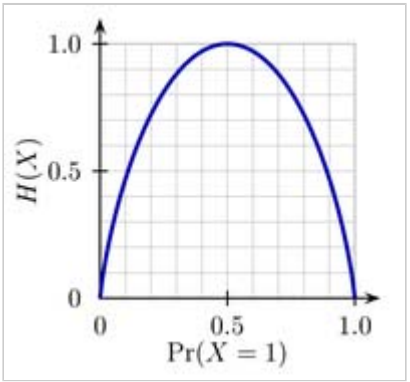
Information entropy

From Wikipedia, the free encyclopedia
(Redirected from Shannon entropy)

Entropy is a concept in thermodynamics (see entropy), statistical mechanics and information theory. Both concepts of entropy have deep links with one another, although it took many years for the development of the theories of statistical mechanics and information theory to make this connection apparent. This article is about **information entropy**, the information-theoretic formulation of entropy. Information entropy is occasionally called **Shannon's entropy** in honor of Claude E. Shannon, who formulated many of the key ideas of information theory.



Entropy of a Bernoulli trial as a function of success probability, often called the **binary entropy function**

Contents

Introduction

The concept of entropy in information theory describes how much information there is in a signal or event. Shannon introduced the idea of information entropy in his 1948 paper "A Mathematical Theory of Communication".

An intuitive understanding of information entropy relates to the amount of *uncertainty* about an event associated with a given probability distribution. As an example, consider a box containing many coloured balls. If the balls are all of different colours and no colour predominates, then our uncertainty about the colour of a randomly drawn ball is maximal. On the other hand, if the box contains more red balls than any other colour, then there is slightly less uncertainty about the result: the ball drawn from the box has more chances of being red (if we were forced to place a bet, we would bet on a red ball). Telling someone the colour of every new drawn ball provides them with more information in the first case than it does in the second case, because there is more uncertainty about what might happen in the first case than there is in the second. Intuitively, if we know the number of balls remaining, and they are all of one color, then there is no uncertainty about what the next ball drawn will be, and therefore there is no information content from drawing the ball. As a result, the entropy of the "signal" (the sequence of balls drawn, as calculated from the probability distribution) is higher in the first case than in the second.

Shannon, in fact, defined entropy as a measure of the average information content associated with a random outcome.

Shannon's definition of information entropy makes this intuitive distinction mathematically precise. His definition satisfies these desiderata:

- The measure should be continuous — i.e., changing the value of one of the probabilities by a very small amount should only change the entropy by a small amount.
- If all the outcomes (ball colours in the example above) are equally likely, then entropy should be maximal. In this case, the entropy increases with the number of outcomes.
- If the outcome is a certainty, then the entropy should be zero.
- The amount of entropy should be the same independently of how the process is regarded as being divided into parts.

(Note: The Shannon/Weaver book makes reference to Tolman (1938) who in turn credits Pauli (1933) with the definition of entropy Shannon used. Elsewhere in statistical mechanics, the literature includes references to von Neumann having derived the same form of entropy in 1927, which may explain why von Neumann favoured the use of the existing term 'entropy'.)

Formal definitions

Shannon defines entropy in terms of a discrete random variable X, with possible states (or outcomes) $x_1 \ldots x_n$ as:

$$H(X) = \sum_{i=1}^{n} p(x_i) \log_2 \left(\frac{1}{p(x_i)} \right) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i),$$

where

$p(x_i) = Pr(X = x_i)$ is the probability of the i^{th} outcome of X.

That is, the entropy of the variable X is the sum, over all possible outcomes x_i of X, of the product of the probability of outcome x_i times the log of the inverse of the probability of x_i (which is also called x_i's *surprisal* - the entropy of X is the expected value of its outcome's surprisal). We can also apply this to a general probability distribution, rather than a discrete-valued event.

Shannon shows that any definition of entropy satisfying his assumptions will be of the form:

$$-K \sum_{i=1}^{n} p(x_i) \log p(x_i).$$

where K is a constant (and is really just a choice of measurement units).

