arXiv:1802.05968v3 [cs.IT] 13 Jun 2019

Information Theory: A Tutorial Introduction

James V Stone, Psychology Department, University of Sheffield, England.

j.v.stone@sheffield.ac.uk

File: main_InformationTheory_JVStone_v4.tex

Abstract

Shannon's mathematical theory of communication defines fundamental limits on how much information can be transmitted between the different components of any man-made or biological system. This paper is an informal but rigorous introduction to the main ideas implicit in Shannon's theory. An annotated reading list is provided for further reading.

1 Introduction

In 1948, Claude Shannon published a paper called *A Mathematical Theory of Communication*[1]. This paper heralded a transformation in our understanding of information. Before Shannon's paper, information had been viewed as a kind of poorly defined miasmic fluid. But after Shannon's paper, it became apparent that information is a well-defined and, above all, *measurable* quantity. Indeed, as noted by Shannon,

> *A basic idea in information theory is that information can be treated very much like a physical quantity, such as mass or energy.*
> Caude Shannon, 1985.

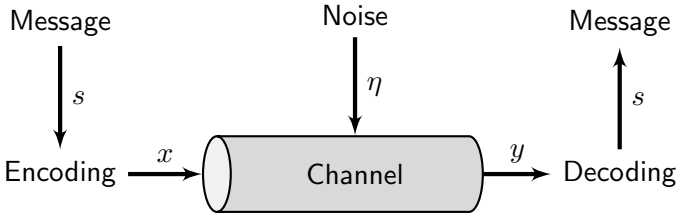


Figure 1: The communication channel. A message (data) is encoded before being used as input to a communication channel, which adds noise. The channel output is decoded by a receiver to recover the message.

Information theory defines definite, unbreachable limits on precisely how much information can be communicated between any two components of any system, whether this system is man-made or natural. The theorems of information theory are so important that they deserve to be regarded as the *laws* of information[2, 3, 4].

The basic laws of information can be summarised as follows. For any communication channel (Figure 1): 1) there is a definite upper limit, the *channel capacity*, to the amount of information that can be communicated through that channel, 2) this limit shrinks as the amount of noise in the channel increases, 3) this limit can very nearly be reached by judicious packaging, or encoding, of data.

2 Finding a Route, Bit by Bit

Information is usually measured in *bits*, and one bit of information allows you to choose between two equally probable, or *equiprobable*, alternatives. In order to understand why this is so, imagine you are standing at the fork in the road at point A in Figure 2, and that you want to get to the point marked D. The fork at A represents two equiprobable alternatives, so if I tell you to go left then you have received one bit of information. If we represent my instruction with a *binary digit* (0=left and 1=right) then this binary digit provides you with one bit of information, which tells you which road to choose.

Now imagine that you come to another fork, at point B in Figure 2. Again, a binary digit (1=right) provides one bit of information, allowing you to choose the correct road, which leads to C. Note that C is one of four possible interim destinations that you could

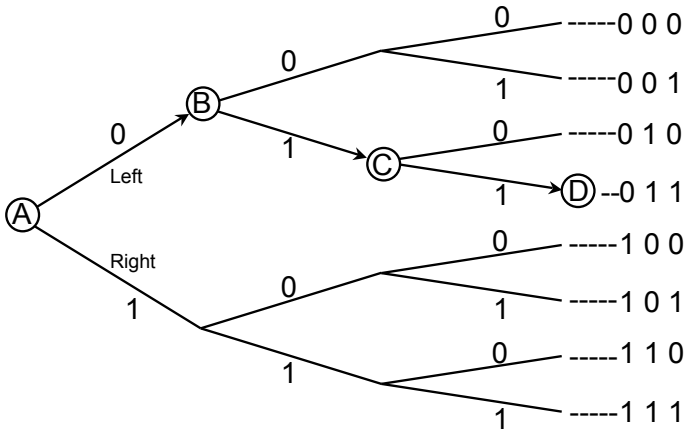


Figure 2: For a traveller who does not know the way, each fork in the road requires one bit of information to make a correct decision. The 0s and 1s on the right-hand side summarise the instructions needed to arrive at each destination; a left turn is indicated by a 0 and a right turn by a 1.

have reached after making two decisions. The two binary digits that allow you to make the correct decisions provided two bits of information, allowing you to choose from four (equiprobable) alternatives; 4 equals $2 \times 2 = 2^2$.

A third binary digit (1=right) provides you with one more bit of information, which allows you to again choose the correct road, leading to the point marked D. There are now eight roads you could have chosen from when you started at A, so three binary digits (which provide you with three bits of information) allow you to choose from eight equiprobable alternatives, which also equals $2 \times 2 \times 2 = 2^3 = 8$.

We can restate this in more general terms if we use n to represent the number of forks, and m to represent the number of final destinations. If you have come to n forks then you have effectively chosen from $m = 2^n$ final destinations. Because the decision at each fork requires one bit of information, n forks require n bits of information.

Viewed from another perspective, if there are $m = 8$ possible destinations then the number of forks is $n = 3$, which is the *logarithm* of 8. Thus, $3 = \log_2 8$ is the number of forks implied by eight destinations. More generally, the logarithm of m is the power to which 2 must be raised in order to obtain m; that is, $m = 2^n$. Equivalently, given a number m, which we wish to express as a logarithm, $n = \log_2 m$. The subscript $_2$ indicates that we are using logs to the base 2 (all logarithms in this book use base 2 unless stated otherwise).

3 Bits Are Not Binary Digits

The word *bit* is derived from *binary digit*, but a bit and a binary digit are fundamentally different types of quantities. A binary digit is the value of a binary variable, whereas a bit is an *amount of information*. To mistake a binary digit for a bit is a category error. In this case, the category error is not as bad as mistaking marzipan for justice, but it is analogous to mistaking a pint-sized bottle for a pint of milk. Just as a bottle can contain between zero and one pint, so a binary digit (when averaged over both of its possible states) can convey between zero and one bit of information.

