# What is Entropy?

Jeff Gill, jgill@ucdavis.edu

## 1 Entropy in Information Theory

There are many definitions of information in various literatures but all of them have the same property of distinction from a message. If a message is the physical manifestation of information transmission (language, character set, salutations, headers/trailers, body, etc.), then the *information* is the substantive content independent of transmission processes. For example, in baseball the third base coach transmits a message to the runner on second base through the use of hand-gestures. The information contained in the hand gestures is independent of the actual gesture to the base-runner. If you could condense a set of messages down to a minimal finite length, then the information content would be the number of digits required by these alternate sets of messages. In this case information and message would be identical. Information and message are never exactly the same. For example, the DNA strand required to uniquely determine the biological characteristics of a human being contains far more codes than is minimally sufficient to transmit the required genetic information. In information theory entropy is a measure of the average amount of information required to describe the distribution of some random variable of interest (Cover and Thomas 1991). More generally, information can be thought of as a measure of reduction in uncertainty given a specific message (Shannon 1948, Ayres 1994, p.44).

Suppose we wanted to identify a particular voter by serial information on this person's characteristics. We are allowed to ask a consecutive set of yes/no questions (i.e. like the common guessing game). As we get answers to our series of questions we gradually converge (hopefully, depending on our skill) on the desired voter. Our first question is: does the voter reside in California. Since about 13% of voters in the United States reside in California, a yes answer gives us different information than a no answer. Restated, a yes answer reduces our uncertainty more than a no answer because a yes answer eliminates 87% of the choices whereas a no answer eliminates 13%. If $P_i$ is the probability of the $i^{\text{th}}$ event (residing in California), then the improvement in information is defined as:

$$I_{P_i} = \log_2\left[\frac{1}{P_i}\right] = -\log_2 P_i. \tag{1}$$

The probability is placed in the denominator of (1) because the smaller the probability, the greater the investigative information supplied by a yes answer. The log function is required to obtain some desired properties (discussed below), and is justified by limit theorems (Bevensee 1993, Jaynes

1982, Van Campenhout and Cover 1981). The log is base-2 since there are only two possible answers to our question (yes and no) making the units of information bits. In this example $H_i = -\log_2(0.13) = 2.943416$ bits, whereas if we had asked: does the voter live in Arkansas then an affirmative reply would have increased our information by: $H_i = -\log_2(0.02) = 5.643856$ bits, or about twice as much. However, there is a much smaller probability that we would have gotten an affirmative reply had the question been asked about Arkansas. What Slater (1938) found, and Shannon (1948) later refined was the idea that the "value" of the question was the information returned by a positive response times the probability of a positive response. So the value of the $i^{\text{th}}$ binary-response question is just:

$$H_i = f_i \log_2 \left[ \frac{1}{f_i} \right] = -f_i \log_2 f_i. \tag{2}$$

And the value of a series of $n$ of these questions is:

$$\sum_{i=1}^{n} H_i = k \sum_{i=1}^{n} f_i \log_2 \left[ \frac{1}{f_i} \right] = -k \sum_{i=1}^{n} f_i \log_2 f_i \tag{3}$$

where $f_i$ is the frequency distribution of the $i^{\text{th}}$ yes answer and $k$ is an arbitrary scaling factor that determines choice of units. The arbitrary scaling factor makes the choice of base in the logarithm unimportant since we can change this base by manipulating the constant.[1] The function (3) was actually first introduced by Slater (1938), although similar forms were referenced by Boltzmann (1877) where $f_i$ is the probability that a particle system is in microstate $i$.

We can see that the total improvement in information is the additive value of the series of individual information improvements. So in our simple example we might ask a series of questions narrowing down on the individual of interest. Is the voter in California? Is the voter registered as a Democrat? Does the voter reside in an urban area? Is the voter female?. The total information supplied by this vector of yes/no responses is the total information improvement in units of bits since the response-space is binary. Its important to remember that the information obtained is defined only with regard to a well-defined question having finite, enumerated responses

The link between information and entropy is often confusing due to differing terminology and usage in various literatures. In the thermodynamic sense, information and entropy are complementary: "gain in entropy means loss in information - nothing more" (Lewis 1930). So as uncertainty

[1] For instance if the entropy form were expressed in terms of the natural log, but $\log_2$ was more appropriate for the application (such as above), then setting $k = \frac{1}{\ln 2}$ converts the entropy form to base 2.

about the system increases, information about the configuration of the microstate of a system decreases. In terms of information theory, the distinction was less clear until Shannon (1948) presented the first unambiguous picture. To Shannon the function (3) represented the *expected* uncertainty. In a given communication, possible messages are indexed $m_1, m_2, \ldots, m_k$, and each have an associated probability $p_1, p_2, \ldots, p_k$. These probabilities sum to one, so the scaling factor is simply 1. Thus the Shannon entropy function is the negative expected value of the natural log of the probability mass function:

$$H = -k \sum p_i \ln(p_i). \tag{4}$$

Shannon defined the information in a message as the difference between the entropy before the message and the entropy after the message. If there is no information before the message, then a uniform prior distribution[2] is assigned to the $p_i$ and entropy is at its maximum[3] In this case any result increases our information. Yet, if there is certainty about the result, then a degenerate distribution describes the $m_i$, and the message does not change our information level.[4]. So according to Shannon, the message produces a new assessment about the state of the world and this in turn leads to the assignment of new probabilities ($p_i$) and a subsequent updating of the entropy value.