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. 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 measure came to be taken as a measure of the uncertainty about the realization of a random variable. It thus served as a proxy capturing the concept of information contained in a message as opposed to the portion of the message that is strictly determined (hence predictable) by inherent

structures. For example, redundancy in language structure or statistical properties relating to the occurrence frequencies of letter or word pairs, triplets etc. See Markov chain.

Relationship to thermodynamic entropy

Shannon's definition of entropy is closely related to thermodynamic entropy as defined in physics and chemistry. Boltzmann and Gibbs did considerable work on statistical thermodynamics, which became the inspiration for adopting the word *entropy* in information theory. There are relationships between thermodynamic and informational entropy. In fact, in the view of Jaynes (1957), thermodynamics should be seen as an *application* of Shannon's information theory: the thermodynamic entropy is interpreted as being an estimate of 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. For example, adding heat to a system increases its thermodynamic entropy because it increases the number of possible microscopic states that it could be in, thus making any complete state description longer. (See article: *maximum entropy thermodynamics*). Maxwell's demon (hypothetically) reduces the thermodynamic entropy of a system using information about the states of individual molecules; however, the demon himself increases his own entropy in the process, and so the total entropy does not decrease (which resolves the paradox).

Entropy as information content

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

The entropy rate of a data source means the average number of bits per symbol needed to encode it. Empirically, it seems that entropy of English text is between .6 and 1.3 bits per character, though clearly that will vary from one source of text to another. Shannon's experiments with human predictors show an information rate of between .6 and 1.3 bits per character (http://marknelson.us/2006/08/24/the-hutter-prize/#comment-293), depending on the experimental setup; the PPM compression algorithm can achieve a compression ratio of 1.5 bits per character.

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.

Data compression

Entropy effectively bounds the performance of the strongest lossless (or nearly 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.

Limitations of entropy as information content

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 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 $-\log_2 N$. 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.

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 p_i 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 (s).

For a second order Markov source, the entropy rate is

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

In general the ***b*-ary entropy** of a source $\mathcal{S} = (S,P)$ with source alphabet $S = \{a_1, \ldots, a_n\}$ and discrete probability distribution $P = \{p_1, \ldots, p_n\}$ where p_i is the probability of a_i (say $p_i = p(a_i)$) 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" which is being used as the standard yardstick to measure source alphabets. In information theory, two symbols are necessary and sufficient for an alphabet to be able 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 $\log_b n$.

Alternative definition

Another way to define the entropy function *H* (not using the Markov model) is by proving that *H* is uniquely defined (as earlier mentioned) if and only if *H* satisfies the following conditions:

1. $H(p_1, \ldots, p_n)$ is defined and continuous for all $p_1, \ldots, p_n$ where $p_i \in [0,1]$ for all $i = 1, \ldots, n$ and $p_1 + \ldots + p_n = 1$. (Remark that the function solely depends on the probability distribution, not the alphabet.)

2. For all positive integers *n*, *H* satisfies

$$H\underbrace{\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)}_{n \text{ arguments}} < H\underbrace{\left(\frac{1}{n+1}, \ldots, \frac{1}{n+1}\right)}_{n+1 \text{ arguments}}.$$

3. For positive integers b_i where $b_1 + \ldots + b_k = n$, *H* satisfies

$$H\underbrace{\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)}_{n} = H\underbrace{\left(\frac{b_1}{n}, \ldots, \frac{b_k}{n}\right)}_{k} + \sum_{i=1}^{k} \frac{b_i}{n} H\underbrace{\left(\frac{1}{b_i}, \ldots, \frac{1}{b_i}\right)}_{b_i}.$$

This last functional relationship characterizes the entropy of a system with sub-systems and is in a sense the most important of the three. It demands that the entropy of a system can be calculated from the entropy of its sub-systems if we know how the sub-systems interract with each other.

Assume that we have an ensemble of *n* elements with a uniform distribution on them. If we mentally divide this ensemble into *k* boxes (sub-systems) with b_i elements in each, the entropy can be calculated as a sum of individual entropies of the boxes weighed by the probability of finding oneself in that particular box PLUS the entropy of the system of boxes.