4 Information and Entropy

Consider a coin which lands heads up 90% of the time (i.e. $p(x_h) = 0.9$). When this coin is flipped, we expect it to land heads up ($x = x_h$), so when it does so we are less surprised than when it lands tails up ($x = x_t$). The more improbable a particular outcome is, the

more surprised we are to observe it. If we use logarithms to the base 2 then the Shannon information or *surprisal* of each outcome is measured in bits (see Figure 3a)

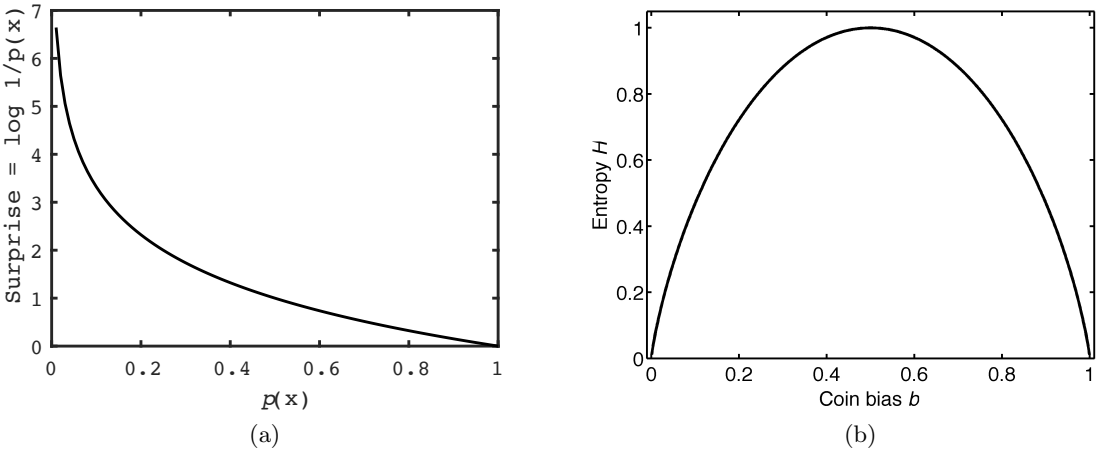
$$\text{Shannon information} \quad = \quad \log \frac{1}{p(x_h)} \text{ bits}, \tag{1}$$

which is often expressed as: information $= -\log\, p(x_h)$ bits.

Entropy is Average Shannon Information. We can represent the outcome of a coin flip as the *random variable* x, such that a head is $x = x_h$ and a tail is $x = x_t$. In practice, we are not usually interested in the surprise of a particular value of a random variable, but we are interested in how much surprise, on average, is associated with the entire set of possible values. The average surprise of a variable x is defined by its probability distribution $p(x)$, and is called the *entropy* of $p(x)$, represented as $H(x)$.

The Entropy of a Fair Coin. The average amount of surprise about the possible outcomes of a coin flip can be found as follows. If a coin is fair or unbiased then $p(x_h) = p(x_t) = 0.5$ then the Shannon information gained when a head or a tail is observed is $\log 1/0.5 = 1$ bit, so the average Shannon information gained after each coin flip is also 1 bit. Because entropy is defined as average Shannon information, the entropy of a fair coin is $H(x) = 1$ bit.

The Entropy of an Unfair (Biased) Coin. If a coin is biased such that the probability of a head is $p(x_h) = 0.9$ then it is easy to predict the result of each coin flip (i.e. with 90% accuracy if we predict a head for each flip). If the outcome is a head then the amount of Shannon information gained is $\log(1/0.9) = 0.15$ bits. But if the outcome is a tail then



(a) (b)

Figure 3: a) Shannon information as surprise. Values of x that are less probable have larger values of surprise, defined as $\log_2(1/p(x))$ bits. b) Graph of entropy $H(x)$ versus coin bias (probability $p(x_h)$ of a head). The entropy of a coin is the average amount of surprise or Shannon information in the distribution of possible outcomes (i.e. heads and tails).

the amount of Shannon information gained is $\log(1/0.1) = 3.32$ bits. Notice that more information is associated with the more surprising outcome. Given that the proportion of flips that yield a head is $p(x_h)$, and that the proportion of flips that yield a tail is $p(x_t)$ (where $p(x_h) + p(x_t) = 1$), the average surprise is

$$H(x) = p(x_h) \log \frac{1}{p(x_h)} + p(x_t) \log \frac{1}{p(x_t)}, \quad (2)$$

which comes to $H(x) = 0.469$ bits, as in Figure 3b. If we define a tail as $x_1 = x_t$ and a head as $x_2 = x_h$ then Equation 2 can be written as

$$H(x) = \sum_{i=1}^{2} p(x_i) \log \frac{1}{p(x_i)} \text{bits}. \quad (3)$$

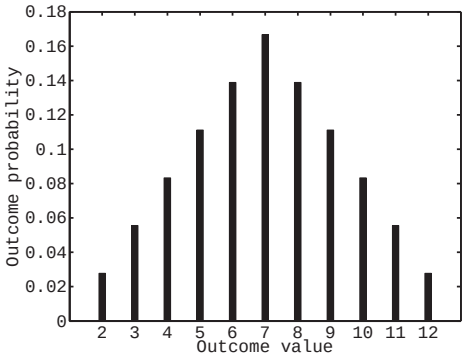
More generally, a random variable x with a probability distribution $p(x) = \{p(x_1), \ldots, p(x_m)\}$ has an entropy of

$$H(x) = \sum_{i=1}^{m} p(x_i) \log \frac{1}{p(x_i)} \text{ bits}. \quad (4)$$

The reason this definition matters is because Shannon's source coding theorem (see Section 7) guarantees that each value of the variable x can be represented with an average of (just over) $H(x)$ binary digits. However, if the values of consecutive values of a random variable are not independent then each value is more predictable, and therefore less surprising, which reduces the information-carrying capability (i.e. entropy) of the variable. This is why it is important to specify whether or not consecutive variable values are *independent*.

Interpreting Entropy. If $H(x) = 1$ bit then the variable x could be used to represent $m = 2^{H(x)}$or 2 equiprobable values. Similarly, if $H(x) = 0.469$ bits then the variable x

(a)



(b)

Figure 4: (a) A pair of dice. (b) Histogram of dice outcome values.

could be used to represent $m = 2^{0.469}$ or 1.38 equiprobable values; as if we had a die with 1.38 sides. At first sight, this seems like an odd statement. Nevertheless, translating entropy into an equivalent number of equiprobable values serves as an intuitive guide for the amount of information represented by a variable.

Dicing With Entropy. Throwing a pair of 6-sided dice yields an *outcome* in the form of an ordered pair of numbers, and there are a total of 36 equiprobable outcomes, as shown in Table 1. If we define an *outcome value* as the sum of this pair of numbers then there are $m = 11$ possible outcome values $A_x = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, represented by the symbols $x_1, \ldots, x_{11}$. These outcome values occur with the frequencies shown in Figure 4b and Table 1. Dividing the frequency of each outcome value by 36 yields the probability P of each outcome value. Using Equation 4, we can use these 11 probabilities to find the entropy

$$\begin{aligned} H(x) &= p(x_1) \log \frac{1}{p(x_1)} + p(x_2) \log \frac{1}{p(x_2)} + \cdots + p(x_{11}) \log \frac{1}{p(x_{11})} \\ &= 3.27 \text{ bits.} \end{aligned}$$

Using the interpretation described above, a variable with an entropy of 3.27 bits can represent $2^{3.27} = 9.65$ equiprobable values.

