result in information theory known as the *data processing inequality* [9, Thm. 2.8.1] for Markov chains.

32 Shannon's Channel Coding Theorem

The mutual information, along with the asymptotic large deviation bound (65), can be used to solve the *information transmission* problem: How can we reliably transmit a source of information at the highest possible speed? This corresponds to the model of Figure 1 in which the information source is to be transmitted though the noisy channel at the maximum possible transmission rate while achieving an arbitrarily reliable communication.

A discrete sequence $\underline{x} = (x_1, x_2, \ldots, x_n)$ input to the channel normally corresponds to a chosen channel *code* and is not random. But Shannon had the brilliant idea to consider the whole set of *all possible codes* that may be used in the communication, and assign an (albeit artificial) probability to each code in such a way that $\underline{x}$ can be considered as a realization of an i.i.d. sequence, as if the code were chosen at random with independent code sequences. This is certainly the first application of the famous "probabilistic method" later attributed to Paul Erdős [16]. In this non-constructive method, known as *random coding*, Shannon considers the average performance over all "random" codes and deduces the existence of at least one "good" code. Roughly sketched, his argument is as follows.

Assume that the sequence $\underline{x} = (x_1, x_2, \ldots, x_n)$ is input to a memoryless noisy channel. Then the corresponding channel output $\underline{y} = (y_1, y_2, \ldots, y_n)$ has the property that it is jointly typical with the channel input $\underline{x}$ with high probability, in the sense of the law of large numbers (see § 13). Therefore, to achieve an arbitrarily reliable communication, it would be sufficient in theory to decode the received signal $\underline{y}$ by selecting the code sequence $\underline{x}$ that is jointly typical with it, provided that the actual transmitted code $\underline{x}$ is the only sequence having this property.

Now if another code sequence $\underline{x}'$ happens to be jointly typical with $\underline{y}$, since the code sequences are chosen to be independent, the actual probability distribution p of the corresponding bivariate random variable (X', Y) is $p(x)p(y)$ while its type q is roughly equal to the joint distribution $p(x, y)$. From § 24 and the asymptotic large deviation bound (65), the probability that this happens is bounded by $\exp\big(-nD(q,p)\big)$ where $D(q,p) = D(p(x,y), p(x)p(y)) = I(X;Y)$ by definition of mutual information (84). For a channel with N code sequences, the total decoding error probability P_e (averaged over all possible codes) is then bounded by

$$P_e \leq N \exp\big(-nI(X;Y)\big). \tag{97}$$

For this error probability to be exponentially small as $n \to +\infty$, it is sufficient that the transmission rate R per symbol be strictly less than the mutual information:

$$R = \frac{\log N}{n} < I(X;Y). \tag{98}$$

In order to maximize the transmission rate, the probability distribution $p(x)$ of code sequences can be chosen so as to maximize $I(X;Y)$. Shannon's channel *capacity*

$$C = \max_{p(x)} I(X;Y) \tag{99}$$

is the maximum possible amount of information transmitted over the communication channel. Thus, there exists a code achieving arbitrarily small probability of decoding error, provided that

$$R < C. \tag{100}$$

This is the celebrated *Shannon's second coding theorem* [43] which provides the best possible performance of any data transmission scheme over a noisy channel: An arbitrarily reliable communication can be achieved so long as the transmission rate does not exceed the channel's capacity.

This revolutionary theorem did change our world. For the first time, it was realized that the transmission noise does not limit the reliability of the communication, only the speed of transmission. Thus, digital communications can achieve almost perfect quality. That alone justifies that Shannon is considered as the father of the digital information age.

Somewhat paradoxically, even though Shannon's channel coding theorem is nonconstructive, it suggest that any code picked at random would be very likely to be almost optimal. However, since $n \to +\infty$ in Shannon's argument, such a code would be impossible to implement in practice. Intensive research is still undertaken today to derive good channel codes, sufficiently complex (that appear 'random') to perform well and at the same time sufficiently simple to be efficiently implemented.

