

ENCYCLOPEDIA OF COGNITIVE SCIENCE

2000

©Macmillan Reference Ltd

## Information Theory

information, entropy, communication, coding, bit, learning

Ghahramani, Zoubin

Zoubin Ghahramani

University College London [United Kingdom](#)

### [Definition]

Information is the reduction of uncertainty. Imagine your friend invites you to dinner for the first time. When you arrive at the building where he lives you find that you have misplaced his apartment number. He lives in a building with 4 floors and 8 apartments on each floor. If a neighbour passing by tells you that your friend lives on the top floor, your uncertainty about where he lives reduces from 32 choices to 8. By reducing your uncertainty, the neighbour has conveyed *information* to you. How can we quantify the amount of information?

Information theory is the branch of mathematics that describes how uncertainty should be quantified, manipulated and represented. Ever since the fundamental premises of information theory were laid down by Claude Shannon in 1949, it has had far reaching implications for almost every field of science and technology. Information theory has also had an important role in shaping theories of perception, cognition, and neural computation. In this article we will cover some of the basic concepts in information theory and how they relate to cognitive science and neuroscience.<sup>1</sup>

## Entropy and Mutual Information

The most fundamental quantity in information theory is *entropy* (Shannon and Weaver, 1949). Shannon borrowed the concept of entropy from thermodynamics where it describes the amount of disorder of a system. In information theory, entropy

---

<sup>1</sup> For more advanced textbooks on information theory see Cover and Thomas (1991) and MacKay (2001).

measures the amount of uncertainty of an unknown or random quantity. The entropy of a random variable  $X$  is defined to be:

$$H(X) = - \sum_{\text{all } x} p(x) \log_2 p(x) ,$$

where the sum is over all values  $x$  that the variable  $X$  can take, and  $p(x)$  is the probability of each of these values occurring. Entropy is measured in *bits* and can be generalised to continuous variables as well, although care must be taken to specify the precision level at which we would like to represent the continuous variable. Returning to our example, if  $X$  is the random variable which describes which apartment your friend lives in, initially it can take on 32 values with equal probability  $p(x)=1/32$ . Since  $\log_2(1/32) = -5$ , the entropy of  $X$  is 5 bits. After the neighbour tells you that he lives on the top floor, the probability of  $X$  drops to 0 for 24 of the 32 values and becomes 1/8 for the other 8 equally probable values. The entropy of  $X$  thus drops to 3 bits (using  $0 \log 0 = 0$ ). The neighbour has therefore conveyed 2 bits of information to you.

This fundamental definition of entropy as a measure of uncertainty can be derived from a small set of axioms. Entropy is the average amount of “surprise” associated with set of events. The amount of “surprise” of a particular event  $x$  is a function of the probability of that event - the less probable an event (e.g. a moose walking down Wall Street), the more surprising it is. The amount of surprise of two independent events (e.g. the Moose, and a solar eclipse) should be the sum of the amount of surprise of each event. These two constraints imply that the surprise of an event is proportional to  $-\log p(x)$ , with the proportionality constant determining what base logarithms are taken in (i.e. base 2 for bits). Averaging over all events according to their respective probabilities, we get the expression for  $H(X)$ .

Entropy in information theory has deep ties to the thermodynamic concept of entropy and, as we’ll see, it can be related to the least number of bits it would take on average to communicate  $X$  from a one location (the sender) to another (the receiver). On the one hand, the concepts of entropy and information are universal, in the sense that a bit of information can refer to the answer to *any* Yes/No question where the two options are equally probable. A megabyte is a megabyte (the answer to about a million Yes/No questions which can potentially distinguish between  $2^{1000000}$  possibilities!) regardless of whether it is used to encode a picture, music, or large quantities of text. On the other hand, entropy is always measured relative to a probability distribution,  $p(x)$ , and for many situations, it is not possible to consider the “true” probability of an event. For example, I may have high uncertainty about the weather tomorrow, but the meteorologist might not. This results in different entropies for the same set of events, defined relative to the subjective beliefs of the entity whose uncertainty we are measuring. This subjective or Bayesian view of probabilities is useful in considering how information communicated between different (biological or artificial) agents changes their beliefs.