Entropy and Uncertainty. Entropy is a measure of *uncertainty*. When our uncertainty is reduced, we gain information, so information and entropy are two sides of the same coin. However, information has a rather subtle interpretation, which can easily lead to confusion.

Average information shares the same definition as entropy, but whether we call a given quantity information or entropy depends on whether it is being given to us or taken away.

Symbol	Sum	Outcome	Frequency	P	Surprisal
x_1	2	1:1	1	0.03	5.17
x_2	3	1:2, 2:1	2	0.06	4.17
x_3	4	1:3, 3:1, 2:2	3	0.08	3.59
x_4	5	2:3, 3:2, 1:4, 4:1	4	0.11	3.17
x_5	6	2:4, 4:2, 1:5, 5:1, 3:3	5	0.14	2.85
x_6	7	3:4, 4:3, 2:5, 5:2, 1:6, 6:1	6	0.17	2.59
x_7	8	3:5, 5:3, 2:6, 6:2, 4:4	5	0.14	2.85
x_8	9	3:6, 6:3, 4:5, 5:4	4	0.11	3.17
x_9	10	4:6, 6:4, 5:5	3	0.08	3.59
x_{10}	11	5:6, 6:5	2	0.06	4.17
x_{11}	12	6:6	1	0.03	5.17

Table 1: A pair of dice have 36 possible outcomes.
Sum: outcome value, total number of dots for a given throw of the dice.
Outcome: ordered pair of dice numbers that could generate each symbol.
Freq: number of different outcomes that could generate each outcome value.
P: the probability that the pair of dice yield a given outcome value (freq/36).
Surprisal: $P \log(1/P)$ bits.

For example, if a variable has high entropy then our initial uncertainty about the value of that variable is large and is, by definition, exactly equal to its entropy. If we are told the value of that variable then, on average, we have been given information equal to the uncertainty (entropy) we had about its value. Thus, receiving an amount of information is equivalent to having exactly the same amount of entropy (uncertainty) taken away.

5 Entropy of Continuous Variables

For discrete variables, entropy is well-defined. However, for all continuous variables, entropy is effectively infinite. Consider the difference between a discrete variable x_d with n possible values and a continuous variable x_c with an uncountably infinite number of possible values; for simplicity, assume that all values are equally probable. The probability of observing each value of the discrete variable is $P_d = 1/m$, so the entropy of x_d is $H(x_d) = \log m$. In contrast, the probability of observing each value of the continuous variable is $P_c = 1/\infty = 0$, so the entropy of x_c is $H(x_c) = \log \infty = \infty$. In one respect, this makes sense, because each value of of a continuous variable is implicitly specified with infinite precision, from which it follows that the amount of information conveyed by each value is infinite. However, this result implies that all continuous variables have the same entropy. In order to assign different values to different variables, all infinite terms are simply ignored, which yields the *differential entropy*

$$H(x_c) = \int p(x_c) \log \frac{1}{p(x_c)} \, dx_c. \tag{5}$$

It is worth noting that the technical difficulties associated with entropy of continuous variables disappear for quantities like mutual information, which involve the difference between two entropies. For convenience, we use the term entropy for both continuous and discrete variables below.

6 Maximum Entropy Distributions

A distribution of values that has as much entropy (information) as theoretically possible is a *maximum entropy distribution*. Maximum entropy distributions are important because, if we wish to use a variable to transmit as much information as possible then we had better make sure it has maximum entropy. For a given variable, the precise form of its maximum entropy distribution depends on the constraints placed on the values of that variable[3]. It will prove useful to summarise three important maximum entropy distributions. These are listed in order of decreasing numbers of constraints below.

The Gaussian Distribution. If a variable x has a fixed variance, but is otherwise unconstrained, then the maximum entropy distribution is the Gaussian distribution (Figure 5a). This is particularly important in terms of energy efficiency because no other distribution can provide as much information at a lower energy cost per bit. If a variable has a Gaussian or *normal* distribution then the probability of observing a particular value x is

$$p(x) = \frac{1}{\sqrt{2\pi v_x}} e^{-(\mu_x - x)^2/(2v_x)}, \quad (6)$$

where $e = 2.7183$. This equation defines the bell-shaped curve in Figure 5a. The term μ_x is the mean of the variable x, and defines the central value of the distribution; we assume that all variables have a mean of zero (unless stated otherwise). The term v_x is the variance of the variable x, which is the square of the standard deviation σ_x of x, and defines the width of the bell curve. Equation 6 is a *probability density function*, and (strictly speaking) $p(x)$ is the *probability density* of x.

The Exponential Distribution. If a variable has no values below zero, and has a fixed mean μ, but is otherwise unconstrained, then the maximum entropy distribution is exponential,

$$p(x) = \frac{1}{\mu} e^{-x/\mu}, \quad (7)$$

which has a variance of $\text{var}(x) = \mu^2$, as shown in Figure 5b.

The Uniform Distribution. If a variable has a fixed lower bound x_{min} and upper bound x_{max}, but is otherwise unconstrained, then the maximum entropy distribution is uniform,

$$p(x) = 1/(x_{max} - x_{min}), \quad (8)$$

which has a variance $(x_{max} - x_{min})^2/12$, as shown in Figure 5c.

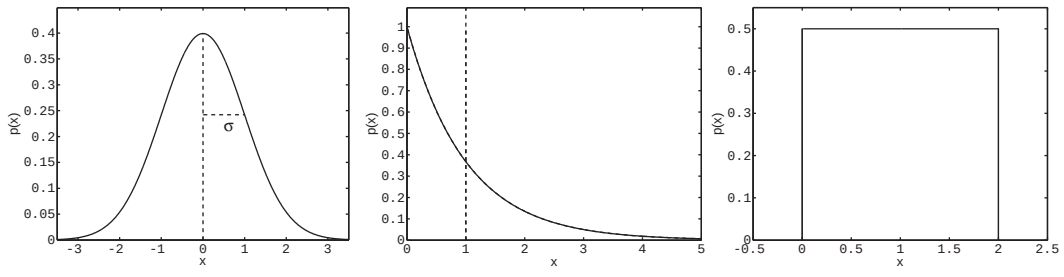


Figure 5: Maximum entropy distributions. a) Gaussian distribution, with mean $\mu = 0$ and a standard deviation $\sigma = 1$ (Equation 6). b) Exponential distribution, with mean indicated by the vertical line (Equation 7). c) A uniform distribution with a range between zero and two (Equation 8).