33 Shannon's Capacity Formula

Perhaps the most emblematic classical expression of information theory is Shannon's capacity formula for a communication channel with additive white Gaussian noise (AWGN). The AWGN model is the basic noise model used to mimic the effect of many random processes that occur in nature, and a very good model for many practical communication links. The capacity is given by (99) where in the AWGN model,

$$Y = X + Z^*, \tag{101}$$

where $Z^* \sim \mathcal{N}(0, N)$ is a Gaussian random variable with zero mean and power N, independent of the transmitted signal X. The maximum in (99) is to be taken over distributions $p(x)$ such that X has limited power P. Here the quantity $\mathrm{SNR} = P/N$ is known as the signal-to-noise ratio (SNR). We have

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) && \text{by (87)} && (102)\\ &= h(Y) - h(Z^*) && \text{by (101)} && (103)\\ &\le h(Y^*) - h(Z^*) && \text{by (53)}, && (104) \end{aligned}$$

where Y^* is a Gaussian random variable having the same power as $Y = X + Z^*$, that is, $P + N$. The upper bound (104) is attained when $X = X^*$ is itself Gaussian, since then $Y = X^* + Z^* = Y^*$ is also Gaussian (as the sum of independent Gaussian variables). Therefore, (104) is the required capacity. From (38) we obtain

$$C = \frac{1}{2}\log\big(2\pi e(P+N)\big) - \frac{1}{2}\log(2\pi eN) = \frac{1}{2}\log(1 + P/N). \tag{105}$$

This is the celebrated Shannon's capacity formula $C = (1/2)\log(1 + \mathrm{SNR})$ that appears in Shannon's 1948 paper. It is often said that Hartley derived a similar rule

twenty years before Shannon, but in fact this is a historical misstatement [38]. However, this formula was discovered independently by at least seven other researchers in the same year 1948! [38] An illustration of a concept whose time has come.

34 The Entropy Power Inequality and a Saddle Point Property

It is well known that the power of the sum of independent zero-mean random variables equals the sum of the individual powers. For the entropy power (41), however, with have the inequality

$$N(X+Y) \geq N(X) + N(Y) \tag{106}$$

for any independent variables X and Y. This is known as the *entropy power inequality* (EPI): The entropy power of the sum of independent random variables *is not less* than the sum of the individual entropy powers.

The EPI was first stated by Shannon in his 1948 paper and is perhaps the most difficult and fascinating inequality in the theory. Shannon's 1948 proof [43] was incomplete; the first rigorous proof was given ten years later by Stam [50], with a quite involved argument based on the Fisher information. In the last section of this paper I will present a simple and recent (2017) proof from [39].

The EPI was initially used by Shannon to evaluate the channel capacity for non-Gaussian channels. It now finds many applications in information theory (to bound performance regions for multi-user source or channel coding problems) and in mathematics (e.g., to prove strong versions of the central limit theorem).

It is particularly interesting to review how Shannon used the EPI to specify the role of the Gaussian distribution for communication problems. From the preceding section (§ 33), we know that the Gaussian X^* maximizes the mutual information $I(X;Y)$ transmitted over a channel with additive Gaussian noise Z^*, i.e.,

$$I(X; X+Z^*) \leq I(X^*; X^*+Z^*) = C. \tag{107}$$

where $C = \frac{1}{2}\log(1+P/N)$. On the other hand, when the additive noise Z of power N is not necessarily Gaussian and the channel input X^* is Gaussian of power P, we have

$$\begin{aligned} I(X^*; X^*+Z) &= h(X^*+Z) - h(Z) && (108)\\ &= \frac{1}{2}\log\big(2\pi e N(X^*+Z)\big) - \frac{1}{2}\log\big(2\pi e N(Z)\big), && (109)\end{aligned}$$

where by the EPI (106), $N(X^*+Z) \geq N(X^*) + N(Z) = P + N(Z)$. Therefore,

$$I(X^*; X^*+Z) \geq \frac{1}{2}\log\big(1+P/N(Z)\big) \geq \frac{1}{2}\log\big(1+P/N\big), \tag{110}$$

where we have used that the entropy power $N(Z)$ does not exceed the actual power N. Combining this with (107) we obtain a *saddle point property of mutual information*:

$$I(X; X+Z^*) \leq \underbrace{I(X^*; X^*+Z^*)}_{C=\frac{1}{2}\log(1+P/N)} \leq I(X^*; X^*+Z). \tag{111}$$

This shows that the *Gaussian* is at the same time the *best signal* X^* (which maximizes information) and the *worst noise* Z^* (which minimizes information).

