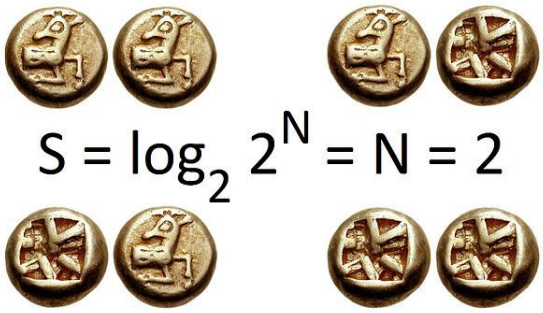
Entropy (information theory)



2 bits of entropy.

In information theory, **entropy** is the average amount of information contained in each message received. Here, *message* stands for an event, sample or character drawn from a distribution or data stream. Entropy thus characterizes our uncertainty about our source of information. (Entropy is best understood as a measure of uncertainty rather than certainty as entropy is larger for more random sources.) The source is also characterized by the probability distribution of the samples drawn from it. The idea here is that the less likely an event is, the more information it provides when it occurs. For some other reasons (explained below) it makes sense to define information as the negative of the logarithm of the probability distribution. The probability distribution of the events, coupled with the information amount of every event, forms a random variable whose average (a.k.a. expected) value is the average amount of information, a.k.a. entropy, generated by this distribution. Because entropy is average information, it is also measured in shannons, nats, or hartleys, depending on the base of the logarithm used to define it.

The logarithm of the probability distribution is useful as measure of information because of its additivity. For instance, flipping a coin provides 1 shannon of information whereas m tosses gather m bits. Generally, you need $\log_2(n)$ bits to represent a variable that can take one of n values. Since 1 of n outcomes is possible when you apply a scale graduated with n marks, you receive $\log_2(n)$ bits of information with every such measurement. The $\log_2(n)$ rule holds only until all outcomes are equally probable. If one of the events occurs more often than others, observation of that event is less informative. Conversely, observing rarer events compensate by providing more information when observed. Since observation of less probable events occurs more rarely, the net effect is that the entropy (thought of as the average information) received from non-uniformly distributed data is less than $\log_2(n)$. Entropy is zero when only one certain outcome is expected. Shannon entropy quantifies all these considerations exactly when a probability distribution of the source is provided. It is important to note that the meaning of the events observed (a.k.a. the meaning of *messages*) do not matter in the definition of entropy. Entropy only takes into account the probability of observing a specific event, so the information it encapsulates is information about the underlying probability distribution, not the meaning of the events themselves.

Generally, "entropy" stands for "disorder" or uncertainty. The entropy we talk about here was introduced by Claude E. Shannon in his 1948 paper "A Mathematical Theory of Communication".[1] We also call it **Shannon entropy** to distinguish from other occurrences of the term, which appears in various parts of physics in different forms. Shannon entropy provides an absolute limit on the best possible average length of lossless encoding or compression of any communication, assuming that[2] the communication may be represented as a sequence of independent and identically distributed random variables.

1 Introduction

Entropy is a measure of *unpredictability* of *information content*. To get an informal, intuitive understanding of the connection between these three English terms, consider the example of a poll on some political issue. Usually, such polls happen because the outcome of the poll isn't already known. In other words, the outcome of the poll is relatively *unpredictable*, and actually performing the poll and learning the results gives some new *information*; these are just different ways of saying that the *entropy* of the poll results is large. Now, consider the case that the same poll is performed a second time shortly after the first poll. Since the result of the first poll is already known, the outcome of the second poll can be predicted well and the results should not contain much new information; in this case the entropy of the second poll results is small.

Now consider the example of a coin toss. When the coin is fair, that is, when the probability of heads is the same as the probability of tails, then the entropy of the coin toss is as high as it could be. This is because there is no way to predict the outcome of the coin toss ahead of time—the best we can do is predict that the coin will come up heads, and our prediction will be correct with probability

1/2. Such a coin toss has one bit of entropy since there are two possible outcomes that occur with equal probability, and learning the actual outcome contains one bit of information. Contrarily, a coin toss with a coin that has two heads and no tails has zero entropy since the coin will always come up heads, and the outcome can be predicted perfectly.

English text has fairly low entropy. In other words, it is fairly predictable. Even if we don't know exactly what is going to come next, we can be fairly certain that, for example, there will be many more e's than z's, that the combination 'qu' will be much more common than any other combination with a 'q' in it, and that the combination 'th' will be more common than 'z', 'q', or 'qu'. After the first few letters one can often guess the rest of the word. Uncompressed, English text has between 0.6 and 1.3 bits of entropy for each character of message.[3][4]

If a compression scheme is lossless—that is, you can always recover the entire original message by decompressing—then a compressed message has the same quantity of information as the original, but communicated in fewer characters. That is, it has more information, or a higher entropy, per character. This means a compressed message has less redundancy. Roughly speaking, Shannon's source coding theorem says that a lossless compression scheme cannot compress messages, on average, to have more than one bit of information per bit of message. The entropy of a message multiplied by the length of that message is a measure of how much information the message contains.

Shannon's theorem also implies that no lossless compression scheme can compress *all* messages. If some messages come out smaller, at least one must come out larger due to the pigeonhole principle. In practical use, this is generally not a problem, because we are usually only interested in compressing certain types of messages, for example English documents as opposed to gibberish text, or digital photographs rather than noise, and it is unimportant if a compression algorithm makes some unlikely or uninteresting sequences larger. However, the problem can still arise even in everyday use when applying a compression algorithm to already compressed data: for example, making a ZIP file of music in the FLAC audio format is unlikely to achieve much extra saving in space.

2 Definition

Named after Boltzmann's H-theorem, Shannon defined the entropy H (Greek letter Eta) of a discrete random variable X with possible values $\{x_1, ..., xn\}$ and probability mass function $P(X)$ as:

$$H(X) = E[I(X)] = E[-\ln(P(X))].$$

Here E is the expected value operator, and I is the information content of X.[5][6] $I(X)$ is itself a random variable.