7 Channel Capacity

A very important (and convenient) channel is the additive channel. As encoded values pass through an additive channel, noise η (eta) is added, so that the channel output is a noisy version y of the channel input x

$$y \quad = \quad x + \eta. \tag{9}$$

The *channel capacity* C is the maximum amount of information that a channel can provide at its output about the input.

The rate at which information is transmitted through the channel depends on the entropies of three variables: 1) the entropy $H(x)$ of the input, 2) the entropy $H(y)$ of the output, 3) the entropy $H(\eta)$ of the noise in the channel. If the output entropy is high then this provides a large potential for information transmission, and the extent to which this potential is realised depends on the input entropy and the level of noise. If the noise is low then the output entropy can be close to the channel capacity. However, channel capacity gets progressively smaller as the noise increases. Capacity is usually expressed in bits per usage (i.e. bits per output), or bits per second (bits/s).

8 Shannon's Source Coding Theorem

Shannon's source coding theorem, described below, applies only to noiseless channels. This theorem is really about re-packaging (encoding) data before it is transmitted, so that, when

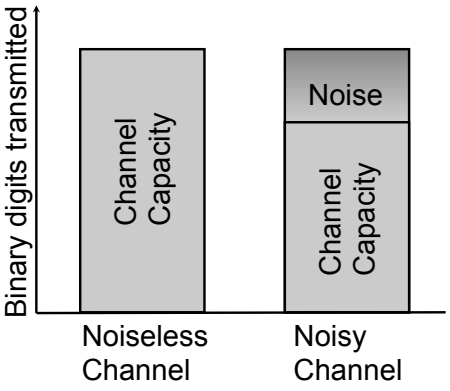


Figure 6: The *channel capacity* of noiseless and noisy channels is the maximum rate at which information can be communicated. If a noiseless channel communicates data at 10 binary digits/s then its capacity is $C = 10$ bits/s. The capacity of a noiseless channel is numerically equal to the rate at which it communicates binary digits, whereas the capacity of a noisy channel is less than this because it is limited by the noise in the channel.

it is transmitted, every datum conveys as much information as possible. This theorem is highly relevant to the biological information processing because it defines definite limits to how efficiently sensory data can be re-packaged. We consider the source coding theorem using binary digits below, but the logic of the argument applies equally well to any channel inputs.

Given that a binary digit can convey a maximum of one bit of information, a noiseless channel which communicates R binary digits per second can communicate information at the rate of up to R bits/s. Because the capacity C is the maximum rate at which it can communicate information from input to output, it follows that the capacity of a noiseless channel is numerically equal to the number R of binary digits communicated per second. However, if each binary digit carries less than one bit (e.g. if consecutive output values are correlated) then the channel communicates information at a lower rate $R < C$.

Now that we are familiar with the core concepts of information theory, we can quote Shannon's source coding theorem in full. This is also known as Shannon's *fundamental theorem for a discrete noiseless channel*, and as the *first fundamental coding theorem*.

> Let a source have entropy H (bits per symbol) and a channel have a capacity C (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate $C/H - \epsilon$ symbols per second over the channel where ϵ is arbitrarily small. It is not possible to transmit at an average rate greater than C/H [symbols/s].
> Shannon and Weaver, 1949[2].
> [Text in square brackets has been added by the author.]

Recalling the example of the sum of two dice, a naive encoding would require 3.46 (log 11) binary digits to represent the sum of each throw. However, Shannon's source coding theorem guarantees that an encoding exists such that an average of (just over) 3.27 (i.e. log 9.65) binary digits per value of s will suffice (the phrase 'just over' is an informal interpretation of Shannon's more precise phrase 'arbitrarily close to').

This encoding process yields inputs with a specific distribution $p(x)$, where there are implicit constraints on the form of $p(x)$ (e.g.power constraints). The shape of the distribution $p(x)$ places an upper limit on the entropy $H(x)$, and therefore on the maximum information that each input can carry. Thus, the capacity of a noiseless channel is defined in terms of the particular distribution $p(x)$ which maximises the amount of information per input

$$C = \max_{p(x)} H(x) \text{ bits per input.} \tag{10}$$

This states that channel capacity C is achieved by the distribution $p(x)$ which makes $H(x)$ as large as possible (see Section 6).

9 Noise Reduces Channel Capacity

Here, we examine how noise effectively reduces the maximum information that a channel can communicate. If the number of equiprobable (signal) input states is m_x then the input entropy is

$$H(x) \;=\; \log\, m_x \,\text{bits}. \tag{11}$$

For example, suppose there are $m_x = 3$ equiprobable input states, say, $x_1 = 100$ and $x_2 = 200$ and $x_3 = 300$, so the input entropy is $H(x) = \log 3 = 1.58$ bits. And if there are $m_\eta = 2$ equiprobable values for the channel noise, say, $\eta_1 = 10$ and $\eta_2 = 20$, then the noise entropy is $H(\eta) = \log 2 = 1.00$ bit.

Now, if the input is $x_1 = 100$ then the output can be one of two equiprobable states, $y_1 = 100 + 10 = 110$ or $y_2 = 100 + 20 = 120$. And if the input is $x_2 = 200$ then the output can be either $y_3 = 210$ or $y_4 = 220$. Finally, if the input is $x_3 = 300$ then the output can be either $y_5 = 310$ or $y_6 = 320$. Thus, given three equiprobable input states and two equiprobable noise values, there are $m_y = 6 (= 3 \times 2)$ equiprobable output states. So the output entropy is $H(y) = \log 6 = 2.58$ bits. However, some of this entropy is due to noise, so not all of the output entropy comprises *information about the input.*

