Information theory

From Wikipedia, the free encyclopedia

Information theory studies the quantification, storage, and communication of information. It was originally proposed by Claude E. Shannon in 1948 to find fundamental limits on signal processing and communication operations such as data compression, in a landmark paper entitled "A Mathematical Theory of Communication". Now this theory has found applications in many other areas, including statistical inference, natural language processing, cryptography, neurobiology,[1] the evolution[2] and function[3] of molecular codes, model selection in ecology,[4] thermal physics,[5] quantum computing, linguistics, plagiarism detection,[6] pattern recognition, and anomaly detection.[7]

A key measure in information theory is "entropy". Entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process. For example, identifying the outcome of a fair coin flip (with two equally likely outcomes) provides less information (lower entropy) than specifying the outcome from a roll of a die (with six equally likely outcomes). Some other important measures in information theory are mutual information, channel capacity, error exponents, and relative entropy.

Applications of fundamental topics of information theory include lossless data compression (e.g. ZIP files), lossy data compression (e.g. MP3s and JPEGs), and channel coding (e.g. for Digital Subscriber Line (DSL)).

The field is at the intersection of mathematics, statistics, computer science, physics, neurobiology, and electrical engineering. Its impact has been crucial to the success of the Voyager missions to deep space, the invention of the compact disc, the feasibility of mobile phones, the development of the Internet, the study of linguistics and of human perception, the understanding of black holes, and numerous other fields. Important sub-fields of information theory include source coding, channel coding, algorithmic complexity theory, algorithmic information theory, information-theoretic security, and measures of information.

Contents

- 1 Overview
- 2 Historical background
- 3 Quantities of information
  - 3.1 Entropy of an information source
  - 3.2 Joint entropy
  - 3.3 Conditional entropy (equivocation)
  - 3.4 Mutual information (transinformation)
  - 3.5 Kullback–Leibler divergence (information gain)
  - 3.6 Other quantities
- 4 Coding theory
  - 4.1 Source theory
    - 4.1.1 Rate
  - 4.2 Channel capacity
    - 4.2.1 Capacity of particular channel models
- 5 Applications to other fields
  - 5.1 Intelligence uses and secrecy applications
  - 5.2 Pseudorandom number generation
  - 5.3 Seismic exploration
  - 5.4 Semiotics
  - 5.5 Miscellaneous applications

- 6 See also
  - 6.1 Applications
  - 6.2 History
  - 6.3 Theory
  - 6.4 Concepts
- 7 References
  - 7.1 The classic work
  - 7.2 Other journal articles
  - 7.3 Textbooks on information theory
  - 7.4 Other books
  - 7.5 MOOC on information theory
- 8 External links

Overview

Information theory studies the transmission, processing, utilization, and extraction of information. Abstractly, information can be thought of as the resolution of uncertainty. In the case of communication of information over a noisy channel, this abstract concept was made concrete in 1948 by Claude Shannon in his paper "A Mathematical Theory of Communication", in which "information" is thought of as a set of possible messages, where the goal is to send these messages over a noisy channel, and then to have the receiver reconstruct the message with low probability of error, in spite of the channel noise. Shannon's main result, the noisy-channel coding theorem showed that, in the limit of many channel uses, the rate of information that is asymptotically achievable is equal to the channel capacity, a quantity dependent merely on the statistics of the channel over which the messages are sent.[1]

Information theory is closely associated with a collection of pure and applied disciplines that have been investigated and reduced to engineering practice under a variety of rubrics throughout the world over the past half century or more: adaptive systems, anticipatory systems, artificial intelligence, complex systems, complexity science, cybernetics, informatics, machine learning, along with systems sciences of many descriptions. Information theory is a broad and deep mathematical theory, with equally broad and deep applications, amongst which is the vital field of coding theory.

Coding theory is concerned with finding explicit methods, called *codes*, for increasing the efficiency and reducing the error rate of data communication over noisy channels to near the Channel capacity. These codes can be roughly subdivided into data compression (source coding) and error-correction (channel coding) techniques. In the latter case, it took many years to find the methods Shannon's work proved were possible. A third class of information theory codes are cryptographic algorithms (both codes and ciphers). Concepts, methods and results from coding theory and information theory are widely used in cryptography and cryptanalysis. *See the article ban (unit) for a historical application*.

