

# Kullback–Leibler divergence

In mathematical statistics, the **Kullback–Leibler divergence**, *D*<sub>KL</sub> (also called **relative entropy**), is a measure of how one probability distribution is different from a second, reference probability distribution.<sup>[1][2]</sup> Applications include characterizing the relative (Shannon) entropy in information systems, randomness in continuous time-series, and information gain when comparing statistical models of inference. In contrast to variation of information, it is a distribution-wise *asymmetric* measure and thus does not qualify as a statistical *metric* of spread – it also does not satisfy the triangle inequality. In the simple case, a relative entropy of 0 indicates that the two distributions in question have identical quantities of information. In simplified terms, it is a measure of surprise, with diverse applications such as applied statistics, fluid mechanics, neuroscience and bioinformatics.

## Contents

### Introduction and context

### Etymology

### Definition

### Basic example

### Interpretations

### Motivation

### Properties

### Examples

Multivariate normal distributions

### Relation to metrics

Fisher information metric

Fisher information metric theorem

Variation of information

### Relation to other quantities of information theory

Self-information

Mutual information

Shannon entropy

Conditional entropy

Cross entropy

### Bayesian updating

Bayesian experimental design

### Discrimination information

Principle of minimum discrimination information

### Relationship to available work

### Quantum information theory

### Relationship between models and reality

### Symmetrised divergence

### Relationship to other probability-distance measures

### Data differencing

### See also

### References

# Introduction and context

Consider two probability distributions *P* and *Q*. Usually, *P* represents the data, the observations, or a probability distribution precisely measured. Distribution *Q* represents instead a theory, a model, a description or an approximation of *P*. The Kullback–Leibler divergence is then interpreted as the average difference of the number of bits required for encoding samples of *P* using a code optimized for *Q* rather than one optimized for *P*.

# Etymology

The relative entropy was introduced by [Solomon Kullback](#) and [Richard Leibler](#) in 1951 as the **directed divergence** between two distributions; Kullback preferred the term **discrimination information**.<sup>[3]</sup> The divergence is discussed in Kullback's 1959 book, *Information Theory and Statistics*.<sup>[2]</sup>

# Definition

For [discrete probability distributions](#) *P* and *Q* defined on the same [probability space](#), ***X***, the relative entropy from *Q* to *P* is defined<sup>[4]</sup> to be

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right).$$

which is equivalent to

$$D_{\text{KL}}(P \parallel Q) = - \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{Q(x)}{P(x)} \right)$$

In other words, it is the [expectation](#) of the logarithmic difference between the probabilities *P* and *Q*, where the expectation is taken using the probabilities *P*. Relative entropy is defined only if for all ***x***, *Q(x) = 0* implies *P(x) = 0* ([absolute continuity](#)). Whenever *P(x)* is zero the contribution of the corresponding term is interpreted as zero because

$$\lim_{x \rightarrow 0^+} x \log(x) = 0.$$

For distributions *P* and *Q* of a [continuous random variable](#), relative entropy is defined to be the integral:<sup>[5]:p. 55</sup>

$$D_{\text{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \left( \frac{p(x)}{q(x)} \right) dx$$

where ***p*** and ***q*** denote the [probability densities](#) of *P* and *Q*.

More generally, if *P* and *Q* are [probability measures](#) over a set ***X***, and *P* is [absolutely continuous](#) with respect to *Q*, then the relative entropy from *Q* to *P* is defined as

$$D_{\text{KL}}(P \parallel Q) = \int_{\mathcal{X}} \log \left( \frac{dP}{dQ} \right) dP,$$

where 



d
P

d
Q




{\displaystyle {\frac {dP}{dQ}}}

 is the [Radon–Nikodym derivative](#) of *P* with respect to *Q*, and provided the expression on the right-hand side exists. Equivalently (by the [chain rule](#)), this can be written as

$$D_{\text{KL}}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) \frac{dP}{dQ} dQ,$$

which is the entropy of  $Q$  relative to  $P$ . Continuing in this case, if  $\mu$  is any measure on  $\mathcal{X}$  for which  $p = \frac{dP}{d\mu}$  and  $q = \frac{dQ}{d\mu}$  exist (meaning that  $p$  and  $q$  are absolutely continuous with respect to  $\mu$ ), then the relative entropy from  $Q$  to  $P$  is given as

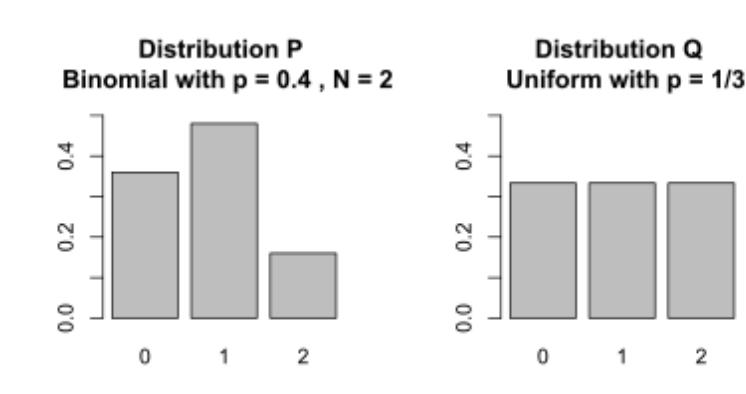
$$D_{\text{KL}}(P \parallel Q) = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right) d\mu.$$

The logarithms in these formulae are taken to base 2 if information is measured in units of bits, or to base  $e$  if information is measured in nats. Most formulas involving relative entropy hold regardless of the base of the logarithm.