When taken from a finite sample, the entropy can explicitly be written as

$$H(X) = \sum_i P(x_i)\, I(x_i) = -\sum_i P(x_i) \log_b P(x_i)$$

where b is the base of the logarithm used. Common values of b are 2, Euler's number e, and 10, and the unit of entropy is bit for $b = 2$, nat for $b = e$, and dit (or digit) for $b = 10$.[7]

In the case of $p(xi) = 0$ for some i, the value of the corresponding summand $0 \log b(0)$ is taken to be 0, which is consistent with the well-known limit:

$$\lim_{p \to 0+} p \log(p) = 0$$

One may also define the conditional entropy of two events X and Y taking values xi and yj respectively, as

$$H(X|Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)}$$

where $p(xi,yj)$ is the probability that $X=xi$ and $Y=yj$. This quantity should be understood as the amount of randomness in the random variable X given that you know the value of Y.

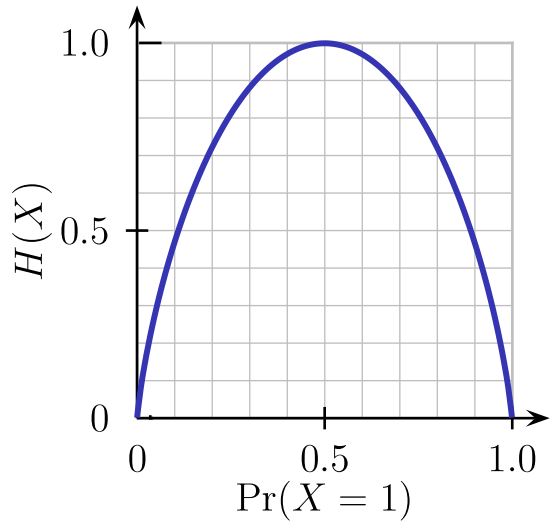
3 Example

Main article: Binary entropy function
Main article: Bernoulli process

Consider tossing a coin with known, not necessarily fair, probabilities of coming up heads or tails; this is known as the Bernoulli process.

The entropy of the unknown result of the next toss of the coin is maximized if the coin is fair (that is, if heads and tails both have equal probability 1/2). This is the situation of maximum uncertainty as it is most difficult to predict the outcome of the next toss; the result of each toss of the coin delivers one full bit of information.

However, if we know the coin is not fair, but comes up heads or tails with probabilities p and q, where $p \neq q$, then there is less uncertainty. Every time it is tossed, one side is more likely to come up than the other. The reduced uncertainty is quantified in a lower entropy: on average each toss of the coin delivers less than one full bit of information.



Entropy H(X) *(i.e. the expected surprisal) of a coin flip, measured in shannons, graphed versus the fairness of the coin Pr(*X*=1), where* X*=1 represents a result of heads.*
Note that the maximum of the graph depends on the distribution. Here, the entropy is at most 1 shannon, and to communicate the outcome of a fair coin flip (2 possible values) will require an average of at most 1 bit. The result of a fair die (6 possible values) would require on average $\log_2 6$ *bits.*

The extreme case is that of a double-headed coin that never comes up tails, or a double-tailed coin that never results in a head. Then there is no uncertainty. The entropy is zero: each toss of the coin delivers no new information as the outcome of each coin toss is always certain. In this respect, entropy can be normalized by dividing it by information length. This ratio is called metric entropy and is a measure of the randomness of the information.

4 Rationale

To understand the meaning of $\sum p_i \log \frac{1}{p_i}$, at first, try to define information function, *I*, in terms of event *i* probability, p_i. How much information do we receive when event *i* is observed? Shannon answered the way we see in the characterization:[8]

1. $I(p) \geq 0$ – information is non-negative quantity
2. $I(1) = 0$ – event that always happens does not communicate information
3. $I(p_1\ p_2) = I(p_1) + I(p_2)$ – additivity of independent events

The latter is a crucial property. It says that joint probability communicates as much information as two individual events separately. Others spell it the way that if first event can happen in one of n equiprobable outcomes and another is 1 of *m* equiprobable outcomes then joint event is 1 of *mn* possible. This means that you need $\log_2(n)$ bit to encode the first value and $\log_2(m)$ to encode the second, you need $\log_2(mn) = \log_2(m) + \log_2(n)$ to encode both. This reveals that log function captures the additivity of information. Indeed, Shannon has discovered that logarithm is the function can serve as information function of event if we put

$$I(p) = \log(1/p)$$

The base of logarithm does not matter. You can choose any. The different units of information (bits for $\log_2$, trits for $\log_3$, nats for ln and so on) are just constant multiples of each other. For instance, in case of tossing a fair coin, heads provides $\log_2(2) = 1$ bit of information. Because of additivity, n tosses provide n bits of information. That is how we measure information indeed.

Now, suppose we have a distribution where every event can happen with probability *pi*. Suppose we have sampled it *N* times and outcome *i* was, thus, seen $n_i = Np_i$ times. The total amount of information we received is

$$\sum n_i I(p_i) = \sum (Np_i) \log(1/p_i)$$

and average amount of information that we receive with every event is N times less

$$\sum p_i \log \frac{1}{p_i}.$$

5 Aspects

5.1 Relationship to thermodynamic entropy

Main article: Entropy in thermodynamics and information theory

The inspiration for adopting the word *entropy* in information theory came from the close resemblance between Shannon's formula and very similar known formulae from statistical mechanics.

In statistical thermodynamics the most general formula for the thermodynamic entropy *S* of a thermodynamic system is the Gibbs entropy,

$$S = -k_\mathrm{B} \sum p_i \ln p_i$$

where *k*B is the Boltzmann constant, and *pi* is the probability of a microstate. The Gibbs entropy was defined by J. Willard Gibbs in 1878 after earlier work by Boltzmann (1872).[9]

The Gibbs entropy translates over almost unchanged into the world of quantum physics to give the von Neumann entropy, introduced by John von Neumann in 1927,

$$S = -k_{\mathrm{B}}\,\mathrm{Tr}(\rho \ln \rho)$$

where ρ is the density matrix of the quantum mechanical system and Tr is the trace.