From this result, one can define a two-person (signal X and noise Z) zero-sum game with mutual information as the payoff function for which a *Nash equilibrium* holds with the Gaussian saddle point (X^*, Z^*) [1]. Similar considerations can be used to establish a certain duality between source and channel coding.

35 MaxEnt vs. MinEnt Principles

Before going through the proof the EPI (106), it is convenient to rewrite it as an extremum property of the differential entropy. Let X and Y be any two independent continuous random variables and let X^* and Y^* be zero-mean Gaussian independent variables.

From the *maximum entropy (MaxEnt) principle* (see (53) in § 21) we know that under a fixed variance constraint, the differential entropy is maximized for a Gaussian variable. Thus under the condition of identical individual *variances*:

$$\mathsf{Var}(X^*) = \mathsf{Var}(X) \quad \text{and} \quad \mathsf{Var}(Y^*) = \mathsf{Var}(Y), \tag{112}$$

the entropy of the linear combination $aX + bY$ is *maximized* for Gaussian variables:

$$h(aX + bY) \leq h(aX^* + bY^*). \tag{113}$$

This is simply a consequence of the fact that $aX^* + bY^*$ is Gaussian (as the sum of independent Gaussian variables) of the same variance as $aX + bY$.

Interestingly, the EPI (106) can be rewritten as a *minimum entropy (MinEnt) principle*: Under the condition of identical individual *entropies*:

$$h(X^*) = h(X) \quad \text{and} \quad h(Y^*) = h(Y), \tag{114}$$

the entropy of the linear combination $aX + bY$ is now *minimized* for Gaussian variables:

$$h(aX + bY) \geq h(aX^* + bY^*). \tag{115}$$

To see this, notice that from the scaling property of the entropy power (42), the EPI (106) can be rewritten as $N(aX+bY) \geq N(aX)+N(bY) = a^2N(X)+b^2N(Y)$. But from (114) we have $N(X) = N(X^*)$ and $N(Y) = N(Y^*)$. Since the entropy power of a Gaussian variable is the same as its power, $a^2N(X)+b^2N(Y) = a^2N(X^*)+ b^2N(Y^*) = N(aX^* + bY^*)$. Thus the EPI can be rewritten as $N(aX + bY) \geq N(aX^* + bY^*)$ which taking logarithms is the same as (115).

In addition, we shall see that the minimum is achieved only for Gaussian variables provided the linear combination is not trivial (a, b are non-zero scalars). This MinEnt form (115) of the EPI finds application in signal processing for blind source separation and deconvolution.

36 A Simple Proof of the Entropy Power Inequality [39]

Lastly, proceed to prove the EPI in the form (115). First, by the scaling property of entropy (27), we can always modify the constants a, b in such a way that X and Y can be assumed to have equal entropies. Then condition (114) writes

$$h(X) = h(Y) = h(X^*) = h(Y^*). \tag{116}$$

In particular, the zero-mean Gaussian variables X^* and Y^* have equal entropies, equal variances, and, therefore, identical normal distributions.

Next, applying by the scaling property of entropy (27) if necessary in (115), we can always assume that a, b have been further normalized such that $a^2 + b^2 = 1$. Then $\widetilde{X} = aX^* + bY^*$ in the right-hand side of (115) is also identically distributed as X^* and Y^*. In fact, the rotation

$$\begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix} = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \begin{pmatrix} X^* \\ Y^* \end{pmatrix} \tag{117}$$

transform i.i.d. Gaussian variables X^*, Y^* into i.i.d. Gaussian variables $\widetilde{X}, \widetilde{Y}$.