Information theory is also used in information retrieval, intelligence gathering, gambling, statistics, and even in musical composition.

Historical background

The landmark event that **established** the discipline of information theory, and brought it to immediate worldwide attention, was the publication of Claude E. Shannon's classic paper "A Mathematical Theory of Communication" in the *Bell System Technical Journal* in July and October 1948.

Prior to this paper, limited information-theoretic ideas had been developed at Bell Labs, all implicitly assuming events of equal probability. Harry Nyquist's 1924 paper, *Certain Factors Affecting Telegraph Speed*, contains a theoretical section quantifying "intelligence" and the "line speed" at which it can be transmitted by a communication system, giving the relation $W = K \log m$ (recalling Boltzmann's constant), where W is the speed of transmission of intelligence, m is the number of different voltage levels to choose from at each time step, and K is a constant. Ralph Hartley's 1928 paper, *Transmission of Information*, uses the word *information* as a measurable quantity, reflecting the receiver's ability to distinguish one sequence of symbols from any other, thus quantifying information as $H = \log S^n = n \log S$, where S was the number of possible symbols, and n the number of symbols in a transmission. The unit of information was therefore the decimal digit, much later renamed the hartley in his honour as a unit or scale or measure of information. Alan Turing in 1940 used similar ideas as part of the statistical analysis of the breaking of the German second world war Enigma ciphers.

Much of the mathematics behind information theory with events of different probabilities were developed for the field of thermodynamics by Ludwig Boltzmann and J. Willard Gibbs. Connections between information-theoretic entropy and thermodynamic entropy, including the important contributions by Rolf Landauer in the 1960s, are explored in *Entropy in thermodynamics and information theory*.

In Shannon's revolutionary and groundbreaking paper, the work for which had been substantially completed at Bell Labs by the end of 1944, Shannon for the first time introduced the qualitative and quantitative model of communication as a statistical process underlying information theory, opening with the assertion that

> "The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point."

With it came the ideas of

- the information entropy and redundancy of a source, and its relevance through the source coding theorem;
- the mutual information, and the channel capacity of a noisy channel, including the promise of perfect loss-free communication given by the noisy-channel coding theorem;
- the practical result of the Shannon–Hartley law for the channel capacity of a Gaussian channel; as well as
- the bit—a new way of seeing the most fundamental unit of information.

Quantities of information

Information theory is based on probability theory and statistics. Information theory often concerns itself with measures of information of the distributions associated with random variables. Important quantities of information are entropy, a measure of information in a single random variable, and mutual information, a measure of information in common between two random variables. The former quantity is a property of the probability distribution of a random variable and gives a limit on the rate at which data generated by independent samples with the given distribution can be reliably compressed. The latter is a property of the joint distribution of two random variables, and is the maximum rate of reliable communication across a noisy channel in the limit of long block lengths, when the channel statistics are determined by the joint distribution.

The choice of logarithmic base in the following formulae determines the unit of information entropy that is used. A common unit of information is the bit, based on the binary logarithm. Other units include the nat, which is based on the natural logarithm, and the hartley, which is based on the common logarithm.

In what follows, an expression of the form $p \log p$ is considered by convention to be equal to zero whenever $p = 0$. This is justified because $\lim_{p \to 0+} p \log p = 0$ for any logarithmic base.

Entropy of an information source

Based on the probability mass function of each source symbol to be communicated, the Shannon entropy H, in units of bits (per symbol), is given by

$$H = -\sum_{i} p_i \log_2(p_i)$$

where p_i is the probability of occurrence of the i-th possible value of the source symbol. This equation gives the entropy in the units of "bits" (per symbol) because it uses a logarithm of base 2, and this base-2 measure of entropy has sometimes been called the "shannon" in his honor. Entropy is also commonly computed using the natural logarithm (base e, where e is Euler's number), which produces a measurement of entropy in "nats" per symbol and sometimes simplifies the analysis by avoiding the need to include extra constants in the formulas. Other bases are also possible, but less commonly used. For example, a logarithm of base $2^8 = 256$ will produce a measurement in bytes per symbol, and a logarithm of base 10 will produce a measurement in decimal digits (or hartleys) per symbol.