At an everyday practical level the links between information entropy and thermodynamic entropy are not evident. Physicists and chemists are apt to be more interested in *changes* in entropy as a system spontaneously evolves away from its initial conditions, in accordance with the second law of thermodynamics, rather than an unchanging probability distribution. And, as the minuteness of Boltzmann's constant *kB* indicates, the changes in *S*/*kB* for even tiny amounts of substances in chemical and physical processes represent amounts of entropy that are extremely large compared to anything in data compression or signal processing. Furthermore, in classical thermodynamics the entropy is defined in terms of macroscopic measurements and makes no reference to any probability distribution, which is central to the definition of information entropy.

At a multidisciplinary level, however, connections can be made between thermodynamic and informational entropy, although it took many years in the development of the theories of statistical mechanics and information theory to make the relationship fully apparent. In fact, in the view of Jaynes (1957), thermodynamic entropy, as explained by statistical mechanics, should be seen as an *application* of Shannon's information theory: the thermodynamic entropy is interpreted as being proportional to the amount of further Shannon information needed to define the detailed microscopic state of the system, that remains uncommunicated by a description solely in terms of the macroscopic variables of classical thermodynamics, with the constant of proportionality being just the Boltzmann constant. For example, adding heat to a system increases its thermodynamic entropy because it increases the number of possible microscopic states of the system that are consistent with the measurable values of its macroscopic variables, thus making any complete state description longer. (See article: *maximum entropy thermodynamics*). Maxwell's demon can (hypothetically) reduce the thermodynamic entropy of a system by using information about the states of individual molecules; but, as Landauer (from 1961) and co-workers have shown, to function the demon himself must increase thermodynamic entropy in the process, by at least the amount of Shannon information he proposes to first acquire and store; and so the total thermodynamic entropy does not decrease (which resolves the paradox). Landauer's principle has implications on the amount of heat a computer must dissipate to process a given amount of information, though modern computers are nowhere near the efficiency limit.

5.2 Entropy as information content

Main article: Shannon's source coding theorem

Entropy is defined in the context of a probabilistic model. Independent fair coin flips have an entropy of 1 bit per flip. A source that always generates a long string of B's has an entropy of 0, since the next character will always be a 'B'.

The entropy rate of a data source means the average number of bits per symbol needed to encode it. Shannon's experiments with human predictors show an information rate between 0.6 and 1.3 bits per character in English;[10] the PPM compression algorithm can achieve a compression ratio of 1.5 bits per character in English text.

From the preceding example, note the following points:

1. The amount of entropy is not always an integer number of bits.

2. Many data bits may not convey information. For example, data structures often store information redundantly, or have identical sections regardless of the information in the data structure.

Shannon's definition of entropy, when applied to an information source, can determine the minimum channel capacity required to reliably transmit the source as encoded binary digits (see caveat below in italics). The formula can be derived by calculating the mathematical expectation of the *amount of information* contained in a digit from the information source. *See also* Shannon-Hartley theorem.

Shannon's entropy measures the information contained in a message as opposed to the portion of the message that is determined (or predictable). *Examples of the latter include redundancy in language structure or statistical properties relating to the occurrence frequencies of letter or word pairs, triplets etc.* See Markov chain.

5.3 Data compression

Main article: Data compression

Entropy effectively bounds the performance of the strongest lossless compression possible, which can be realized in theory by using the typical set or in practice using Huffman, Lempel-Ziv or arithmetic coding. The performance of existing data compression algorithms is often used as a rough estimate of the entropy of a block of data.[11][12] See also Kolmogorov complexity. In practice, compression algorithms deliberately include some judicious redundancy in the form of checksums to protect against errors.

5.4 World's technological capacity to store and communicate entropic information

A 2011 study in *Science* estimates the world's technological capacity to store and communicate optimally compressed information normalized on the most effective compression algorithms available in the year 2007, therefore estimating the entropy of the technologically available sources.[13]

The authors estimate humankind technological capacity to store information (fully entropically compressed) in 1986 and again in 2007. They break the information into three categories - To store information on a medium, to receive information through a one-way broadcast networks, to exchange information through two-way telecommunication networks.[13]

5.5 Limitations of entropy as information content

There are a number of entropy-related concepts that mathematically quantify information content in some way:

- the **self-information** of an individual message or symbol taken from a given probability distribution,
- the **entropy** of a given probability distribution of messages or symbols, and
- the **entropy rate** of a stochastic process.

(The "rate of self-information" can also be defined for a particular sequence of messages or symbols generated by a given stochastic process: this will always be equal to the entropy rate in the case of a stationary process.) Other quantities of information are also used to compare or relate different sources of information.

It is important not to confuse the above concepts. Often it is only clear from context which one is meant. For example, when someone says that the "entropy" of the English language is about 1 bit per character, they are actually modeling the English language as a stochastic process and talking about its entropy *rate*.

Although entropy is often used as a characterization of the information content of a data source, this information content is not absolute: it depends crucially on the probabilistic model. A source that always generates the same symbol has an entropy rate of 0, but the definition of what a symbol is depends on the alphabet. Consider a source that produces the string ABABABABAB... in which A is always followed by B and vice versa. If the probabilistic model considers individual letters as independent, the entropy rate of the sequence is 1 bit per character. But if the sequence is considered as "AB AB AB AB AB..." with symbols as two-character blocks, then the entropy rate is 0 bits per character.

However, if we use very large blocks, then the estimate of per-character entropy rate may become artificially low. This is because in reality, the probability distribution of the sequence is not knowable exactly; it is only an estimate. For example, suppose one considers the text of every book ever published as a sequence, with each symbol being the text of a complete book. If there are N published books, and each book is only published once, the estimate of the probability of each book is $1/N$, and the entropy (in bits) is $-\log_2(1/N) = \log_2(N)$. As a practical code, this corresponds to assigning each book a unique identifier and using it in place of the text of the book whenever one wants to refer to the book. This is enormously useful for talking about books, but it is not so useful for characterizing the information content of an individual book, or of language in general: it is not possible to reconstruct the book from its identifier without knowing the probability distribution, that is, the complete text of all the books. The key idea is that the complexity of the probabilistic model must be considered. Kolmogorov complexity is a theoretical generalization of this idea that allows the consideration of the information content of a sequence independent of any particular probability model; it considers the shortest program for a universal computer that outputs the sequence. A code that achieves the entropy rate of a sequence for a given model, plus the codebook (i.e. the probabilistic model), is one such program, but it may not be the shortest.