In general, the total number m_y of equiprobable output states is $m_y = m_x \times m_\eta$, from which it follows that the output entropy is

$$H(y) \;=\; \log\, m_x + \log\, m_\eta \tag{12}$$

$$\;=\; H(x) + H(\eta)\,\text{bits}. \tag{13}$$

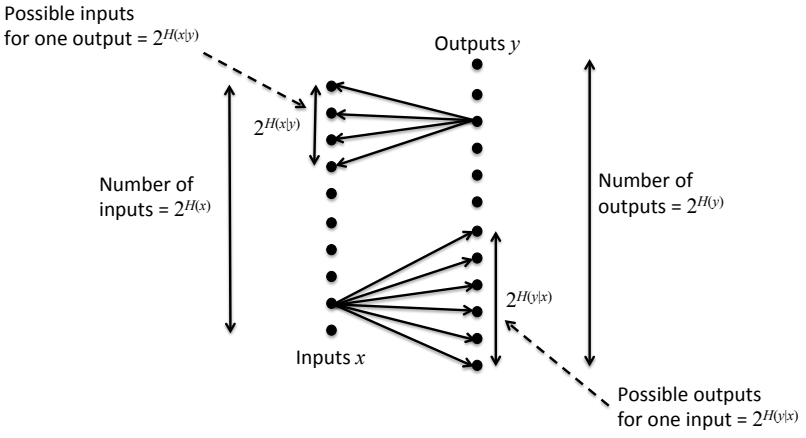


Figure 7: A fan diagram shows how channel noise affects the number of possible outputs given a single input, and *vice versa.* If the noise η in the channel output has entropy $H(\eta) = H(Y|X)$ then each input value could yield one of $2^{H(Y|X)}$ equally probable output values. Similarly, if the noise in the channel input has entropy $H(X|Y)$ then each output value could have been caused by one of $2^{H(X|Y)}$ equally probable input values.

Because we want to explore channel capacity in terms of channel noise, we will pretend to reverse the direction of data along the channel. Accordingly, before we 'receive' an input value, we know that the output can be one of 6 values, so our uncertainty about the input value is summarised by its entropy $H(y) = 2.58$ bits.

Conditional Entropy. Our average uncertainty about the output value given an input value is the *conditional entropy* $H(y|x)$. The vertical bar denotes 'given that', so $H(y|x)$ is, 'the residual uncertainty (entropy) of y given that we know the value of x'.

After we have received an input value, our uncertainty about the output value is reduced from $H(y) = 2.58$ bits to

$$H(y|x) \quad = \quad H(\eta) \quad = \quad \log 2 \quad = \quad 1\text{bit}. \tag{14}$$

Because $H(y|x)$ is the entropy of the channel noise η, we can write it as $H(\eta)$. Equation 14 is true for every input, and it is therefore true for the average input. Thus, for each input, we gain an average of

$$H(y) - H(\eta) \quad = \quad 2.58 - 1 \text{ bits}, \tag{15}$$

about the output, which is the *mutual information* between x and y.

10 Mutual Information

The mutual information $I(x, y)$ between two variables, such as a channel input x and output y, is the average amount of information that each value of x provides about y

$$I(x, y) \quad = \quad H(y) - H(\eta) \text{ bits}. \tag{16}$$

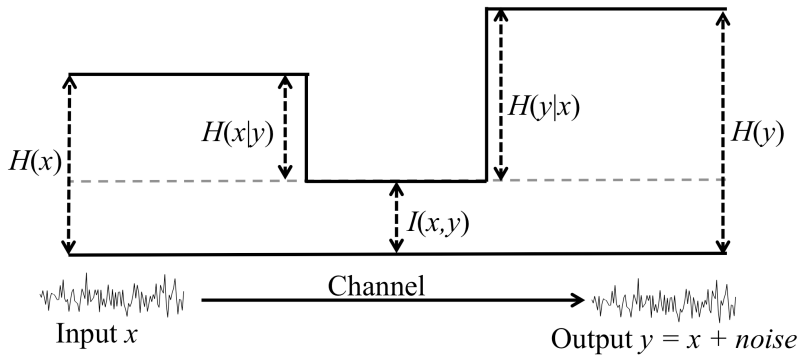


Figure 8: The relationships between information theoretic quantities. Noise refers to noise η in the output, which induces uncertainty $H(y|x) = H(\eta)$ regarding the output given the input; this noise also induces uncertainty $H(x|y)$ regarding the input given the output. The mutual information is $I(x, y) = H(x) - H(x|y) = H(y) - H(y|x)$ bits.

Somewhat counter-intuitively, the average amount of information gained about the output when an input value is received is the same as the average amount of information gained about the input when an output value is received, $I(x, y) = I(y, x)$. This is why it did not matter when we pretended to reverse the direction of data through the channel. These quantities are summarised in Figure 8.

11 Shannon's Noisy Channel Coding Theorem

All practical communication channels are noisy. To take a trivial example, the voice signal coming out of a telephone is not a perfect copy of the speaker's voice signal, because various electrical components introduce spurious bits of noise into the telephone system.

As we have seen, the effects of noise can be reduced by using error correcting codes. These codes reduce errors, but they also reduce the rate at which information is communicated. More generally, any method which reduces the effects of noise also reduces the rate at which information can be communicated. Taking this line of reasoning to its logical conclusion seems to imply that the only way to communicate information with zero error is to reduce the effective rate of information transmission to zero, and in Shannon's day this was widely believed to be true. But Shannon proved that information can be communicated, with vanishingly small error, at a rate which is limited only by the channel capacity.

Now we give Shannon's *fundamental theorem for a discrete channel with noise*, also known as the *second fundamental coding theorem*, and as *Shannon's noisy channel coding theorem*[2]:

> Let a discrete channel have the capacity C and a discrete source the entropy per second H. If $H \leq C$ there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If $H \geq C$ it is possible to encode the source so that the equivocation is less than $H - C + \epsilon$ where ϵ is arbitrarily small. There is no method of encoding which gives an equivocation less than $H - C$.

(The word 'equivocation' means the average uncertainty that remains regarding the value of the input after the output is observed, i.e. the conditional entropy $H(X|Y)$). In essence, Shannon's theorem states that it is possible to use a communication channel to communicate information with a low error rate ϵ (epsilon), at a rate arbitrarily close to the channel

capacity of C bits/s, but it is not possible to communicate information at a rate greater than C bits/s.

The capacity of a noisy channel is defined as

$$C = \max_{p(x)} I(x, y) \quad (17)$$

$$= \max_{p(x)} [H(y) - H(y|x)] \text{ bits.} \quad (18)$$

If there is no noise (i.e. if $H(y|x) = 0$) then this reduces to Equation 10, which is the capacity of a noiseless channel. The *data processing inequality* states that, no matter how sophisticated any device is, the amount of information $I(x, y)$ in its output about its input cannot be greater than the amount of information $H(x)$ in the input.

12 The Gaussian Channel

If the noise values in a channel are drawn independently from a Gaussian distribution (i.e. $\eta \sim \mathcal{N}(\mu_\eta, v_\eta)$, as defined in Equation 6) then this defines a *Gaussian channel*.