The simplicity of the Shannon entropy function belies its theoretical and practical importance. Shannon (1948, Appendix 2) showed that (4) is the only function that satisfies the following three desirable properties:

1. H is continuous in $(p_1, p_2, \ldots, p_n)$.

2. If the $p_i$ are uniformly distributed, then H is at its maximum and is monotonically increasing with n.

3. If a set of alternatives can be reformulated as multiple, consecutive sets of alternatives, then the first H should equal the weighted sum of the consecutive H values: $H(p_1, p_2, p_3) = H(p_1, 1 - p_1) + (1 - p_1)H(p_2, p_3)$.

These properties mean that the Shannon entropy function is well-behaved with regard to relative information comparisons. Up until this point only the discrete form of the entropy formulation

[2]The uniform prior distribution as applied provides the greatest entropy since no single event is more likely to occur. Thus the uniform distribution of events provides the minimum information possible with which to decode the message. This application of the uniform distribution does not imply that this is a "no information" assumption since equally likely outcomes is certainly a type of information. A great deal of controversy and discussion has focused around the erroneous treatment of the uniform distribution as a zero-based information source (Dale 1991, Stigler 1986).

[3]$H = -\sum \frac{1}{n} \ln(\frac{1}{n}) = \ln(n)$, so entropy increases logarithmically with the number of equally likely alternatives.

[4]$H = -\sum_{i=1}^{n-1}(0) - \log(1) = 0$

has been discussed. Since entropy is also a measure of uncertainty in probability distributions, its important to have a formulation that applies to continuous probability distribution functions. The following entropy formula also satisfies all of the properties enumerated above:

$$H = \int_{-\infty}^{+\infty} f(x)\ln(f(x))dx. \tag{5}$$

This form is often referred to as the *differential entropy* of a random variable $X$ with a known probability density function $f(x)$. There is one important distinction between (5) and (4): the discrete case measures the absolute uncertainty of a random variable given a set of probabilities, whereas the continuous case this uncertainty is measured relative to the chosen coordinate system (Ryu 1993). So transformations of the coordinate system require transformation in the probability distribution function with the associated jacobian. In addition, the continuous case can have infinite entropy unless additional, "side" conditions are imposed (see Jaynes 1968).

There has been considerable debate about the nature of information with regard to entropy (Ayres 1994, Jaynes 1957, Ruelle 1991, Tribus 1961, 1979). A major question arises as to *whose* entropy is it? The sender? The receiver? The transmission equipment? These are often philosophical debates rather than practical considerations as Shannon's focus was clearly on the receiver as the unit of entropy measure. Jaynes (1982) however considers the communication channel and its limitations as the determinants of entropy. Some, like Aczél and Daróczy (1975), view entropy as a descriptor of a stochastic event, and the traditional thermodynamic definition of entropy is as a measure of uncertainty in microstates. For the purposes of this work, the Shannon formulation (4) is used.

## 2 References


Aczél J. and Z. Daróczy. 1975. *On Measures of Information and Their Characterizations.* New York: Academic Press.

Ayres, David. 1994. *Information, Entropy, and Progress.* New York: American Institute of Physics Press.

Bevensee, Robert M. 1993. *Maximum Entropy Solutions to Scientific Problems.* Englewood Cliffs, NJ: Prentice Hall.

Boltzmann, L. 1877. "Uber die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Warmetheorie und der Wahrscheinlichkeitsrechnung, respective den Satzen uber das Wermegleichgewicht." *Wein Ber* 76: 373-95.

Cover, Thomas M., Joy A. Thomas. 1991. *Elements of Information Theory.* New York: John Wiley and Sons.

Dale, Andrew I. 1991. *A History of Inverse Probability: From Thomas Bayes to Karl Pearson.* New York: Spring-Verlag.

Jaynes, Edwin T. 1957. "Information Theory and Statistical Mechanics." *Physics Review* 106: 620-30.

Jaynes, Edwin T. 1968 "Prior Probabilities." *IEEE Transmissions on Systems Science and Cybernetics* SSC(4): 227-41.

Jaynes, Edwin T. 1982. "On the Rationale of Maximum-Entropy Methods." *Proceedings of the IEEE* 70(9): 939-52.

Lewis, Gilbert N. 1930. "The Symmetry of Time in Physics ." *Science* LXXI: 569-77.

Robert, Claudine. 1990. "An Entropy Concentration Theorem: Applications in Artificial Intelligence and Descriptive Statistics." *Journal of Applied Probability* 27: 303-13.

Ruelle, David. 1991. *Chance and Chaos.* Princeton: Princeton University Press.

Ryu, Hang K. 1993. "Maximum Entropy Estimation of Density and Regression Functions." *Journal of Econometrics* 56: 397-440.

Shannon, Claude. 1948. "A Mathematical Theory of Communication." *Bell System Technology Journal* 27: 379-423, 623-56.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Tribus, Myron. 1961. "Information Theory as the Basis for Thermostatics and Thermodynamics." *Journal of Applied Mechanics* 28: 1-8.

Tribus, Myron. 1979. "Thirty Years of Information Theory." In *The Maximum Entropy Formalism*, eds. Raphael D. Levine & Myron Tribus. Cambridge: MIT Press.

Van Campenhout, Jan M. and Thomas M Cover. 1981. "Maximum Entropy and Conditional Probability" *IEEE Transactions on Information Theory* 27(4): 483-9.