Intuitively, the entropy H_X of a discrete random variable X is a measure of the amount of *uncertainty* associated with the value of X when only its distribution is known.

The entropy of a source that emits a sequence of N symbols that are independent and identically distributed (iid) is $N \cdot H$ bits (per message of N symbols). If the source data symbols are identically distributed but not independent, the entropy of a message of length N will be less than $N \cdot H$.

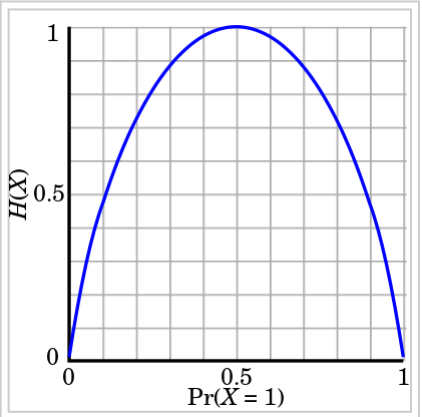
Suppose one transmits 1000 bits (0s and 1s). If the value of each of these bits is known to the receiver (has a specific value with certainty) ahead of transmission, it is clear that no information is transmitted. If, however, each bit is independently equally likely to be 0 or 1, 1000 shannons of information (more often called bits) have been transmitted. Between these two extremes, information can be quantified as follows. If $\mathbb{X}$ is the set of all messages $\{x_1, \ldots, x_n\}$ that X could be, and $p(x)$ is the probability of some $x \in \mathbb{X}$, then the entropy, H, of X is defined:[8]

$$H(X) = \mathbb{E}_X[I(x)] = -\sum_{x \in \mathbb{X}} p(x) \log p(x).$$

(Here, $I(x)$ is the self-information, which is the entropy contribution of an individual message, and $\mathbb{E}_X$ is the expected value.) A property of entropy is that it is maximized when all the messages in the message space are equiprobable $p(x) = 1/n$; i.e., most unpredictable, in which case $H(X) = \log n$.

The special case of information entropy for a random variable with two outcomes is the **binary entropy function**, usually taken to the logarithmic base 2, thus having the shannon (Sh) as unit:

$$H_{\mathrm{b}}(p) = -p \log_2 p - (1-p) \log_2 (1-p).$$



The entropy of a Bernoulli trial as a function of success probability, often called the **binary entropy function**, $H_{\mathrm{b}}(p)$. The entropy is maximized at 1 bit per trial when the two possible outcomes are equally probable, as in an unbiased coin toss.

Joint entropy

The **joint entropy** of two discrete random variables X and Y is merely the entropy of their pairing: (X, Y). This implies that if X and Y are independent, then their joint entropy is the sum of their individual entropies.

For example, if (X, Y) represents the position of a chess piece — X the row and Y the column, then the joint entropy of the row of the piece and the column of the piece will be the entropy of the position of the piece.

$$H(X, Y) = \mathbb{E}_{X,Y}[-\log p(x, y)] = -\sum_{x,y} p(x, y) \log p(x, y)$$

Despite similar notation, joint entropy should not be confused with **cross entropy**.

Conditional entropy (equivocation)

The **conditional entropy** or **conditional uncertainty** of X given random variable Y (also called the **equivocation** of X about Y) is the average conditional entropy over Y:[9]

$$H(X|Y) = \mathbb{E}_Y[H(X|y)] = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y) = -\sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(y)}.$$

Because entropy can be conditioned on a random variable or on that random variable being a certain value, care should be taken not to confuse these two definitions of conditional entropy, the former of which is in more common use. A basic property of this form of conditional entropy is that:

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

Mutual information (transinformation)

Mutual information measures the amount of information that can be obtained about one random variable by observing another. It is important in communication where it can be used to maximize the amount of information shared between sent and received signals. The mutual information of X relative to Y is given by:

$$I(X; Y) = \mathbb{E}_{X,Y}[SI(x, y)] = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)}$$

where SI (*S*pecific mutual *I*nformation) is the pointwise mutual information.

A basic property of the mutual information is that

$$I(X; Y) = H(X) - H(X|Y).$$

That is, knowing *Y*, we can save an average of $I(X; Y)$ bits in encoding *X* compared to not knowing *Y*.