We now use the classical *inverse transform sampling* method: Let Φ be the cumulative distribution function (c.d.f.) of X^* and F be the c.d.f. of X. Then

$$\begin{aligned} F(x) &= \Phi\left(\Phi^{-1}\big(F(x)\big)\right) && (118)\\ &= \mathbb{P}\left(X^* \le \Phi^{-1}\big(F(x)\big)\right) && (119)\\ &= \mathbb{P}\big(\Phi(X^*) \le F(x)\big) && (120)\\ &= \mathbb{P}\left(F^{-1}\big(\Phi(X^*)\big) \le x\right). && (121) \end{aligned}$$

Thus letting $T = F^{-1} \circ \Phi$, the variable $T(X^*)$ has the same distribution as X, and we can write $X = T(X^*)$ in the above expressions. This so-called *transport argument* can also be applied to Y. Thus, we can always assume that

$$X = T(X^*) \quad \text{and} \quad Y = U(Y^*) \tag{122}$$

for some "transportation" functions T and U. The EPI (115) now writes

$$h\big(aT(X^*) + bU(Y^*)\big) \ge h(aX^* + bY^*), \tag{123}$$

where the right-hand side equals $h(\widetilde{X})$. By the inverse of rotation (117)

$$\begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix}, \tag{124}$$

the EPI (115) is equivalent to the inequality:

$$h\big(a\,T(a\widetilde{X} - b\widetilde{Y}) + b\,U(b\widetilde{X} + a\widetilde{Y})\big) \ge h(\widetilde{X}). \tag{125}$$

We now proceed to prove (125). Since conditioning reduces entropy (92),

$$h\big(a\,T(a\widetilde{X} - b\widetilde{Y}) + b\,U(b\widetilde{X} + a\widetilde{Y})\big) \ge h\big(a\,T(a\widetilde{X} - b\widetilde{Y}) + b\,U(b\widetilde{X} + a\widetilde{Y})\big|\widetilde{Y}\big). \tag{126}$$

Apply the change of variable formula (25) in the right-hand side, where $\widetilde{Y}$ is fixed so that $T_{\widetilde{Y}}(\widetilde{X}) = a\,T(a\widetilde{X} - b\widetilde{Y}) + b\,U(b\widetilde{X} + a\widetilde{Y})$ is a function of $\widetilde{X}$ alone with derivative $a^2T'(a\widetilde{X} - b\widetilde{Y}) + b^2U'(b\widetilde{X} + a\widetilde{Y})$. Then

$$\begin{aligned} &h\big(a\,T(a\widetilde{X} - b\widetilde{Y}) + b\,U(b\widetilde{X} + a\widetilde{Y})\big|\widetilde{Y}\big)\\ &\qquad = h(\widetilde{X} \mid \widetilde{Y}) + \mathbb{E}\log\big(a^2T'(a\widetilde{X} - b\widetilde{Y}) + b^2U'(b\widetilde{X} + a\widetilde{Y})\big) && (127)\\ &\qquad = h(\widetilde{X}) + \mathbb{E}\log\big(a^2T'(X^*) + b^2U'(Y^*)\big). && (128) \end{aligned}$$

It remains to prove that the second term $\mathbb{E}\log\big(a^2T'(X^*)+b^2U'(Y^*)\big)$ is nonnegative. By the concavity property of the logarithm,

$$\mathbb{E}\log\big(a^2T'(X^*)+b^2U'(Y^*)\big) \geq \mathbb{E}\big(a^2\log T'(X^*)+b^2\log U'(Y^*)\big) \tag{129}$$

$$= a^2\,\mathbb{E}\log T'(X^*)+b^2\,\mathbb{E}\log U'(Y^*), \tag{130}$$

where from (25) and (116),

$$a^2\,\mathbb{E}\log T'(X^*)+b^2\,\mathbb{E}\log U'(Y^*) = a^2\big(h(T(X^*))-h(X^*)\big)+b^2\big(h(U(Y^*))-h(Y^*)\big)$$

$$= a^2\big(h(X)-h(X^*)\big)+b^2\big(h(Y)-h(Y^*)\big) \tag{131}$$

$$= 0. \tag{132}$$

This ends the proof of the EPI.