While entropy is useful in determining the uncertainty in a single variable, it does not tell us how much uncertainty we have in one variable given knowledge of another. For this we need to define the *conditional entropy* of  $X$  given  $Y$ :

$$H(X | Y) = - \sum_{\text{all } x, y} p(x, y) \log_2 p(x | y)$$

where  $p(x | y)$  denotes the probability of  $x$  given that we have observed  $y$ . Building on this definition, the *mutual information* between two variables is the reduction in uncertainty in one variable given another variable. Mutual information can be written in three different ways:

$$I(X; Y) = H(X) - H(X | Y) = H(Y) - H(Y | X) = H(X) + H(Y) - H(X, Y)$$

where we see that the mutual information between two variables is symmetric:  $I(X; Y) = I(Y; X)$ . If the random variables  $X$  and  $Y$  are *independent*, that is, if the probability of their taking on values  $x$  and  $y$  is  $p(x, y) = p(x)p(y)$ , then they have zero mutual information. Similarly, if one can determine  $X$  exactly from  $Y$  and vice-versa, the mutual information is equal to the entropy of either of the two variables.

## Source Coding

Consider the problem of transmitting a sequence of symbols across a communication line using a binary representation. We assume that the symbols come from a finite alphabet (e.g. letters and punctuation marks of text, or levels from 0-255 of grey scale image patches) and that the communication line is noise-free. We further assume that the symbols are produced by a *source* which emits each symbol  $x$  randomly with some known probability  $p(x)$ . How many bits do we need to transmit per symbol so that the receiver can perfectly decode the sequence of symbols?

If there are  $N$  symbols in the alphabet then we could assign a distinct binary string (called *codeword*) of length  $L$  to each symbol as long as  $2^L > N$ , suggesting that we would need at most  $\log_2 N + 1$  bits. But we can do much better than this by assigning shorter codewords to more probable symbols and longer codewords to less probable ones. Shannon's **noiseless source coding theorem** states that if the source has entropy  $H(X)$  then there exists a decodable prefix code having an average length  $L$  per symbol such that

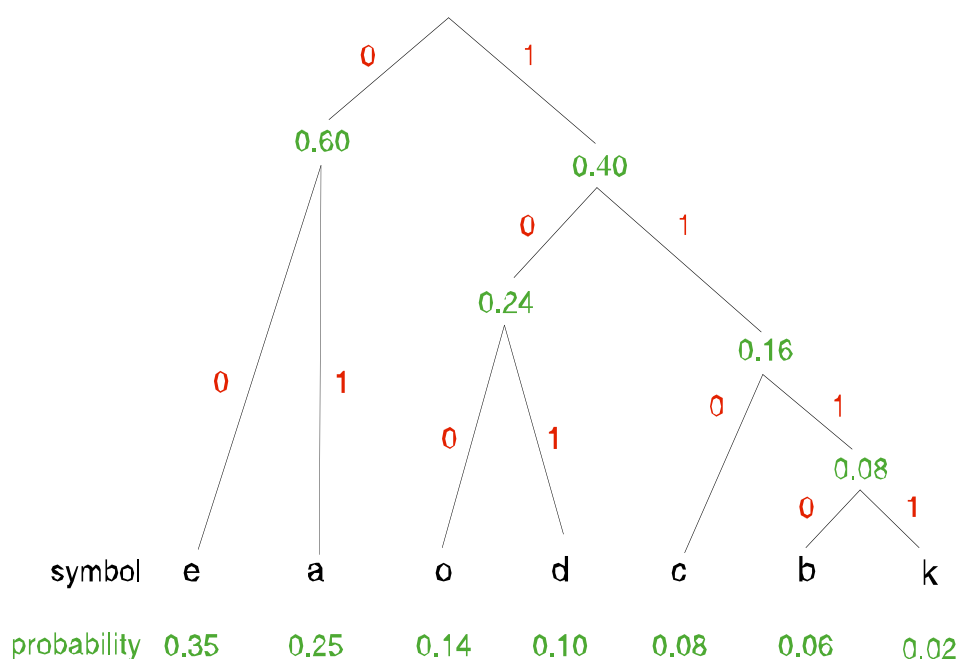
$$H(X) \leq L < H(X) + 1$$

Moreover, no uniquely decodable code exists having a smaller average length. In a *prefix* code, no codeword starts with another codeword, so the message can be decoded unambiguously as it comes in. This result places a lower bound on how many bits are required to compress a sequence of symbols losslessly. A closely related concept is the Kolmogorov complexity of a finite string, defined as the length in bits of the shortest program which when run on a universal computer will cause the string to be output.