Mutual information is symmetric:

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

Mutual information can be expressed as the average Kullback–Leibler divergence (information gain) between the posterior probability distribution of *X* given the value of *Y* and the prior distribution on *X*:

$$I(X; Y) = \mathbb{E}_{p(y)}[D_{\mathrm{KL}}(p(X|Y = y) \| p(X))].$$

In other words, this is a measure of how much, on the average, the probability distribution on *X* will change if we are given the value of *Y*. This is often recalculated as the divergence from the product of the marginal distributions to the actual joint distribution:

$$I(X;Y) = D_{\mathrm{KL}}(p(X,Y) \| p(X)p(Y)).$$

Mutual information is closely related to the log-likelihood ratio test in the context of contingency tables and the multinomial distribution and to Pearson's χ^2 test: mutual information can be considered a statistic for assessing independence between a pair of variables, and has a well-specified asymptotic distribution.

Kullback–Leibler divergence (information gain)

The **Kullback–Leibler divergence** (or **information divergence**, **information gain**, or **relative entropy**) is a way of comparing two distributions: a "true" probability distribution *p(X)*, and an arbitrary probability distribution *q(X)*. If we compress data in a manner that assumes *q(X)* is the distribution underlying some data, when, in reality, *p(X)* is the correct distribution, the Kullback–Leibler divergence is the number of average additional bits per datum necessary for compression. It is thus defined

$$D_{\mathrm{KL}}(p(X) \| q(X)) = \sum_{x \in X} -p(x) \log q(x) \ - \sum_{x \in X} -p(x) \log p(x) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}.$$

Although it is sometimes used as a 'distance metric', KL divergence is not a true metric since it is not symmetric and does not satisfy the triangle inequality (making it a semi-quasimetric).

Another interpretation of the KL divergence is the "unnecessary surprise" introduced by a prior from the truth: suppose a number *X* is about to be drawn randomly from a discrete set with probability distribution *p(x)*. If Alice knows the true distribution *p(x)*, while Bob believes (has a prior) that the distribution is *q(x)*, then Bob will be more surprised than Alice, on average, upon seeing the value of *X*. The KL divergence is the (objective) expected value of Bob's (subjective) surprisal minus Alice's surprisal, measured in bits if the *log* is in base 2. In this way, the extent to which Bob's prior is "wrong" can be quantified in terms of how "unnecessarily surprised" it's expected to make him.

Other quantities

Other important information theoretic quantities include Rényi entropy (a generalization of entropy), differential entropy (a generalization of quantities of information to continuous distributions), and the conditional mutual information.

Coding theory

Coding theory is one of the most important and direct applications of information theory. It can be subdivided into source coding theory and channel coding theory. Using a statistical description for data, information theory quantifies the number of bits needed to describe the data, which is the information entropy of the source.

- Data compression (source coding): There are two formulations for the compression problem:

1. lossless data compression: the data must be reconstructed exactly;
2. lossy data compression: allocates bits needed to reconstruct the data, within a specified fidelity level measured by a distortion function. This subset of Information theory is called rate–distortion theory.

A picture showing scratches on the readable surface of a CD-R. Music and data CDs are coded using error correcting codes and thus can still be read even if they have minor scratches using error detection and correction.

- Error-correcting codes (channel coding): While data compression removes as much redundancy as possible, an error correcting code adds just the right kind of redundancy (i.e., error correction) needed to transmit the data efficiently and faithfully across a noisy channel.

This division of coding theory into compression and transmission is justified by the information transmission theorems, or source–channel separation theorems that justify the use of bits as the universal currency for information in many contexts. However, these theorems only hold in the situation where one transmitting user wishes to communicate to one receiving user. In scenarios with more than one transmitter (the multiple-access channel), more than one receiver (the broadcast channel) or intermediary "helpers" (the relay channel), or more general networks, compression followed by transmission may no longer be optimal. Network information theory refers to these multi-agent communication models.

Source theory

Any process that generates successive messages can be considered a **source** of information. A memoryless source is one in which each message is an independent identically distributed random variable, whereas the properties of ergodicity and stationarity impose less restrictive constraints. All such sources are stochastic. These terms are well studied in their own right outside information theory.