The equality case in (115) can be easily settled in this proof. If the linear combination is not trivial (that is, if both a and b are nonzero scalars), equality holds in the concavity inequality (129) if and only if $T'(X^*) = U'(Y^*)$. Since X^* and Y^* are independent Gaussian variables, this implies that the derivatives T' and U' are constant and equal, hence X and Y in (122) are Gaussian. Thus equality holds in (115) only for Gaussian variables.

37 Conclusion

Who else but Shannon himself can conclude on his life's work in information theory?

> *"I didn't think in the first stages that it was going to have a great deal of impact. I enjoyed working on this kind of a problem, as I have enjoyed working on many other problems, without any notion of either financial gain or gain in the sense of being famous and so on; and I think indeed that most scientists are oriented that way, that they are working because they like the game."* [49]

References

[1] Nelson M. Blachman, "Communication as a Game," 1957 IRE WESCON Convention Record, pt. 2, 1957, pp. 61–66.

[2] Ludwig Boltzmann, "Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen," *Sitzungsberichte Kaiserlichen Akademie der Wissenschaften*, Wien Mathematisch Naturwissenschaftliche Classe, Vol. 66, pp. 275–370, 1872.

[3] Léon Brillouin, *Science and Information Theory*, Academic Press: New York, U.S.A., 1956.

[4] Samuel Hawksley Burbury, "Boltzmann's Minimum Function," *Nature*, Vol. 51, No. 1308, Nov. 1894, p. 78.

[5] John Parker Burg, *Maximum Entropy Spectral Analysis*, Ph.D. thesis, Department of Geophysics, Stanford University, Stanford, California, U.S.A., May 1975.

[6] Herman Chernoff, "A Measure of the Asymptotic Efficiency of Tests of a Hypothesis Based on a Sum of Observations," *Annals of Mathematical Statistics*, Vol. 23, No. 4, 1952, pp. 493–507.

[7] Adrian Cho, "A Fresh Take on Disorder, Or Disorderly Science?," *Science*, Vol. 297, No. 5585, 2002, pp. 1268–1269.

[8] Rudolf Clausius, *The Mechanical Theory of Heat with its Applications to the Steam Engine and to Physical Properties of Bodies*, Ninth Memoir (1865): "On Several Convenient Forms of the Fundamental Equations of the Mechanical Theory of Heat," pp. 327-365, 1865.

[9] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2nd ed., 2006.

[10] Harald Cramér, *Mathematical Methods of Statistics*, Princeton University Press.: Princeton, NJ, U.S.A., 1946.

[11] Imre Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten". Magyar. Tud. Akad. Mat. Kutato Int. Kozl., Vol. 8, 1963, pp. 85–108.

[12] Georges Darmois, "Sur les limites de la dispersion de certaines estimations," *Revue de l'Institut International de Statistique*, Vol. 13, 1945, pp. 9–15.

[13] Joseph Leo Doob, "Review of C. E. Shannon's 'A Mathematical Theory of Communication'," *Mathematical Reviews*, vol. 10, p. 133, Feb. 1949.

[14] John P. Dougherty, "Foundations of Non-equilibrium Statistical Mechanics," *Philosophical Transactions: Physical Sciences and Engineering*, Royal Society London, Vol. 346, No. 1680, pp. 259-305, Feb. 1994.

[15] Francis Ysidro Edgeworth, "On the Probable Errors of Frequency-Constants," *Journal of the Royal Statistical Society*, Vol. 71, No. 2,3,4, 1908, pp. 381-397, 499-512, 651-678.

[16] Paul Erdős, "Graph Theory and Probability," *Canadian Journal of Mathematics*, Vol. 11, No. 0, 1959, pp. 34–38.

[17] Robert Mario Fano, *Transmission of Information: A Statistical Theory of Communications.* Cambridge, Mass: MIT Press, 1961.