Given that $y = x + \eta$, if we want $p(y)$ to be Gaussian then we should ensure that $p(x)$ and $p(\eta)$ are Gaussian, because the sum of two independent Gaussian variables is also Gaussian[3]. So, $p(x)$ must be (iid) Gaussian in order to maximise $H(x)$, which maximises $H(y)$, which maximises $I(x, y)$. Thus, if each input, output, and noise variable is (iid) Gaussian then the average amount of information communicated per output value is the channel capacity, so that $I(x, y) = C$ bits. This is an informal statement of *Shannon's continuous noisy channel coding theorem for Gaussian channels*. We can use this to express capacity in terms of the input, output, and noise.

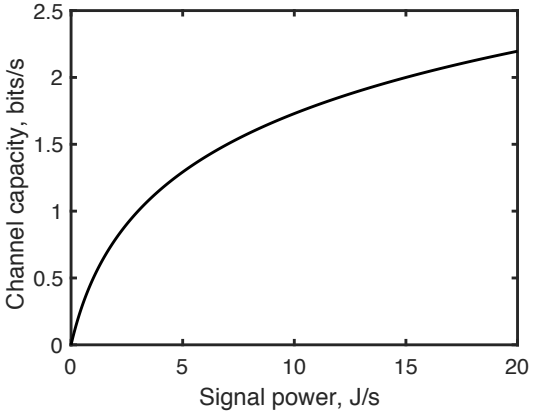


Figure 9: Gaussian channel capacity C (Equation 23) increases slowly with signal power S, which equals signal power here because $N = 1$.

If the channel input $x \sim \mathcal{N}(\mu_x, v_x)$ then the continuous analogue (integral) of Equation 4 yields the *differential entropy*

$$H(x) \quad = \quad (1/2) \log\ 2\pi e v_x \text{ bits.} \tag{19}$$

The distinction between differential entropy effectively disappears when considering the difference between entropies, and we will therefore find that we can safely ignore this distinction here. Given that the channel noise is iid Gaussian, its entropy is

$$H(\eta) \quad = \quad (1/2) \log\ 2\pi e v_\eta \text{ bits.} \tag{20}$$

Because the output is the sum $y = x + \eta$, it is also iid Gaussian with variance $v_y = v_x + v_\eta$, and its entropy is

$$H(y) \quad = \quad (1/2) \log\ 2\pi e (v_x + v_\eta) \text{ bits.} \tag{21}$$

Substituting Equations 20 and 21 into Equation 16 yields

$$I(x, y) \quad = \quad \frac{1}{2} \log \left(1 + \frac{v_x}{v_\eta} \right) \text{ bits,} \tag{22}$$

which allows us to choose one out of $m = 2^I$ equiprobable values. For a Gaussian channel, $I(x, y)$ attains its maximal value of C bits.

The variance of any signal with a mean of zero is equal to its *power*, which is the rate at which energy is expended per second, and the physical unit of power is measured in *Joules* per second (J/s) or *Watts*, where 1 Watt = 1 J/s. Accordingly, the signal power is $S = v_x$ J/s, and the noise power is $N = v_\eta$ J/s. This yields Shannon's famous equation for the capacity of a Gaussian channel

$$C \quad = \quad \frac{1}{2} \log \left(1 + \frac{S}{N} \right) \text{ bits,} \tag{23}$$

where the ratio of variances S/N is the *signal to noise ratio* (SNR), as in Figure 9. It is worth noting that, given a Gaussian signal obscured by Gaussian noise, the probability of detecting the signal is[5]

$$P \quad = \quad \frac{1}{2} \log \left(1 + \text{erf} \left(\sqrt{\frac{S}{8N}} \right) \right), \tag{24}$$

where erf is the cumulative distribution function of a Gaussian.

13 Fourier Analysis

If a sinusoidal signal has a *period* of λ seconds then it has a frequency of $f = 1/\lambda$ periods per second, measured in *Hertz* (Hz). A sinusoid with a frequency of W Hz can be represented perfectly if its value is measured at the *Nyquist sample rate*[6] of $2W$ times per second. Indeed, *Fourier analysis* allows almost any signal x to be represented as a mixture of sinusoidal *Fourier components* $x(f) : (f = 0, \ldots, W)$, shown in Figure 10. A signal which includes frequencies between 0 Hz and WHz has a *bandwidth* of W Hz.

Fourier Analysis. Fourier analysis allows any signal to be represented as a weighted sum of sine and cosine functions (see Section 13). More formally, consider a signal x with a value x_t at time t, which spans a time interval of T seconds. This signal can be represented as a weighted average of sine and cosine functions

$$x_t \;=\; x_0 \;+\; \sum_{n=1}^{\infty} a_n \cos(f_n t) \;+\; \sum_{n=1}^{\infty} b_n \sin(f_n t), \tag{25}$$

where $f_n = 2\pi n/T$ represents frequency, a_n is the Fourier coefficient (amplitude) of a cosine with frequency f_n, and b_n is the Fourier coefficient of a sine with frequency f_n; and x_0 represents the background amplitude (usually assumed to be zero). Taken over all frequencies, these pairs of coefficients represent the *Fourier transform* of x.

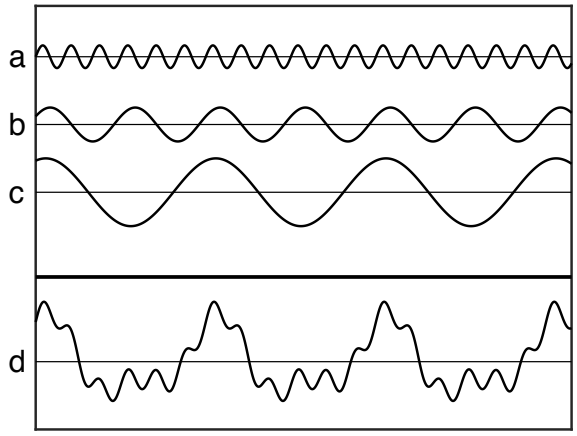


Figure 10: Fourier analysis decomposes the signal x in d into a unique set of sinusoidal *Fourier components* $x(f)$ $(f = 0, \ldots, W\text{Hz})$ in a-c, where d=a+b+c.

The Fourier coefficients can be found from the integrals

$$a_n = \frac{2}{T}\int_0^T x_t \cos(f_n t)\, dt \tag{26}$$

$$b_n = \frac{2}{T}\int_0^T x_t \sin(f_n t)\, dt. \tag{27}$$

Each coefficient a_n specifies how much of the signal x consists of a cosine at the frequency f_n, and b_n specifies how much consists of a sine. Each pair of coefficients specifies the power and *phase* of one frequency component; the power at frequency f_n is $S_f = (a_n^2 + b_n^2)$, and the phase is $\arctan(b_n/a_n)$. If x has a bandwidth of WHz then its *power spectrum* is the set of W values $S_0, \ldots, S_W$.