For example, the Fibonacci sequence is 1, 1, 2, 3, 5, 8, 13, ... Treating the sequence as a message and each number as a symbol, there are almost as many symbols as there are characters in the message, giving an entropy of approximately $\log_2(n)$. So the first 128 symbols of the Fibonacci sequence has an entropy of approximately 7 bits/symbol. However, the sequence can be expressed using a formula [$F(n) = F(n-1) + F(n-2)$ for $n=\{3,4,5,...\}$, $F(1)=1$, $F(2)=1$] and this formula has a much lower entropy and applies to any length of the Fibonacci sequence.

5.6 Limitations of entropy as a measure of unpredictability

In cryptanalysis, entropy is often roughly used as a measure of the unpredictability of a cryptographic key. For example, a 128-bit key that is randomly generated has 128 bits of entropy. It takes (on average) 2^{128-1} guesses to break by brute force. If the key's first digit is 0, and the others random, then the entropy is 127 bits, and it takes (on average) 2^{127-1} guesses.

However, entropy fails to capture the number of guesses required if the possible keys are not of equal probability.[14][15] If the key is half the time "password" and half the time a true random 128-bit key, then the entropy is approximately 65 bits. Yet half the time the key

may be guessed on the first try, if your first guess is "password", and on average, it takes around 2^{126} guesses (not 2^{65-1}) to break this password.

Similarly, consider a 1000000-digit binary one-time pad. If the pad has 1000000 bits of entropy, it is perfect. If the pad has 999999 bits of entropy, evenly distributed (each individual bit of the pad having 0.999999 bits of entropy) it may still be considered very good. But if the pad has 999999 bits of entropy, where the first digit is fixed and the remaining 999999 digits are perfectly random, then the first digit of the ciphertext will not be encrypted at all.

5.7 Data as a Markov process

A common way to define entropy for text is based on the Markov model of text. For an order-0 source (each character is selected independent of the last characters), the binary entropy is:

$$H(\mathcal{S}) = -\sum p_i \log_2 p_i,$$

where *pi* is the probability of *i*. For a first-order Markov source (one in which the probability of selecting a character is dependent only on the immediately preceding character), the **entropy rate** is:

$$H(\mathcal{S}) = -\sum_i p_i \sum_j \ p_i(j) \log_2 p_i(j),$$

where *i* is a **state** (certain preceding characters) and $p_i(j)$ is the probability of *j* given *i* as the previous character.

For a second order Markov source, the entropy rate is

$$H(\mathcal{S}) = -\sum_i p_i \sum_j p_i(j) \sum_k p_{i,j}(k) \ \log_2 \ p_{i,j}(k).$$

5.8 *b*-ary entropy

In general the ***b*-ary entropy** of a source $\mathcal{S}$ = (*S*,*P*) with source alphabet *S* = {a_1, ..., *an*} and discrete probability distribution *P* = {p_1, ..., *pn*} where *pi* is the probability of *ai* (say *pi* = *p*(*ai*)) is defined by:

$$H_b(\mathcal{S}) = -\sum_{i=1}^{n} p_i \log_b p_i,$$

Note: the *b* in "*b*-ary entropy" is the number of different symbols of the *ideal alphabet* used as a standard yardstick to measure source alphabets. In information theory, two symbols are necessary and sufficient for an alphabet to encode information. Therefore, the default is to let *b* = 2 ("binary entropy"). Thus, the entropy of the source alphabet, with its given empiric probability distribution, is a number equal to the number (possibly fractional) of symbols of the "ideal alphabet", with an optimal probability distribution, necessary to encode for each symbol of the source alphabet. Also note that "optimal probability distribution" here means a uniform distribution: a source alphabet with *n* symbols has the highest possible entropy (for an alphabet with *n* symbols) when the probability distribution of the alphabet is uniform. This optimal entropy turns out to be log*b*(*n*).

6 Efficiency

A source alphabet with non-uniform distribution will have less entropy than if those symbols had uniform distribution (i.e. the "optimized alphabet"). This deficiency in entropy can be expressed as a ratio called efficiency:

$$\eta(X) = -\sum_{i=1}^{n} \frac{p(x_i) \log_b(p(x_i))}{\log_b(n)}$$

Efficiency has utility in quantifying the effective use of a communications channel. This formulation is also referred to as the normalized entropy, as the entropy is divided by the maximum entropy $\log_b(n)$.

7 Characterization

Shannon entropy is characterized by a small number of criteria, listed below. Any definition of entropy satisfying these assumptions has the form

$$-K \sum_{i=1}^{n} p_i \log(p_i)$$

where *K* is a constant corresponding to a choice of measurement units.

In the following, *pi* = Pr (*X* = *xi*) and $H_n(p_1, \ldots, p_n) = H(X)$.

7.1 Continuity

The measure should be continuous, so that changing the values of the probabilities by a very small amount should only change the entropy by a small amount.

7.2 Symmetry

The measure should be unchanged if the outcomes *xi* are re-ordered.

$$H_n(p_1, p_2, \ldots) = H_n(p_2, p_1, \ldots)$$

7.3 Maximum

The measure should be maximal if all the outcomes are equally likely (uncertainty is highest when all possible events are equiprobable).

$$H_n(p_1, \ldots, p_n) \leq H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_b(n).$$

For equiprobable events the entropy should increase with the number of outcomes.

$$H_n\left(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n}\right) = \log_b(n) < \log_b(n+1) = H_{n+1}\left(\underbrace{\frac{1}{n+1}, \ldots, \frac{1}{n+1}}_{n+1}\right)$$

7.4 Additivity

The amount of entropy should be independent of how the process is regarded as being divided into parts.

This last functional relationship characterizes the entropy of a system with sub-systems. It demands that the entropy of a system can be calculated from the entropies of its sub-systems if the interactions between the sub-systems are known.