Rate

Information **rate** is the average entropy per symbol. For memoryless sources, this is merely the entropy of each symbol, while, in the case of a stationary stochastic process, it is

$$r = \lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2}, X_{n-3}, \ldots);$$

that is, the conditional entropy of a symbol given all the previous symbols generated. For the more general case of a process that is not necessarily stationary, the *average rate* is

$$r = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots X_n);$$

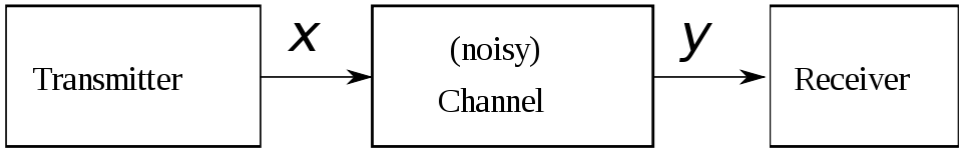
that is, the limit of the joint entropy per symbol. For stationary sources, these two expressions give the same result.[10]

It is common in information theory to speak of the "rate" or "entropy" of a language. This is appropriate, for example, when the source of information is English prose. The rate of a source of information is related to its redundancy and how well it can be compressed, the subject of **source coding**.

Channel capacity

Communications over a channel—such as an ethernet cable—is the primary motivation of information theory. As anyone who's ever used a telephone (mobile or landline) knows, however, such channels often fail to produce exact reconstruction of a signal; noise, periods of silence, and other forms of signal corruption often degrade quality. How much information can one hope to communicate over a noisy (or otherwise imperfect) channel?

Consider the communications process over a discrete channel. A simple model of the process is shown below:



Here X represents the space of messages transmitted, and Y the space of messages received during a unit time over our channel. Let $p(y|x)$ be the conditional probability distribution function of Y given X. We will consider $p(y|x)$ to be an inherent fixed property of our communications channel (representing the nature of the **noise** of our channel). Then the joint distribution of X and Y is completely determined by our channel and by our choice of $f(x)$, the marginal distribution of messages we choose to send over the channel. Under these constraints, we would like to maximize the rate of information, or the **signal**, we can communicate over the channel. The appropriate measure for this is the mutual information, and this maximum mutual information is called the **channel capacity** and is given by:

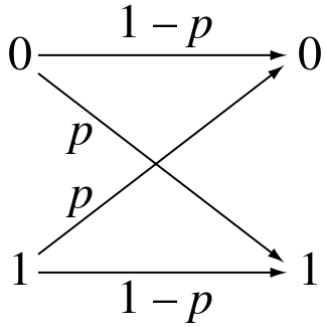
$$C = \max_f I(X;Y).$$

This capacity has the following property related to communicating at information rate R (where R is usually bits per symbol). For any information rate $R < C$ and coding error $\varepsilon > 0$, for large enough N, there exists a code of length N and rate $\geq$ R and a decoding algorithm, such that the maximal probability of block error is $\leq \varepsilon$; that is, it is always possible to transmit with arbitrarily small block error. In addition, for any rate $R > C$, it is impossible to transmit with arbitrarily small block error.

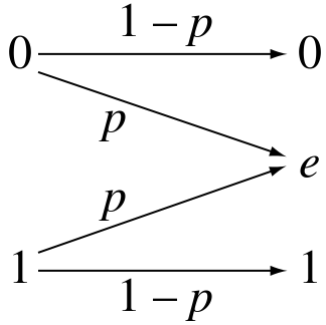
Channel coding is concerned with finding such nearly optimal codes that can be used to transmit data over a noisy channel with a small coding error at a rate near the channel capacity.

Capacity of particular channel models

- A continuous-time analog communications channel subject to Gaussian noise — see Shannon–Hartley theorem.
- A binary symmetric channel (BSC) with crossover probability p is a binary input, binary output channel that flips the input bit with probability p. The BSC has a capacity of $1 - H_\mathrm{b}(p)$ bits per channel use, where H_b is the binary entropy function to the base 2 logarithm:



- A binary erasure channel (BEC) with erasure probability p is a binary input, ternary output channel. The possible channel outputs are *0*, *1*, and a third symbol 'e' called an erasure. The erasure represents complete loss of information about an input bit. The capacity of the BEC is *1 - p* bits per channel use.