An extremely useful property of Fourier analysis is that, when applied to *any* variable, the resultant Fourier components are mutually uncorrelated[7], and, when applied to any Gaussian variable, these Fourier components are also mutually independent. This means that the entropy of any Gaussian variable can be estimated by adding up the entropies of its Fourier components, which can be used to estimate the mutual information between Gaussian variables.

Consider a variable $y = x + \eta$, which is the sum of a Gaussian signal x with variance S, and Gaussian noise with variance N. If the highest frequency in y is W Hz, and if values of x are transmitted at the Nyquist rate of $2W$ Hz, then the channel capacity is $2WC$ bits per second, (where C is defined in Equation 23). Thus, when expressed in terms of bits per second, this yields a channel capacity of

$$C = W \log\left(1 + \frac{S}{N}\right) \text{ bits/s.} \tag{28}$$

If the signal power of Fourier component $x(f)$ is $S(f)$, and the noise power of component $\eta(f)$ is $N(f)$ then the signal to noise ratio is $S(f)/N(f)$ (see Section 13). The mutual information at frequency f is therefore

$$I(x(f), y(f)) = \log\left(1 + \frac{S(f)}{N(f)}\right) \text{ bits/s.} \tag{29}$$

Because the Fourier components of any Gaussian variable are mutually independent, the mutual information between Gaussian variables can be obtained by summing $I(x(f), y(f))$ over frequency

$$I(x, y) = \int_{f=0}^{W} I(x(f), y(f))\, df \quad \text{bits/s.} \tag{30}$$

If each Gaussian variable x, y and η is also iid then $I(x, y) = C$ bits/s, otherwise $I(x, y) < C$ bits/s[2]. If the peak power at all frequencies is a constant k then it can be shown

that $I(x, y)$ is maximised when $S(f) + N(f) = k$, which defines a flat power spectrum. Finally, if the signal spectrum is sculpted so that the signal plus noise spectrum is flat then the logarithmic relation in Equation 23 yields improved, albeit still diminishing, returns[7] $C \propto (S/N)^{1/3}$ bits/s.

14 A Very Short History of Information Theory

Even the most gifted scientist cannot command an original theory out of thin air. Just as Einstein could not have devised his theories of relativity if he had no knowledge of Newton's work, so Shannon could not have created information theory if he had no knowledge of the work of Boltzmann (1875) and Gibbs (1902) on thermodynamic entropy, Wiener (1927) on signal processing, Nyquist (1928) on sampling theory, or Hartley (1928) on information transmission[8].

Even though Shannon was not alone in trying to solve one of the key scientific problems of his time (i.e. how to define and measure information), he was alone in being able to produce a complete mathematical theory of information: a theory that might otherwise have taken decades to construct. In effect, Shannon single-handedly accelerated the rate of scientific progress, and it is entirely possible that, without his contribution, we would still be treating information as if it were some ill-defined vital fluid.

15 Key Equations

Logarithms use base 2 unless stated otherwise.

Entropy

$$H(x) = \sum_{i=1}^{m} p(x_i) \log \frac{1}{p(x_i)} \text{ bits} \tag{31}$$

$$H(x) = \int_x p(x) \log \frac{1}{p(x)} \, dx \text{ bits} \tag{32}$$

Joint entropy

$$H(x, y) = \sum_{i=1}^{m} \sum_{j=1}^{m} p(x_i, y_j) \log \frac{1}{p(x_i, y_j)} \text{ bits} \tag{33}$$

$$H(x, y) = \int_x \int_y p(y, x) \log \frac{1}{p(y, x)} \, dy \, dx \text{ bits} \tag{34}$$

$$H(x, y) = I(x, y) + H(x|y) + H(y|x) \text{ bits} \tag{35}$$

Conditional Entropy

$$H(y|x) = \sum_{i=1}^{m}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{1}{p(x_i|y_j)} \text{ bits} \quad (36)$$

$$H(y|x) = \sum_{i=1}^{m}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{1}{p(y_j|x_i)} \text{ bits} \quad (37)$$

$$H(x|y) = \int_y \int_x p(x, y) \log \frac{1}{p(x|y)} \, dx \, dy \text{ bits} \quad (38)$$

$$H(y|x) = \int_y \int_x p(x, y) \log \frac{1}{p(y|x)} \, dx \, dy \text{ bits} \quad (39)$$

$$H(x|y) = H(x, y) - H(y) \text{ bits} \quad (40)$$

$$H(y|x) = H(x, y) - H(x) \text{ bits} \quad (41)$$

from which we obtain the *chain rule for entropy*

$$H(x, y) = H(x) + H(y|x) \text{ bits} \quad (42)$$

$$= H(y) + H(x|y) \text{ bits} \quad (43)$$

Marginalisation

$$p(x_i) = \sum_{j=1}^{m} p(x_i, y_j), \qquad p(y_j) = \sum_{i=1}^{m} p(x_i, y_j) \quad (44)$$

$$p(x) = \int_y p(x, y) \, dy, \qquad p(y) = \int_x p(x, y) \, dx \quad (45)$$

Mutual Information

$$I(x, y) = \sum_{i=1}^{m}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \text{ bits} \quad (46)$$

$$I(x, y) = \int_y \int_x p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dx \, dy \text{ bits} \quad (47)$$

$$I(x, y) = H(x) + H(y) - H(x, y) \quad (48)$$

$$= H(x) - H(x|y) \quad (49)$$

$$= H(y) - H(y|x) \quad (50)$$

$$= H(x, y) - [H(x|y) + H(y|x)] \text{ bits} \quad (51)$$

If $y = x + \eta$, with x and y (not necessarily iid) Gaussian variables then

$$I(x,y) \quad = \quad \int_{f=0}^{W} \log\left(1 + \frac{S(f)}{N(f)}\right) df \quad \text{bits/s}, \tag{52}$$

where W is the bandwidth, $S(f)/N(f)$ is the signal to noise ratio of the signal and noise Fourier components at frequency f (Section 13), and data are transmitted at the Nyquist rate of $2W$ samples/s.

Channel Capacity

$$C \quad = \quad \max_{p(x)} I(x,y) \text{ bits per value.} \tag{53}$$

If the channel input x has variance S, the noise η has variance N, and both x and η are iid Gaussian variables then $I(x,y) = C$, where

$$C \quad = \quad \frac{1}{2}\log\left(1 + \frac{S}{N}\right) \text{ bits per value}, \tag{54}$$

where the ratio of variances S/N is the signal to noise ratio.