Given an ensemble of *n* uniformly distributed elements that are divided into *k* boxes (sub-systems) with b_1, ..., *bk* elements each, the entropy of the whole ensemble should be equal to the sum of the entropy of the system of boxes and the individual entropies of the boxes, each weighted with the probability of being in that particular box.

For positive integers *bi* where $b_1 + \ldots + bk = n$,

$$H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = H_k\left(\frac{b_1}{n}, \ldots, \frac{b_k}{n}\right) + \sum_{i=1}^{k} \frac{b_i}{n} H_{b_i}\left(\frac{1}{b_i}, \ldots, \frac{1}{b_i}\right)$$

Choosing $k = n$, $b_1 = \ldots = bn = 1$ this implies that the entropy of a certain outcome is zero: $H_1(1) = 0$. This implies that the efficiency of a source alphabet with *n* symbols can be defined simply as being equal to its *n*-ary entropy. See also Redundancy (information theory).

8 Further properties

The Shannon entropy satisfies the following properties, for some of which it is useful to interpret entropy as the amount of information learned (or uncertainty eliminated) by revealing the value of a random variable *X*:

- Adding or removing an event with probability zero does not contribute to the entropy:

$$H_{n+1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n)$$

- It can be confirmed using the Jensen inequality that

$$H(X) = \mathrm{E}\left[\log_b\left(\frac{1}{p(X)}\right)\right] \leq \log_b\left(\mathrm{E}\left[\frac{1}{p(X)}\right]\right) = \log_b(n)$$

  This maximal entropy of log*b*(*n*) is effectively attained by a source alphabet having a uniform probability distribution: uncertainty is maximal when all possible events are equiprobable.

- The entropy or the amount of information revealed by evaluating (*X*,*Y*) (that is, evaluating *X* and *Y* simultaneously) is equal to the information revealed by conducting two consecutive experiments: first evaluating the value of *Y*, then revealing the value of *X* given that you know the value of *Y*. This may be written as

$$H[(X, Y)] = H(X|Y) + H(Y) = H(Y|X) + H(X).$$

- If *Y=f(X)* where *f* is deterministic, then *H*(*f*(*X*)|*X*) = 0. Applying the previous formula to *H*(*X*, *f*(*X*)) yields

$$H(X) + H(f(X)|X) = H(f(X)) + H(X|f(X)),$$

  so $H(f(X)) \leq H(X)$, thus the entropy of a variable can only decrease when the latter is passed through a deterministic function.

- If *X* and *Y* are two independent experiments, then knowing the value of *Y* doesn't influence our knowledge of the value of *X* (since the two don't influence each other by independence):

$$H(X|Y) = H(X).$$

- The entropy of two simultaneous events is no more than the sum of the entropies of each individual event, and are equal if the two events are independent. More specifically, if *X* and *Y* are two random variables on the same probability space, and *(X,Y)* denotes their Cartesian product, then

$$H[(X, Y)] \leq H(X) + H(Y).$$

Proving this mathematically follows easily from the previous two properties of entropy.

9 Extending discrete entropy to the continuous case: differential entropy

Main article: Differential entropy

The Shannon entropy is restricted to random variables taking discrete values. The corresponding formula for a continuous random variable with probability density function *f(x)* with finite or infinite support $\mathbb{X}$ on the real line is defined by analogy, using the above form of the entropy as an expectation:

$$h[f] = \mathrm{E}[-\ln(f(x))] = -\int_{\mathbb{X}} f(x)\ln(f(x))\,dx.$$

This formula is usually referred to as the **continuous entropy**, or differential entropy. A precursor of the continuous entropy $h[f]$ is the expression for the functional H in the H-theorem of Boltzmann.

Although the analogy between both functions is suggestive, the following question must be set: is the differential entropy a valid extension of the Shannon discrete entropy? Differential entropy lacks a number of properties that the Shannon discrete entropy has – it can even be negative – and thus corrections have been suggested, notably limiting density of discrete points.

To answer this question, we must establish a connection between the two functions:

We wish to obtain a generally finite measure as the bin size goes to zero. In the discrete case, the bin size is the (implicit) width of each of the n (finite or infinite) bins whose probabilities are denoted by p_n. As we generalize to the continuous domain, we must make this width explicit.

To do this, start with a continuous function f discretized into bins of size Δ . By the mean-value theorem there exists a value x_i in each bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx$$

and thus the integral of the function f can be approximated (in the Riemannian sense) by

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{\Delta\to 0}\sum_{i=-\infty}^{\infty} f(x_i)\Delta$$

where this limit and "bin size goes to zero" are equivalent.

We will denote

$$H^{\Delta} := -\sum_{i=-\infty}^{\infty} f(x_i)\Delta\log\left(f(x_i)\Delta\right)$$

and expanding the logarithm, we have

$$H^{\Delta} = -\sum_{i=-\infty}^{\infty} f(x_i)\Delta\log(f(x_i)) - \sum_{i=-\infty}^{\infty} f(x_i)\Delta\log(\Delta).$$

As $\Delta \to 0$, we have

$$\sum_{i=-\infty}^{\infty} f(x_i)\Delta \to \int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\sum_{i=-\infty}^{\infty} f(x_i)\Delta\log(f(x_i)) \to \int_{-\infty}^{\infty} f(x)\log f(x)\,dx.$$

But note that $\log(\Delta) \to -\infty$ as $\Delta \to 0$, therefore we need a special definition of the differential or continuous entropy:

$$h[f] = \lim_{\Delta\to 0}\left(H^{\Delta} + \log\Delta\right) = -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx,$$

which is, as said before, referred to as the **differential entropy**. This means that the differential entropy *is not* a limit of the Shannon entropy for $n \to \infty$. Rather, it differs from the limit of the Shannon entropy by an infinite offset.

It turns out as a result that, unlike the Shannon entropy, the differential entropy is *not* in general a good measure of uncertainty or information. For example, the differential entropy can be negative; also it is not invariant under continuous co-ordinate transformations.