[18] ———, interview by Arthur L. Norberg, Charles Babbage Institute, Center for the History of Information Processing, University of Minnesota, Minneapolis, 20-21 April 1989. See also [17, p. vii].

[19] ———, interview by Aftab, Cheung, Kim, Thakkar, Yeddanapudi, 6.933 Project History, Massachusetts Institute of Technology, Nov. 2001.

[20] Ronald Aylmer Fisher, "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London A* Vol. 222, 1922, pp. 309–368.

[21] Maurice Fréchet, "Sur l'extension de certaines évaluations statistiques au cas de petit échantillons," *Revue de l'Institut International de Statistique*, Vol. 11, 1943, pp. 182–205.

[22] Howard Gardner, *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books, 1987, p. 144.

[23] Josiah Willard Gibbs, *Elementary Principles in Statistical Mechanics (developed with especial reference to the rational foundation of thermodynamics)*, Dover Publications: New York, U.S.A., 1902. (Theorem I p. 129)

[24] Friedrich-Wilhelm Hagemeyer, *Die Entstehung von lnformationskonzepten in der Nachrichtentechnik: Eine Fallstudie zur Theoriebildung in der Technik in Industrie und Kriegsforschung* [The Origin of Information Theory Concepts in Communication Technology: Case Study for Engineering Theory-Building in Industrial and Military Research], Doctoral Dissertation, Freie Universität Berlin, Nov. 8, 1979, 570 pp.

[25] Ralph Vinton Lyon Hartley, "Transmission of Information," *Bell System Technical Journal*, July 1928, pp. 535–563.

[26] Jan Havrda and František Charvát, "Quantification Method of Classification Processes: Concept of Structural α-Entropy," *Kybernetika*, Vol. 3, No. 1, 1967, pp. 30–34.

[27] Erik Hollnagel and David D. Woods, *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*, Taylor & Francis: Boca Raton, FL, U.S.A., 2005 (p.11).

[28] Stig Hjalmars, "Evidence for Boltzmann's H as a Capital Eta", *American Journal of Physics*, Vol. 45, No. 2, Feb. 1977, p. 214.

[29] Edwin Thompson Jaynes, "Information Theory and Statistical Mechanics," *Physical Review* Vol. 106, No. 4, pp. 620–630, May 1957 and Vol. 108, No. 2, pp. 171–190, Oct. 1957.

[30] Yu. L. Klimontovich, *Statistical Theory of Open Systems. I. A Unified Approach to Kinetic Description of Processes in Active Systems*. Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995 (p. 25).

[31] Ronald R. Kline, *The Cybernetics Moment: Or Why We Call Our Age the Information Age*. Johns Hopkins University Press: Baltimore, MD, U.S.A., 2015, xi+ 336 pp. (p. 123).

[32] Andreï Nikolaïevitch Kolmogorov, "Combinatorial Foundations of Information Theory and the Calculus of Probabilities," talk at the *International Mathematical Congress* (Nice, 1970), *Russian Mathematical Surveys*, Vol. 38, No. 4, pp. 29–40, 1983.

[33] Solomon Kullback and Richard A. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79-86, 1951.

[34] Harry Nyquist, "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, pp. 324–346, and "Certain Topics in Telegraph Transmission Theory," *Transactions of the American Institute of Electrical Engineers*, Vol. 47, April 1928, pp. 617–644.

[35] John Robinson Pierce, "The Early Days of Information Theory," *IEEE Transactions on Information Theory*, vol. IT-19, no. 1, Jan. 1973, pp. 3–8.

[36] Calyampudi Radakrishna Rao, "Information and the Accuracy Attainable in the Estimation of Statistical Parameters," *Bulletin of the Calcutta Mathematical Society*, Vol. 37, 1945, pp. 81–89.

[37] Alfréd Rényi, "On Measures of Entropy and Information," in *Proc. 4th Berkeley Symp. Math., Stat. Prob.*, Berkeley, California, U.S.A., 20 June–30 July 1960, University of California Press, Vol. 1, 1960, pp. 547–561.