Applications to other fields

Intelligence uses and secrecy applications

Information theoretic concepts apply to cryptography and cryptanalysis. Turing's information unit, the ban, was used in the Ultra project, breaking the German Enigma machine code and hastening the end of World War II in Europe. Shannon himself defined an important concept now called the unicity distance. Based on the redundancy of the plaintext, it attempts to give a minimum amount of ciphertext necessary to ensure unique decipherability.

Information theory leads us to believe it is much more difficult to keep secrets than it might first appear. A brute force attack can break systems based on asymmetric key algorithms or on most commonly used methods of symmetric key algorithms (sometimes called secret key algorithms), such as block ciphers. The security of all such methods currently comes from the assumption that no known attack can break them in a practical amount of time.

Information theoretic security refers to methods such as the one-time pad that are not vulnerable to such brute force attacks. In such cases, the positive conditional mutual information between the plaintext and ciphertext (conditioned on the key) can ensure proper transmission, while the unconditional mutual information between the plaintext and ciphertext remains zero, resulting in absolutely secure communications. In other words, an eavesdropper would not be able to improve his or her guess of the plaintext by gaining knowledge of the ciphertext but not of the key. However, as in any other cryptographic system, care must be used to correctly apply even information-theoretically secure methods; the Venona project was able to crack the one-time pads of the Soviet Union due to their improper reuse of key material.

Pseudorandom number generation

Pseudorandom number generators are widely available in computer language libraries and application programs. They are, almost universally, unsuited to cryptographic use as they do not evade the deterministic nature of modern computer equipment and software. A class of improved random number generators is termed cryptographically secure pseudorandom number generators, but even they require random seeds external to the software to work as intended. These can be obtained via extractors, if done carefully. The measure of sufficient randomness in extractors is min-entropy, a value related to Shannon entropy through Rényi entropy; Rényi entropy is also used in evaluating randomness in cryptographic systems. Although related, the distinctions among these measures mean that a random variable with high Shannon entropy is not necessarily satisfactory for use in an extractor and so for cryptography uses.

Seismic exploration

One early commercial application of information theory was in the field of seismic oil exploration. Work in this field made it possible to strip off and separate the unwanted noise from the desired seismic signal. Information theory and digital signal processing offer a major improvement of resolution and image clarity over previous

analog methods.[11]

Semiotics

Concepts from information theory such as redundancy and code control have been used by semioticians such as Umberto Eco and Ferruccio Rossi-Landi to explain ideology as a form of message transmission whereby a dominant social class emits its message by using signs that exhibit a high degree of redundancy such that only one message is decoded among a selection of competing ones.[12]

Miscellaneous applications

Information theory also has applications in gambling and investing, black holes, and bioinformatics.

See also

- Algorithmic probability
- Algorithmic information theory
- Bayesian inference
- Communication theory
- Constructor theory - a generalization of information theory that includes quantum information
- Inductive probability
- Minimum message length
- Minimum description length
- List of important publications
- Philosophy of information

Applications

- Active networking
- Cryptanalysis
- Cryptography
- Cybernetics
- Entropy in thermodynamics and information theory
- Gambling
- Intelligence (information gathering)
- Seismic exploration

History

- Hartley, R.V.L.
- History of information theory
- Shannon, C.E.
- Timeline of information theory
- Yockey, H.P.

Theory

- Coding theory
- Detection theory
- Estimation theory
- Fisher information
- Information algebra
- Information asymmetry
- Information field theory
- Information geometry
- Information theory and measure theory
- Kolmogorov complexity
- Logic of information
- Network coding
- Philosophy of Information
- Quantum information science
- Semiotic information theory
- Source coding

- Unsolved Problems

Concepts

- Ban (unit)
- Channel capacity
- Channel (communications)
- Communication source
- Conditional entropy
- Covert channel
- Decoder
- Differential entropy
- Encoder
- Information entropy
- Joint entropy
- Kullback–Leibler divergence
- Mutual information
- Pointwise mutual information (PMI)
- Receiver (information theory)
- Redundancy
- Rényi entropy
- Self-information
- Unicity distance
- Variety