10 Relative entropy

Main article: Generalized relative entropy

Another useful measure of entropy that works equally well in the discrete and the continuous case is the **relative entropy** of a distribution. It is defined as the Kullback–Leibler divergence from the distribution to a reference measure m as follows. Assume that a probability distribution p is absolutely continuous with respect to a measure m, i.e. is of the form $p(dx) = f(x)m(dx)$ for some non-negative m-integrable function f with m-integral 1, then the relative entropy can be defined as

$$D_{\mathrm{KL}}(p\|m) = \int \log(f(x))p(dx) = \int f(x)\log(f(x))m(dx).$$

In this form the relative entropy generalises (up to change in sign) both the discrete entropy, where the measure m is the counting measure, and the differential entropy, where the measure m is the Lebesgue measure. If the measure m is itself a probability distribution, the relative entropy is non-negative, and zero if $p = m$ as measures. It is defined for any measure space, hence coordinate independent and invariant under co-ordinate reparameterizations

if one properly takes into account the transformation of the measure m. The relative entropy, and implicitly entropy and differential entropy, do depend on the "reference" measure m.

11 Use in combinatorics

Entropy has become a useful quantity in combinatorics.

11.1 Loomis-Whitney inequality

A simple example of this is an alternate proof of the Loomis-Whitney inequality: for every subset $A \subseteq \mathbf{Z}^d$, we have

$$|A|^{d-1} \leq \prod_{i=1}^{d} |P_i(A)|$$

where *Pi* is the orthogonal projection in the ith coordinate:

$$P_i(A) = \{(x_1, ..., x_{i-1}, x_{i+1}, ..., x_d) : (x_1, ..., x_d) \in A\}.$$

The proof follows as a simple corollary of Shearer's inequality: if X_1, ..., *Xd* are random variables and S_1, ..., *Sn* are subsets of {1, ..., *d*} such that every integer between *1* and *d* lies in exactly *r* of these subsets, then

$$H[(X_1, ..., X_d)] \leq \frac{1}{r} \sum_{i=1}^{n} H[(X_j)_{j \in S_i}]$$

where $(X_j)_{j \in S_i}$ is the Cartesian product of random variables *Xj* with indexes *j* in *Si* (so the dimension of this vector is equal to the size of *Si*).

We sketch how Loomis-Whitney follows from this: Indeed, let X be a uniformly distributed random variable with values in A and so that each point in A occurs with equal probability. Then (by the further properties of entropy mentioned above) $H(X) = \log|A|$, where $|A|$ denotes the cardinality of A. Let $Si = \{1, 2, ..., i-1, i+1, ..., d\}$. The range of $(X_j)_{j \in S_i}$ is contained in $Pi(A)$ and hence $H[(X_j)_{j \in S_i}] \leq \log |P_i(A)|$. Now use this to bound the right side of Shearer's inequality and exponentiate the opposite sides of the resulting inequality you obtain.

11.2 Approximation to binomial coefficient

For integers $0 < k < n$ let $q = k/n$. Then

$$\frac{2^{nH(q)}}{n+1} \leq \binom{n}{k} \leq 2^{nH(q)},$$

where

$$H(q) = -q \log_2(q) - (1-q) \log_2(1-q). \text{ [16]}$$

Here is a sketch proof. Note that $\binom{n}{k} q^{qn} (1-q)^{n-nq}$ is one term of the expression

$$\sum_{i=0}^{n} \binom{n}{i} q^i (1-q)^{n-i} = (q + (1-q))^n = 1.$$

Rearranging gives the upper bound. For the lower bound one first shows, using some algebra, that it is the largest term in the summation. But then,

$$\binom{n}{k} q^{qn} (1-q)^{n-nq} \geq \frac{1}{n+1}$$

since there are *n*+1 terms in the summation. Rearranging gives the lower bound.

A nice interpretation of this is that the number of binary strings of length n with exactly k many 1's is approximately $2^{nH(k/n)}$.[17]

12 See also

- Conditional entropy
- Cross entropy – is a measure of the average number of bits needed to identify an event from a set of possibilities between two probability distributions
- Entropy (arrow of time)
- Entropy encoding – a coding scheme that assigns codes to symbols so as to match code lengths with the probabilities of the symbols.
- Entropy estimation
- Entropy power inequality
- Entropy rate
- Fisher information
- Hamming distance
- History of entropy
- History of information theory
- Information geometry
- Joint entropy – is the measure how much entropy is contained in a joint system of two random variables.
- Kolmogorov-Sinai entropy in dynamical systems
- Levenshtein distance
- Mutual information
- Negentropy

- Perplexity
- Qualitative variation – other measures of statistical dispersion for nominal distributions
- Quantum relative entropy – a measure of distinguishability between two quantum states.
- Rényi entropy – a generalisation of Shannon entropy; it is one of a family of functionals for quantifying the diversity, uncertainty or randomness of a system.
- Shannon index
- Theil index
- Typoglycemia

13 References

[1] Shannon, Claude E. (July–October 1948). "A Mathematical Theory of Communication". *Bell System Technical Journal* **27** (3): 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x. (PDF)

[2] Goise, François & Olla, Stefano (2008). *Entropy methods for the Boltzmann equation: lectures from a special semester at the Centre Émile Borel, Institut H. Poincaré, Paris, 2001*. Springer. p. 14. ISBN 978-3-540-73704-9.

[3] Schneier, B: *Applied Cryptography*, Second edition, page 234. John Wiley and Sons.

[4] Shannon, C. E. (January 1951). "Prediction and Entropy of Printed English". *Bell System Technical Journal* **30** (1): 50–64. doi:10.1002/j.1538-7305.1951.tb01366.x. Retrieved 30 March 2014.

[5] Borda, Monica (2011). *Fundamentals in Information Theory and Coding*. Springer. p. 11. ISBN 978-3-642-20346-6.

[6] Han, Te Sun & Kobayashi, Kingo (2002). *Mathematics of Information and Coding*. American Mathematical Society. pp. 19–20. ISBN 978-0-8218-4256-0.

[7] Schneider, T.D, Information theory primer with an appendix on logarithms, National Cancer Institute, 14 April 2007.

[8] Carter, Tom (March 2014). *An introduction to information theory and entropy*. Santa Fe. Retrieved Aug 2014.