Further Reading

Applebaum D (2008)[9]. Probability and Information: An Integrated Approach. *A thorough introduction to information theory, which strikes a good balance between intuitive and technical explanations.*

Avery J (2012)[10]. Information Theory and Evolution. *An engaging account of how information theory is relevant to a wide range of natural and man-made systems, including evolution, physics, culture and genetics. Includes interesting background stories on the development of ideas within these different disciplines.*

Baeyer HV (2005)[11]. Information: The New Language of Science *Erudite, wide-ranging, and insightful account of information theory. Contains no equations, which makes it very readable.*

Cover T and Thomas J (1991)[12]. Elements of Information Theory. *Comprehensive, and highly technical, with historical notes and an equation summary at the end of each chapter.*

Ghahramani Z (2002). Information Theory. Encyclopedia of Cognitive Science. *An excellent, brief overview of information.*

Gleick J (2012)[13]. The Information. *An informal introduction to the history of ideas and people associated with information theory.*

Guizzo EM (2003)[14]. The Essential Message: Claude Shannon and the Making of Information Theory. Master's Thesis, Massachusetts Institute of Technology. *One of the few accounts of Shannon's role in the development of information theory. See* `http://dspace.mit.edu/bitstream/handle/1721.1/39429/54526133.pdf`.

Laughlin, SB (2006). The Hungry Eye: Energy, Information and Retinal Function, *Excellent lecture on the energy cost of Shannon information in eyes. See* `http://www.crsltd.com/guest-talks/crs-guest-lecturers/simon-laughlin`.

MacKay DJC (2003)[15]. Information Theory, Inference, and Learning Algorithms. *The modern classic on information theory. A very readable text that roams far and wide over many topics. The book's web site (below) also has a link to an excellent series of video lectures by MacKay. Available free online at* `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

Pierce JR (1980)[8]. An Introduction to Information Theory: Symbols, Signals and Noise. Second Edition. *Pierce writes with an informal, tutorial style of writing, but does not flinch from presenting the fundamental theorems of information theory. This book provides a good balance between words and equations.*

Reza FM (1961)[3]. An Introduction to Information Theory. *A more comprehensive and mathematically rigorous book than Pierce's book, it should be read only after first reading Pierce's more informal text.*

Seife C (2007)[16]. Decoding the Universe: How the New Science of Information Is Explaining Everything in the Cosmos, From Our Brains to Black Holes. *A lucid and engaging account of the relationship between information, thermodynamic entropy and quantum computing. Highly recommended.*

Shannon CE and Weaver W (1949)[2]. The Mathematical Theory of Communication. University of Illinois Press. *A surprisingly accessible book, written in an era when information theory was known only to a privileged few. This book can be downloaded from* `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html`

Soni, J and Goodman, R (2017)[17]. A mind at play: How Claude Shannon invented the information age *A biography of Shannon.*

Stone, JV (2015)[4] Information Theory: A Tutorial Introduction. *A more extensive introduction than the current article.*

For the complete novice, the videos at the online Kahn Academy provide an excellent introduction. Additionally, the online Scholarpedia web page by Latham and Rudi provides

a lucid technical account of mutual information:
`http://www.scholarpedia.org/article/Mutual_information`.

Finally, some historical perspective is provided in a long interview with Shannon conducted in 1982: `http://www.ieeeghn.org/wiki/index.php/Oral-History:Claude_E._Shannon`.

References

[1] CE Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[2] CE Shannon and W Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[3] FM Reza. *Information Theory*. New York, McGraw-Hill, 1961.

[4] JV Stone. *Information Theory: A Tutorial Introduction*. Sebtel Press, Sheffield, England, 2015.

[5] SR Schultz. Signal-to-noise ratio in neuroscience. *Scholarpedia*, 2(6):2046, 2007.

[6] H. Nyquist. Certain topics in telegraph transmission theory. *Proceedings of the IEEE*, 90(2):280–305, 1928.

[7] F Rieke, D Warland, RR de Ruyter van Steveninck, and W Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA, 1997.

[8] JR Pierce. *An introduction to information theory: Symbols, signals and noise*. Dover, 1980.

[9] D Applebaum. *Probability and Information An Integrated Approach, 2nd Edition*. Cambridge University Press, Cambridge, 2008.

[10] J Avery. *Information Theory and Evolution*. World Scientific Publishing, 2012.

[11] HV Baeyer. *Information: The New Language of Science*. Harvard University Press, 2005.

[12] TM Cover and JA Thomas. *Elements of Information Theory*. New York, John Wiley and Sons, 1991.

[13] J Gleick. *The Information*. Vintage, 2012.

[14] EM Guizzo. The essential message: Claude Shannon and the making of information theory. *http://dspace.mit.edu/bitstream/handle/1721.1/39429/54526133.pdf. Massachusetts Institute of Technology*, 2003.

[15] DJC MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[16] C Seife. *Decoding the Universe: How the New Science of Information Is Explaining Everything in the Cosmos, From Our Brains to Black Holes.* Penguin, 2007.

[17] J Soni and R Goodman. *A mind at play: How Claude Shannon invented the information age.* Simon and Schuster, 2017.

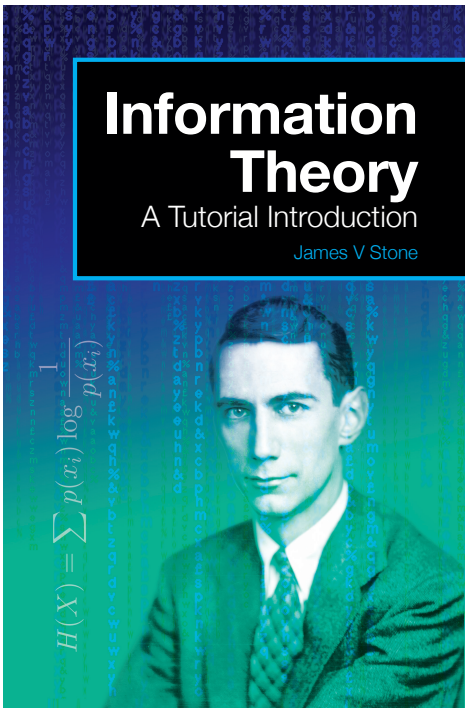


Figure 11: This paper is derived from the book Information Theory. For details, see https://jim-stone.staff.shef.ac.uk/BookInfoTheory/InfoTheoryBookMain.html