References

1. F. Rieke; D. Warland; R Ruyter van Steveninck; W Bialek (1997). *Spikes: Exploring the Neural Code*. The MIT press. ISBN 978-0262681087.
2. cf. Huelsenbeck, J. P., F. Ronquist, R. Nielsen and J. P. Bollback (2001) Bayesian inference of phylogeny and its impact on evolutionary biology, *Science* **294**:2310-2314
3. Rando Allikmets, Wyeth W. Wasserman, Amy Hutchinson, Philip Smallwood, Jeremy Nathans, Peter K. Rogan, Thomas D. Schneider (http://alum.mit.edu/www/toms/), Michael Dean (1998) Organization of the ABCR gene: analysis of promoter and splice junction sequences, *Gene* **215**:1, 111-122
4. Burnham, K. P. and Anderson D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition* (Springer Science, New York) ISBN 978-0-387-95364-9.
5. Jaynes, E. T. (1957) Information Theory and Statistical Mechanics (http://bayes.wustl.edu/), *Phys. Rev.* **106**:620
6. Charles H. Bennett, Ming Li, and Bin Ma (2003) Chain Letters and Evolutionary Histories (http://sciamdigital.com/index.cfm?fa=Products.ViewIssuePreview&ARTICLEID_CHAR=08B64096-0772-4904-9D48227D5C9FAC75), *Scientific American* **288**:6, 76-81
7. David R. Anderson (November 1, 2003). "Some background on why people in the empirical sciences may want to better understand the information-theoretic methods" (pdf). Retrieved 2010-06-23.
8. Fazlollah M. Reza (1994) [1961]. *An Introduction to Information Theory*. Dover Publications, Inc., New York. ISBN 0-486-68210-2.
9. Robert B. Ash (1990) [1965]. *Information Theory*. Dover Publications, Inc. ISBN 0-486-66521-6.
10. Jerry D. Gibson (1998). *Digital Compression for Multimedia: Principles and Standards*. Morgan Kaufmann. ISBN 1-55860-369-7.
11. The Corporation and Innovation, Haggerty, Patrick, Strategic Management Journal, Vol. 2, 97-118 (1981)
12. Semiotics of Ideology, Noth, Winfried, Semiotica, Issue 148,(1981)

The classic work

- Shannon, C.E. (1948), "A Mathematical Theory of Communication", *Bell System Technical Journal*, 27, pp. 379–423 & 623–656, July & October, 1948. PDF. (http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6773024)
  Notes and other formats. (http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html)
- R.V.L. Hartley, "Transmission of Information" (http://www.dotrose.com/etext/90_Miscellaneous/transmission_of_information_1928b.pdf), *Bell System Technical Journal*, July 1928
- Andrey Kolmogorov (1968), "Three approaches to the quantitative definition of information" in International Journal of Computer Mathematics.

Other journal articles

- J. L. Kelly, Jr., Saratoga.ny.us (http://www.racing.saratoga.ny.us/kelly.pdf), "A New Interpretation of Information Rate" *Bell System Technical Journal*, Vol. 35, July 1956, pp. 917–26.
- R. Landauer, IEEE.org (http://ieeexplore.ieee.org/search/wrapper.jsp?arnumber=615478), "Information is Physical" *Proc. Workshop on Physics and Computation PhysComp'92* (IEEE Comp. Sci.Press, Los Alamitos, 1993) pp. 1–4.
- R. Landauer, IBM.com (http://www.research.ibm.com/journal/rd/441/landauerii.pdf), "Irreversibility and Heat Generation in the Computing Process" *IBM J. Res. Develop.* Vol. 5, No. 3, 1961
- Timme, Nicholas; Alford, Wesley; Flecker, Benjamin; Beggs, John M. (2012). "Multivariate information measures: an experimentalist's perspective". arXiv:1111.6857 [cs.IT].