[9] Compare: Boltzmann, Ludwig (1896, 1898). Vorlesungen über Gastheorie : 2 Volumes – Leipzig 1895/98 UB: O 5262-6. English version: Lectures on gas theory. Translated by Stephen G. Brush (1964) Berkeley: University of California Press; (1995) New York: Dover ISBN 0-486-68455-5

[10] Mark Nelson (24 August 2006). "The Hutter Prize". Retrieved 2008-11-27.

[11] T. Schürmann and P. Grassberger, Entropy Estimation of Symbol Sequences, *CHAOS*,Vol. 6, No. 3 (1996) 414–427

[12] T. Schürmann, Bias Analysis in Entropy Estimation J. Phys. A: Math. Gen. 37 (2004) L295-L301.

[13] "The World's Technological Capacity to Store, Communicate, and Compute Information", Martin Hilbert and Priscila López (2011), Science (journal), 332(6025), 60–65; free access to the article through here: martinhilbert.net/WorldInfoCapacity.html

[14] Massey, James (1994). "Guessing and Entropy". "Proc. IEEE International Symposium on Information Theory". Retrieved December 31, 2013.

[15] Malone, David; Sullivan, Wayne (2005). "Guesswork is not a Substitute for Entropy". "Proceedings of the Information Technology & Telecommunications Conference". Retrieved December 31, 2013.

[16] Aoki, New Approaches to Macroeconomic Modeling. page 43.

[17] Probability and Computing, M. Mitzenmacher and E. Upfal, Cambridge University Press

14 Further reading

14.1 Textbooks on information theory

- Arndt, C. *Information Measures, Information and its Description in Science and Engineering* (Springer Series: Signals and Communication Technology), 2004, ISBN 978-3-540-40855-0
- Ash, RB. *Information Theory*. New York: Interscience, 1965. ISBN 0-470-03445-9. New York: Dover 1990. ISBN 0-486-66521-6
- Gallager, R. *Information Theory and Reliable Communication*. New York: John Wiley and Sons, 1968. ISBN 0-471-29048-3
- Goldman, S. *Information Theory*. New York: Prentice Hall, 1953. New York: Dover 1968 ISBN 0-486-62209-6, 2005 ISBN 0-486-44271-3
- Cover, TM, Thomas, JA. *Elements of information theory*, 1st Edition. New York: Wiley-Interscience, 1991. ISBN 0-471-06259-6.

  2nd Edition. New York: Wiley-Interscience, 2006. ISBN 0-471-24195-4.

- MacKay, DJC. *Information Theory, Inference, and Learning Algorithms* Cambridge: Cambridge University Press, 2003. ISBN 0-521-64298-1
- Martin, Nathaniel F.G. & England, James W. (2011). *Mathematical Theory of Entropy*. Cambridge University Press. ISBN 978-0-521-17738-2.
- Mansuripur, M. *Introduction to Information Theory*. New York: Prentice Hall, 1987. ISBN 0-13-484668-0
- Pierce, JR. "An introduction to information theory: symbols, signals and noise". Dover (2nd Edition). 1961 (reprinted by Dover 1980).
- Reza, F. *An Introduction to Information Theory*. New York: McGraw-Hill 1961. New York: Dover 1994. ISBN 0-486-68210-2
- Shannon, CE. Warren Weaver. *The Mathematical Theory of Communication.* Univ of Illinois Press, 1949. ISBN 0-252-72548-4
- Stone, JV. Chapter 1 of book "Information Theory: A Tutorial Introduction", University of Sheffield, England, 2014. ISBN 978-0956372857.

15 External links

- Hazewinkel, Michiel, ed. (2001), "Entropy", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Introduction to entropy and information on Principia Cybernetica Web
- *Entropy* an interdisciplinary journal on all aspect of the entropy concept. Open access.
- Information is not entropy, information is not uncertainty ! – a discussion of the use of the terms "information" and "entropy".
- I'm Confused: How Could Information Equal Entropy? – a similar discussion on the bionet.info-theory FAQ.
- Description of information entropy from "Tools for Thought" by Howard Rheingold
- A java applet representing Shannon's Experiment to Calculate the Entropy of English
- Slides on information gain and entropy
- *An Intuitive Guide to the Concept of Entropy Arising in Various Sectors of Science* – a wikibook on the interpretation of the concept of entropy.
- Calculator for Shannon entropy estimation and interpretation
- A Light Discussion and Derivation of Entropy
- Network Event Detection With Entropy Measures, Dr. Raimund Eimann, University of Auckland, PDF; 5993 kB – a PhD thesis demonstrating how entropy measures may be used in network anomaly detection.