## Huffman Codes

We can achieve the code length described by Shannon’s noiseless coding theorem using a very simple algorithm. The idea is to create a prefix code which uses shorter codewords for more frequent symbols and longer codewords for less frequent ones. First we combine the two least frequent symbols, summing their frequencies, into a new symbol. We do this repeatedly until we only have one symbol. The result is a variable depth tree with the original symbols at the leaves. We’ve illustrated this using an alphabet of 7 symbols  $\{a,b,c,d,e,o,k\}$  with differing probabilities (Figure 1). The codeword for each symbol is the sequences of left (0) and right (1) moves required to reach that symbol from the top of the tree.

Notice that in this example we have 7 symbols, so the naive fixed-length code would require 3 bits per symbol ( $2^3 = 8 \geq 7$ ). The Huffman code (which is variable-length) requires on average 2.48 bits; while the entropy gives a lower bound of 2.41 bits. The fact that it is a prefix code makes it easy to decode a string symbol by symbol by starting from the top of the tree and moving down left or right every time a new bit arrives. For example, try decoding: 1010011010010100.



**Figure 1: Huffman coding**

If we want to improve on this to get closer to the entropy bound, we can code blocks of several symbols at a time. Many practical coding schemes are based on forming blocks of symbols and coding each block separately. Using blocks also makes it possible to correct for errors introduced by noise in the communication channel.

## Information Transmission along a Noisy Channel

In the real world, communication channels suffer from noise. When transmitting data onto a mobile phone, listening to a person in a crowded room, or playing a DVD movie, there are random fluctuations in signal quality, background noise, or disk rotation speed, which we cannot control. A channel can be simply characterised by the conditional probability of the received symbols given the transmitted symbol:  $P(r | t)$ . This noise limits the information capacity of the channel, which is defined to be the maximum over all possible distributions over the transmitted symbols  $T$  of the mutual information between the transmitted and received symbol,  $R$  :

$$C = \max_{p(T)} I(T; R) .$$

For example, if the symbols are binary and the channel has no noise, then the channel capacity is 1 bit per symbol (corresponding to transmitting 0 and 1 with equal probability). However, if 10% of the time, a 0 transmitted is received as a 1, and 10% of the time a 1 transmitted is received as a 0, then the channel capacity is only 0.53 bits/symbol.

This probability of error couldn't be tolerated in most real applications. Does this mean that this channel is unusable? Not if one uses the trick of building redundancy into the transmitted signal in the form of an *error-correcting code* so that the receiver can then decode the intended message (Figure 2). One simple scheme is a *repetition code*. For example, encode the symbols by transmitting three repetitions of each; decode them by taking blocks of three and outputting the majority vote. This reduces the error probability from 10% to 2.7%, at the cost of reducing the rate at which the original symbols are transmitted to 1/3. If we want to achieve an error probability approaching zero, do we need to transmit at a rate approaching zero? The remarkable answer is No, proven by Shannon in the **channel coding theorem**. It states that all rates below channel capacity are achievable, i.e. that there are codes which transmit at that rate and have a maximum probability of error approaching zero. Conversely, if a code has probability of error approaching zero, it must have rate less than or channel capacity. Unfortunately Shannon's channel coding theorem does not say *how* to design codes that approach zero error probability near the channel capacity. Of course, codes with this property are more sophisticated than the repetition code, and finding good error correcting codes that can be decoded in reasonable time is an active area of research. Shannon's result is of immense practical significance since it shows us that we can have essentially perfect communication over a noisy channel.

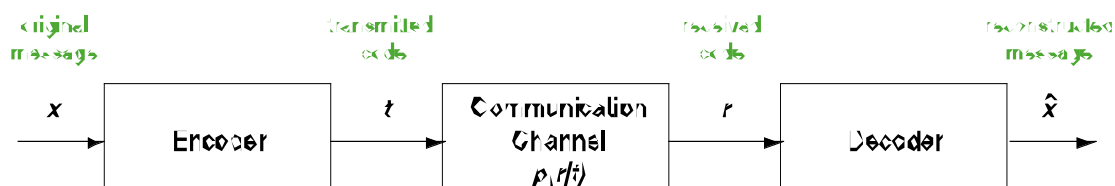


Figure 2: A noisy communication channel.