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, Thomas; Thomas, Joy A. (2006). *Elements of information theory* (2nd ed.). New York: Wiley-Interscience. ISBN 0-471-24195-4.
- Csiszar, I, Korner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems* Akademiai Kiado: 2nd edition, 1997. ISBN 963-05-7440-3
- MacKay, David J. C.. *Information Theory, Inference, and Learning Algorithms (http://www.inference.phy.cam.ac.uk/mackay/itila/book.html)* Cambridge: Cambridge University Press, 2003. ISBN 0-521-64298-1
- Mansuripur, M. *Introduction to Information Theory*. New York: Prentice Hall, 1987. ISBN 0-13-484668-0
- McEliece, R. *The Theory of Information and Coding*". *Cambridge, 2002. ISBN 978-0521831857*
- 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, Claude; Weaver, Warren (1949). *The Mathematical Theory of Communication* (PDF). Urbana, Illinois: University of Illinois Press. ISBN 0-252-72548-4. LCCN 49-11922.
- Stone, JV. Chapter 1 of book "Information Theory: A Tutorial Introduction" (http://jim-stone.staff.shef.ac.uk/BookInfoTheory/InfoTheoryBookMain.html), University of Sheffield, England, 2014. ISBN 978-0956372857.
- Yeung, RW. *A First Course in Information Theory (http://iest2.ie.cuhk.edu.hk/~whyeung/book/)* Kluwer Academic/Plenum Publishers, 2002. ISBN 0-306-46791-7.
- Yeung, RW. *Information Theory and Network Coding (http://iest2.ie.cuhk.edu.hk/~whyeung/book2/)* Springer 2008, 2002. ISBN 978-0-387-79233-0

Other books

- Leon Brillouin, *Science and Information Theory*, Mineola, N.Y.: Dover, [1956, 1962] 2004. ISBN 0-486-43918-6
- James Gleick, *The Information: A History, a Theory, a Flood*, New York: Pantheon, 2011. ISBN 978-0-375-42372-7
- A. I. Khinchin, *Mathematical Foundations of Information Theory*, New York: Dover, 1957. ISBN 0-486-60434-9
- H. S. Leff and A. F. Rex, Editors, *Maxwell's Demon: Entropy, Information, Computing*, Princeton University Press, Princeton, New Jersey (1990). ISBN 0-691-08727-X
- Robert K. Logan. *What is Information? - Propagating Organization in the Biosphere, the Symbolosphere, the Technosphere and the Econosphere*,

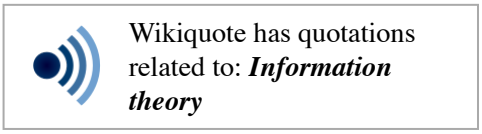
Toronto: DEMO Publishing.

- Tom Siegfried, *The Bit and the Pendulum*, Wiley, 2000. ISBN 0-471-32174-5
- Charles Seife, *Decoding the Universe*, Viking, 2006. ISBN 0-670-03441-X
- Jeremy Campbell, *Grammatical Man*, Touchstone/Simon & Schuster, 1982, ISBN 0-671-44062-4
- Henri Theil, *Economics and Information Theory*, Rand McNally & Company - Chicago, 1967.

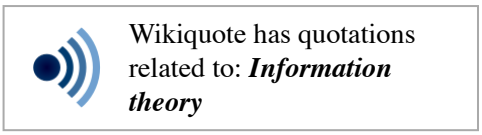
- Escolano, Suau, Bonev, *Information Theory in Computer Vision and Pattern Recognition (http://www.springer.com/computer/image+processing/book/978-1-84882-296-2)*, Springer, 2009. ISBN 978-1-84882-296-2

MOOC on information theory

- Raymond W. Yeung, "Information Theory (http://www.inc.cuhk.edu.hk/InformationTheory/index.html)" (The Chinese University of Hong Kong)



External links



- Erill I. (2012), "A gentle introduction to information content in transcription factor binding sites (http://erilllab.umbc.edu/files/2016/04/Introduction_Information_Theory.pdf)" (University of Maryland, Baltimore County)
- Hazewinkel, Michiel, ed. (2001), "Information", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Lambert F. L. (1999), "Shuffled Cards, Messy Desks, and Disorderly Dorm Rooms - Examples of Entropy Increase? Nonsense! (http://jchemed.chem.wisc.edu/Journal/Issues/1999/Oct/abs1385.html)", *Journal of Chemical Education*
- Schneider T. D. (2014), "Information Theory Primer (http://alum.mit.edu/www/toms/paper/primer)"
- Srinivasa, S., "A Review on Multivariate Mutual Information (http://www.nd.edu/~jnl/ee80653/tutorials/sunil.pdf)"
- IEEE Information Theory Society (http://www.itsoc.org/index.html) and ITSoc review articles (http://www.itsoc.org/review.html)

Categories: Information theory | Cybernetics | Formal sciences | Information Age

---