[38] Olivier Rioul and José Carlos Magossi, "On Shannon's Formula and Hartley's Rule: Beyond the Mathematical Coincidence," in *Entropy*, Special Issue on *Information, Entropy and their Geometric Structures*, Vol. 16, No. 9, pp. 4892-4910, Sept. 2014.

[39] Olivier Rioul, "Yet Another Proof of the Entropy Power Inequality," *IEEE Transactions on Information Theory*, Vol. 63, No. 6, June 2017, pp. 3595–3599.

[40] Claude Elwood Shannon, "A Symbolic Analysis of Relay and Switching Circuits," Thesis (Master of Science), M.I.T., August 10, 1937. In *Transactions American Institute of Electrical Engineers*, Vol. 57, 1938, pp. 713–723.

[41] ——, "An Algebra for Theoretical Genetics," Ph.D. Dissertation, Department of Mathematics, Massachusetts Institute of Technology, April 15, 1940, 69 pp.

[42] ——, "A Mathematical Theory of Cryptography," Memorandum MM 45-110-02, Sept. 1, 1945, Bell Laboratories, 114 pp. Declassified and reprinted as "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, Vol. 28, 1949, pp. 656–715.

[43] ——, "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27, July and October 1948, pp. 379–423 and 623–656.

[44] —— (with Warren Weaver), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, U.S.A., 1949, vi+117 pp.

[45] ——, "The Bandwagon," *IRE Transactions on Information Theory*, Editorial, Vol. 2, No. 1, March 1956, p. 3.

[46] ——, interview by Friedrich-Wilhelm Hagemeyer, Winchester, Massachusetts, U.S.A., February 28, 1977.

[47] ——, interview by Robert Price, Winchester, Massachusetts, U.S.A., IEEE History Center, July 28, 1982. Partly published in F. W. Ellersick, "A Conversation with Claude Shannon," IEEE Communications Magazine, Vol. 22, 1984, pp. 123–126.

[48] ——, interview by Anthony Liversidge, *Omni magazine*, August 1987.

[49] ——, TV interview, circa 1990.

[50] Adriaan Johannes Stam, "Some Inequalities Satisfied by the Quantities of Information of Fisher and Shannon," *Information and Control*, Vol. 2, No. 2, 1959, pp. 101–112.

[51] Leó Szilárd, "Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen." [On the Decrease of Entropy in a Thermodynamic System by the Intervention of Intelligent Beings], *Zeitchrift für Physik*, Vol. 53, 1929, 840–856.

[52] Libb Thims, "Thermodynamics ≠ Information Theory: Science's Greatest Sokal Affair," *Journal of Human Thermodynamics*, Vol. 8, No. 1, Dec. 2012, pp. 1–120.

[53] Myron Tribus and Edward C. McIrvine, "Energy and Information," *Scientific American*, Vol. 225, 1971, pp. 179–188.

[54] Myron Tribus, "A Tribute to Edwin T. Jaynes," in *Proceedings of the 18th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, Garching, Germany, 1998, pp. 11–20.

[55] Constantino Tsallis, "Possible Generalization of Boltzmann-Gibbs Statistics," *Journal of Statistical Physics*, Vol. 52, Nos. 1/2, 1988, pp. 479–487.

[56] Johann von Neumann, *Mathematische Grundlagen der Quantenmechanik*, Verlag Von Julius Springer: Berlin, Germany, 1932.

[57] Warren Weaver, "The Mathematics of Communication," *Scientific American*, vol. 181, no. 1, July 1949, pp. 11–15.

[58] Norbert Wiener, "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series (with Engineering Applications)," M.I.T., Feb. 1, 1942. Published by Technology Press and John Wiley & Sons, 1949.

[59] ———, *Cybernetics*, Chapter III: Time series, Information and Communication, John Wiley & Sons: New York, NY, U.S.A., 1948, pp. 10–11.

[60] ———, "What is Information Theory?" *IRE Transactions on Information Theory*, Editorial, Vol. 2, No. 2, July 1956, p. 48.

[61] IEC 80000-13:2008, Quantities and units – Part 13: Information science and technology, International Organization for Standardization.