16 Text and image sources, contributors, and licenses

16.1 Text

- **Entropy (information theory)** *Source:* http://en.wikipedia.org/wiki/Entropy_(information_theory)?oldid=630746842 *Contributors:* Tobias Hoevekamp, Derek Ross, Bryan Derksen, The Anome, Ap, PierreAbbat, Rade Kutil, Waveguy, B4hand, Youandme, Olivier, Stevertigo, Michael Hardy, Kku, Mkweise, Ahoerstemeier, Snoyes, AugPi, Rick.G, Ww, Sbwoodside, Dysprosia, Jitse Niesen, Fibonacci, Paul-L, Omegatron, Jeffq, Noeckel, Robbot, Tomchiukc, Benwing, Netpilot43556, Rursus, Bkell, Tobias Bergemann, Stirling Newberry, Giftlite, Boaz, Peruvianllama, Brona, Romanpoet, Jabowery, Christopherlin, Neilc, Gubbubu, Beland, OverlordQ, MarkSweep, Karol Langner, Wiml, Sctfn, Zeman, Abdull, TheObtuseAngleOfDoom, Rich Farmbrough, ArnoldReinhold, ESkog, MisterSheik, Jough, Guettarda, Cretog8, Army1987, Foobaz, Flammifer, Sligocki, PAR, Cburnett, Jheald, Tomash, Oleg Alexandrov, Linas, Shreevatsa, LOL, Bkwillwm, Male1979, Ryan Reich, Btyner, Marudubshinki, Graham87, BD2412, Jetekus, Grammarbot, Nanite, Sjö, Rjwilmsi, Thomas Arelatensis, Nneonneo, Erkcan, Alejo2083, Mfeadler, Srleffler, Chobot, Flashmorbid, Wavelength, Alpt, Kymacpherson, Ziddy, Kimchi.sg, Afelton, Buster79, Brandon, Hakeem.gadi, DmitriyV, GrinBot, SmackBot, InverseHypercube, Fulldecent, IstvanWolf, Diegotorquemada, Mcld, Ohnoitsjamie, Dauto, Kurykh, Gutworth, Nbarth, DHN-bot, Colonies Chris, Jdthood, Javalenok, CorbinSimpson, Robma, Radagast83, Cybercobra, Mrander, DMacks, FilippoSidoti, Daniel.Cardenas, Michael Rogers, Andrei Stroe, Ohconfucius, Snowgrouse, Dmh, Ninjagecko, JoseREMY, Severoon, Nonsuch, Phancy Physicist, Seanmadsen, Shockem, Ryan256, Dan Gluck, Kencf0618, Dwmalone, AlainD, Ylloh, CmdrObot, Hanspi, CBM, Mcstrother, Citrus538, Neonleonb, FilipeS, Tkircher, Farzaneh, Blaisorblade, Ignoramibus, Michael C Price, Alexnye, SteveMcCluskey, Nearfar, Thijs!bot, WikiC, Edchi, EdJohnston, D.H, Phy1729, Jvstone, Seaphoto, Heysan, Zylorian, Dougher, Husond, OhanaUnited, Time3000, Shaul1, Coffee2theorems, Magioladitis, RogierBrussee, VoABot II, Albmont, Swpb, First Harmonic, JaGa, Kestasjk, Tommy Herbert, Pax:Vobiscum, R'n'B, CommonsDelinker, Coppertwig, Policron, Jobonki, Jvpwiki, Ale2006, Idioma-bot, Cuzkatzimhut, Trevorgoodchild, Aelkiss, Trachten, Saigyo, Kjells, DragonLord, Mermanj, Spinningspark, PhysPhD, Bowsmand, Michel.machado, TimProof, Maxlittle2007, Hirstormandy, Neil Smithline, Dailyknowledge, Mdsam2, EnOreg, Algorithms, Svick, AlanUS, Melcombe, Rinconsoleao, Alksentrs, Schuermann, Vql, Djr32, Blueyeru, TedDunning, Musides, Ra2007, Qwfp, Johnuniq, Kace7, Porphyro, Addbot, Deepmath, Landon1980, Olli Niemitalo, Hans de Vries, Mv240, MrVanBot, Jill-Jênn, Favonian, ChenzwBot, Wikomidia, Numbo3-bot, Ehrenkater, Tide rolls, Lightbot, Fryed-peach, Eastereaster, Luckas-bot, Yobot, Sobec, Cassandra Cathcart, AnomieBOT, Jim1138, Zandr4, Mintrick, Informationtheory, Belkovich, ArthurBot, Xqbot, Gusshoekey, Br77rino, Almabot, GrouchoBot, Omnipaedista, RibotBOT, Ortvolute, Entropeter, Constructive editor, FrescoBot, Hobsonlane, GEBStgo, Mhadi.afrasiabi, Orubt, Rc3002, HRoestBot, Cesarth73, RedBot, Cfpcompte, Pmagrass, Mduteil, Lotje, BlackAce48, Angelorf, 777sms, CobraBot, Duoduoduo, Aoidh, Spakin, Jann.poppinga, Mitch.mcquoid, Fitoschido, Gopsy, Racerx11, Mo ainm, Hhhippo, Purplie, Quondum, SporkBot, Music Sorter, Erianna, Elsehow, ChuispastonBot, Sigma0 1, DASHBotAV, ClueBot NG, Tschijnmotschau, Mesoderm, Helpful Pixie Bot, Bibcode Bot, BG19bot, Guy vandegrift, Eli of Athens, Hushaohan, Trombonechamp, Manoguru, Muhammad Shuaib Nadwi, BattyBot, ChrisGualtieri, Marek marek, VLReeder77, Jrajniak89, Cerabot, Fourshade, Frosty, SFK2, Szzoli, Chrislgarry, I am One of Many, Jamesmcmahon0, Altroware, OhGodItsSoAmazing, Suderpie, Orehet, Monkbot, Visme, Donen1937, WikiRambala and Anonymous: 295

16.2 Images

- **File:Binary_entropy_plot.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/2/22/Binary_entropy_plot.svg *License:* CC-BY-SA-3.0 *Contributors:* original work by Brona, published on Commons at Image:Binary entropy plot.png. Converted to SVG by Alessio Damato *Original artist:* Brona and Alessio Damato
- **File:Crypto_key.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/6/65/Crypto_key.svg *License:* CC-BY-SA-3.0 *Contributors:* Own work based on image:Key-crypto-sideways.png by MisterMatt originally from English Wikipedia *Original artist:* MesserWoland
- **File:Entropy_flip_2_coins.jpg** *Source:* http://upload.wikimedia.org/wikipedia/commons/d/d4/Entropy_flip_2_coins.jpg *License:* CC-BY-SA-3.0 *Contributors:* File:Ephesos_620-600_BC.jpg *Original artist:* http://www.cngcoins.com/
- **File:Fisher_iris_versicolor_sepalwidth.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/4/40/Fisher_iris_versicolor_sepalwidth.svg *License:* CC-BY-SA-3.0 *Contributors:* en:Image:Fisher iris versicolor sepalwidth.png *Original artist:* en:User:Qwfp (original); Pbroks13 (talk) (redraw)
- **File:Question_book-new.svg** *Source:* http://upload.wikimedia.org/wikipedia/en/9/99/Question_book-new.svg *License:* ? *Contributors:* Created from scratch in Adobe Illustrator. Based on Image:Question book.png created by User:Equazcion *Original artist:* Tkgd2007
- **File:Symbol_template_class.svg** *Source:* http://upload.wikimedia.org/wikipedia/en/5/5c/Symbol_template_class.svg *License:* ? *Contributors:* ? *Original artist:* ?

16.3 Content license

- Creative Commons Attribution-Share Alike 3.0