Efficiency

A source alphabet encountered in practice should be found to have a probability distribution which is less than optimal. If the source alphabet has *n* symbols, then it can be compared to an "optimized alphabet" with *n* symbols, whose probability distribution is uniform. The ratio of the entropy of the source alphabet with the entropy of its optimized version is the efficiency of the source alphabet, which can be expressed as a percentage.

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).

Derivation of Shannon's entropy

Since the entropy was given as a definition, it does not need to be derived. On the other hand, a "derivation" can be given which gives a sense of the motivation for the definition as well as the link to thermodynamic entropy.

Q. Given a roulette with *n* pockets which are all equally likely to be landed on by the ball, what is the probability of obtaining a distribution (A_1, A_2, …, A_n) where A_i is the number of times pocket *i* was landed on and

$$P = \sum_{i=1}^{n} A_i$$

is the total number of ball-landing events?

A. The probability is a multinomial distribution, viz.

$$p = \frac{\Omega}{T} = \frac{P!}{A_1!\ A_2!\ A_3!\ \cdots\ A_n!} \left(\frac{1}{n}\right)^P$$

where

$$\Omega = \frac{P!}{A_1!\ A_2!\ A_3!\ \cdots\ A_n!}$$

is the number of possible combinations of outcomes (for the events) which fit the given distribution, and

$$T = n^P$$

is the number of all possible combinations of outcomes for the set of *P* events.

Q. And what is the entropy?

A. The entropy of the distribution is obtained from the logarithm of Ω:

$$\begin{aligned} H = \log \Omega &= \log \frac{P!}{A_1!\ A_2!\ A_3! \cdots\ A_n!} \\ &= \log P! - \log A_1! - \log A_2! - \log A_3! - \cdots - \log A_n! \\ &= \sum_{i}^{P} \log i - \sum_{i}^{A_1} \log i - \sum_{i}^{A_2} \log i - \cdots - \sum_{i}^{A_n} \log i \end{aligned}$$

The summations can be approximated closely by being replaced with integrals:

$$H = \int_1^P \log x\, dx - \int_1^{A_1} \log x\, dx - \int_1^{A_2} \log x\, dx - \cdots - \int_1^{A_n} \log x\, dx.$$

The integral of the logarithm is

$$\int \log x\, dx = x \log x - \int x\, \frac{dx}{x} = x \log x - x.$$

So the entropy is

$$\begin{aligned} H &= (P\log P - P + 1) - (A_1\log A_1 - A_1 + 1) - (A_2\log A_2 - A_2 + 1) - \cdots - (A_n\log \\ &= (P\log P + 1) - (A_1\log A_1 + 1) - (A_2\log A_2 + 1) - \cdots - (A_n\log A_n \\ &= P\log P - \sum_{x=1}^{n} A_x \log A_x + (1-n) \end{aligned}$$

By letting $p_x = A_x/P$ and doing some simple algebra we obtain:

$$H = (1-n) - \sum_{x=1}^{n} p_x \log p_x$$

and the term $(1 - n)$ can be dropped since it is a constant, independent of the p_x distribution. The result is

$$H = -\sum_{x=1}^{n} p_x \log p_x \,.$$

Thus, the Shannon entropy is a consequence of the equation

$$H = \log \Omega$$

which relates to Boltzmann's definition,

$$\mathcal{S} = k \ln \Omega,$$

of thermodynamic entropy, where k is the Boltzmann constant.