## Information Theory and Learning Systems

Information theory has played an important role in the study of learning systems. Just as information theory deals with quantifying information regardless of its physical medium of transmission, learning theory deals with understanding systems that learn irrespective of whether they are biological or artificial. Learning systems can be broadly categorised by the amount of information they receive from the environment in their supervision signal. In *unsupervised learning*, the goal of the system is to learn from sensory data with no supervision. This can be achieved by casting the unsupervised learning problem as one of discovering a code for the system's sensory data which is as efficient as possible. Thus the family of concepts – entropy, Kolmogorov complexity, and the general notion of description length – can be used to formalise unsupervised learning problems. We know from the source coding theorem that the most efficient code for a data source is one that uses  $-\log_2 p(x)$  bits per symbol  $x$ . Therefore, discovering the optimal coding scheme for a set of sensory data is equivalent to the problem of learning what the true probability distribution  $p(x)$  of the data is. If at some stage we have an estimate  $q(x)$  of this distribution we can use this estimate instead of the true probabilities to code the data. However, we incur a loss in efficiency measured by the *relative entropy* between the two probability distributions  $p$  and  $q$ :

$$D(p\|q) = \sum_{\text{all } x} p(x) \log_2 \frac{p(x)}{q(x)},$$

which is also known as the Kullback-Leibler divergence. This measure is the inefficiency in bits of coding messages with respect to a probability distribution  $q$  instead of the true probability distribution  $p$ , and is zero if and only if  $p = q$ . Many unsupervised learning systems can be designed from the principle of minimising this relative entropy.

## Information Theory in Cognitive Science and Neuroscience

The term “information processing system” has often been used to describe the brain. Indeed, information theory can be used to understand a variety of functions of the brain. We mention a few examples here.

- In neurophysiological experiments where a sensory stimulus is varied and the spiking activity of a neuron is recorded, mutual information can be used to infer what the neuron is coding for. Furthermore, the mutual information for different coding schemes can be compared, for example, to test whether the exact spike timing is used for information transmission (Rieke, et al. 1999).
- Information theory has been used to study both perceptual phenomena (Attneave, 1954) and the neural substrate of early visual processing (Barlow, 1961). It has been argued that the representations found in visual cortex arise from principles of redundancy reduction and optimal coding (Olshausen and Field, 1996).

- Communication via natural language occurs over a channel with limited capacity. Estimates of the entropy of natural language can be used to determine how much ambiguity/surprise there is in the next word following a stream of previous words giraffe and learning methods based on entropy can be used to model language (Berger et al., 1996).
- Redundant information arrives from multiple sensory sources (e.g. vision and audition) and over time (e.g. a series of frames of a movie). Decoding theory can be used to determine how this information should be combined optimally and whether the human system does so (Ghahramani, 1995).
- The human movement control system must cope with noise in motor neurons and in the muscles. Different ways of coding the motor command result in more or less variability in the movement (Harris and Wolpert, 1998).

Information theory lies at the core of our understanding of computing, communication, knowledge representation, and action. Like in many other fields of science, the basic concepts of information theory have played, and will continue to play, an important role in cognitive science and neuroscience.

#### References:

1. Attneave, F. (1954) Informational aspects of visual perception. *Psychological Review*, **61** 183-193.
2. Barlow, H.B. (1961). The coding of sensory messages. Chapter XIII. In *Current Problems in Animal Behaviour*, Thorpe and Zangwill (Eds), Cambridge University Press, pp. 330-360.
3. Berger, A. Della Pietra, S, and Della Pietra, V. A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1):39-71, 1996.
4. Cover, T. M. and Thomas, J.A. (1991) *Elements of Information Theory*. Wiley, New York.
5. Ghahramani, Z. (1995) *Computation and Psychophysics of Sensorimotor Integration*. Ph.D. Thesis, Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology.
6. Harris CM & Wolpert DM (1998) Signal-dependent noise determines motor planning. *Nature* **394**: 780-784
7. MacKay, D.J.C. (2001) *Information Theory, Inference and Learning Algorithms*. <http://wol.ra.phy.cam.ac.uk/mackay/itprnn/book.html>.
8. Olshausen, B.A. & Field, D.J. (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature*, **381**: 607-609.
9. Rieke, F. Warland, D. de Ruyter van Steveninck, R. and William, B. (1999) *Spikes: Exploring the Neural Code*. MIT Press.
10. Shannon, C.E and Weaver, W. W. (1949) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.