Properties of Shannon's information entropy

We write $H(X)$ as $H_n(p_1,...,p_n)$. The Shannon entropy satisfies the following properties:

- For any n, $H_n(p_1,...,p_n)$ is a continuous and symmetric function on variables p_1, p_2,...,p_n.
- Event of probability zero does not contribute to the entropy, i.e. for any n,

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

- Entropy is maximized when the probability distribution is uniform. For all n,

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

Following from the Jensen inequality,

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

- If $p_{ij}, 1 \leq i \leq m, 1 \leq j \leq n$ are non-negative real numbers summing up to one, and $q_i = \sum_{j=1}^{n} p_{ij}$, then

$$H_{mn}(p_{11}, \ldots, p_{mn}) = H_m(q_1, \ldots, q_m) + \sum_{i=1}^{m} q_i H_n\Big(\frac{p_{i1}}{q_i}, \ldots, \frac{p_{in}}{q_i}\Big).$$

If we partition the *mn* outcomes of the random experiment into *m* groups with each group containing *n* elements, we can do the experiment in two steps: first, determine the group to which the actual outcome belongs; then, find the outcome in that group. The probability that you will observe group *i* is q_i. The conditional probability distribution function for group *i* is $p_{i1}/q_i,\ldots,p_{in}/q_i$). The entropy

$$H_n\Big(\frac{p_{i1}}{q_i}, \ldots, \frac{p_{in}}{q_i}\Big)$$

is the entropy of the probability distribution conditioned on group *i*. This property means that the total information is the sum of the information gained in the first step, $H_m(q_1,\ldots, q_n)$, and a weighted sum of the entropies conditioned on each group.

Khinchin in 1957 showed that the only function satisfying the above assumptions is of the form

$$H_n(p_1, \ldots, p_n) = -k \sum_{i=1}^{n} p_i \log p_i,$$

where *k* is a positive constant representing the desired unit of measurement.

Extending discrete entropy to the continuous case: differential entropy

The Shannon entropy is restricted to finite sets. It seems that the formula

$$h[f] = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx, \quad (*)$$

where *f* denotes a probability density function on the real line, is analogous to the Shannon entropy and could thus be viewed as an extension of the Shannon entropy to the domain of real numbers. Formula (*) is usually referred to as the **continuous entropy**, or differential entropy. Although the analogy between both functions is suggestive, the following question must be set: is the Boltzmann entropy a valid extension of the Shannon entropy? 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 as shown in the figure. As the figure indicates, 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

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

and expanding the logarithm, we have

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

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

As $\Delta \to 0$, we have

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

and so

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

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 $n \to \infty$

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.

More useful for the continuous case is the **relative entropy** of a distribution, defined as the Kullback-Leibler divergence from the distribution to a reference measure $m(x)$,

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

The relative entropy carries over directly from discrete to continuous distributions, and is invariant under

co-ordinate reparametrisations.

References

This article incorporates material from Shannon's entropy on PlanetMath, which is licensed under the GFDL.

See also

- **Binary entropy function** - the entropy of a Bernoulli trial with probability of success *p*
- **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
- **Joint entropy** - is the measure how much entropy is contained in a joint system of two random variables.
- **Entropy encoding** - a coding scheme that assigns codes to symbols so as to match code lengths with the probabilities of the symbols.
- **Kolmogorov-Sinai entropy** in dynamical systems
- **Rényi entropy** - a generalisation of information entropy; it is one of a family of functionals for quantifying the diversity, uncertainty or randomness of a system.
- **Perplexity**
- **Quantum relative entropy** - a measure of distinguishability between two quantum states.
- **Theil index**

External links

- Information is not entropy, information is not uncertainty ! (http://www.lecb.ncifcrf.gov/~toms/information.is.not.uncertainty.html) - a discussion of the use of the terms "information" and "entropy".
- I'm Confused: How Could Information Equal Entropy? (http://www.ccrnp.ncifcrf.gov/~toms/bionet.info-theory.faq.html#Information.Equal.Entropy) - a similar discussion on the bionet.info-theory FAQ.
- Description of information entropy from "Tools for Thought" by Howard Rheingold (http://www.rheingold.com/texts/tft/6.html)
- A java applet representing Shannon's Experiment to Calculate the Entropy of English (http://math.ucsd.edu/~crypto/java/ENTROPY/)

Retrieved from "http://en.wikipedia.org/wiki/Information_entropy"

Categories: PlanetMath sourced articles | Entropy | Information theory | Statistics | Randomness