Various conventions exist for referring to  $D_{\text{KL}}(P \parallel Q)$  in words. Often it is referred to as the divergence *between*  $P$  and  $Q$ , but this fails to convey the fundamental asymmetry in the relation. Sometimes, as in this article, it may be described as the divergence of  $P$  *from*  $Q$  or as the divergence *from*  $Q$  *to*  $P$ . This reflects the asymmetry in Bayesian inference, which starts *from* a prior  $Q$  and updates *to* the posterior  $P$ . Another common way to refer to  $D_{\text{KL}}(P \parallel Q)$  is as the relative entropy of  $P$  *with respect to*  $Q$ .

## Basic example

Kullback<sup>[2]</sup> gives the following example (Table 2.1, Example 2.1). Let  $P$  and  $Q$  be the distributions shown in the table and figure.  $P$  is the distribution on the left side of the figure, a binomial distribution with  $N = 2$  and  $p = 0.4$ .  $Q$  is the distribution on the right side of the figure, a discrete uniform distribution with the three possible outcomes  $x = 0, 1$ , or  $2$  (i.e.  $\mathcal{X} = \{0, 1, 2\}$ ), each with probability  $p = 1/3$ .



<b>x</b>	<b>0</b>	<b>1</b>	<b>2</b>
Distribution $P(x)$	9/25	12/25	4/25
Distribution $Q(x)$	1/3	1/3	1/3

Relative entropies  $D_{\text{KL}}(P \parallel Q)$  and  $D_{\text{KL}}(Q \parallel P)$  are calculated as follows. This example uses the natural log with base  $e$ , designated **ln** to get results in nats (see units of information).

$$\begin{aligned}
D_{\text{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \ln \left( \frac{P(x)}{Q(x)} \right) \\
&= \frac{9}{25} \ln \left( \frac{9/25}{1/3} \right) + \frac{12}{25} \ln \left( \frac{12/25}{1/3} \right) + \frac{4}{25} \ln \left( \frac{4/25}{1/3} \right) \\
&= \frac{1}{25} (32 \ln(2) + 55 \ln(3) - 50 \ln(5)) \approx 0.0852996
\end{aligned}$$

$$\begin{aligned}
D_{\text{KL}}(Q \parallel P) &= \sum_{x \in \mathcal{X}} Q(x) \ln \left( \frac{Q(x)}{P(x)} \right) \\
&= \frac{1}{3} \ln \left( \frac{1/3}{9/25} \right) + \frac{1}{3} \ln \left( \frac{1/3}{12/25} \right) + \frac{1}{3} \ln \left( \frac{1/3}{4/25} \right) \\
&= \frac{1}{3} (-4 \ln(2) - 6 \ln(3) + 6 \ln(5)) \approx 0.097455
\end{aligned}$$

## Interpretations

The relative entropy from  $Q$  to  $P$  is often denoted  $D_{\text{KL}}(P \parallel Q)$ .

In the context of machine learning,  $D_{\text{KL}}(P \parallel Q)$  is often called the information gain achieved if  $P$  would be used instead of  $Q$  which is currently used. By analogy with information theory, it is called the **relative entropy** of  $P$  with respect to  $Q$ . In the context of coding theory,  $D_{\text{KL}}(P \parallel Q)$  can be constructed by measuring the expected number of extra bits required to code samples from  $P$  using a code optimized for  $Q$  rather than the code optimized for  $P$ .

Expressed in the language of Bayesian inference,  $D_{\text{KL}}(P \parallel Q)$  is a measure of the information gained by revising one's beliefs from the prior probability distribution  $Q$  to the posterior probability distribution  $P$ . In other words, it is the amount of information lost when  $Q$  is used to approximate  $P$ .<sup>[6]</sup> In applications,  $P$  typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while  $Q$  typically represents a theory, model, description, or approximation of  $P$ . In order to find a distribution  $Q$  that is closest to  $P$ , we can minimize KL divergence and compute an information projection.

Relative entropy is a special case of a broader class of statistical divergences called  $f$ -divergences as well as the class of Bregman divergences. It is the only such divergence over probabilities that is a member of both classes. Although it is often intuited as a way of measuring the distance between probability distributions, the Kullback–Leibler divergence is not a true metric. It does not obey the Triangle Inequality, and in general  $D_{\text{KL}}(P \parallel Q)$  does not equal  $D_{\text{KL}}(Q \parallel P)$ . However, its infinitesimal form, specifically its Hessian, gives a metric tensor known as the Fisher information metric.

Arthur Hobson proved that relative entropy is the only measure of difference between probability distributions that satisfies some desired properties, which are the canonical extension to those appearing in a commonly used characterization of entropy.<sup>[7]</sup> Consequently, mutual information is the only measure of mutual dependence that obeys certain related conditions, since it can be defined in terms of Kullback–Leibler divergence.

## Motivation

In information theory, the Kraft–McMillan theorem establishes that any directly decodable coding scheme for coding a message to identify one value  $\mathbf{x}_i$  out of a set of possibilities  $\mathbf{X}$  can be seen as representing an implicit probability distribution  $q(\mathbf{x}_i) = 2^{-\ell_i}$  over  $\mathbf{X}$ , where  $\ell_i$  is the length of the code for  $\mathbf{x}_i$  in bits. Therefore, relative entropy can be interpreted as the expected extra message-length per datum that must be communicated if a code that is optimal for a given (wrong) distribution  $Q$  is used, compared to using a code based on the true distribution  $P$ .

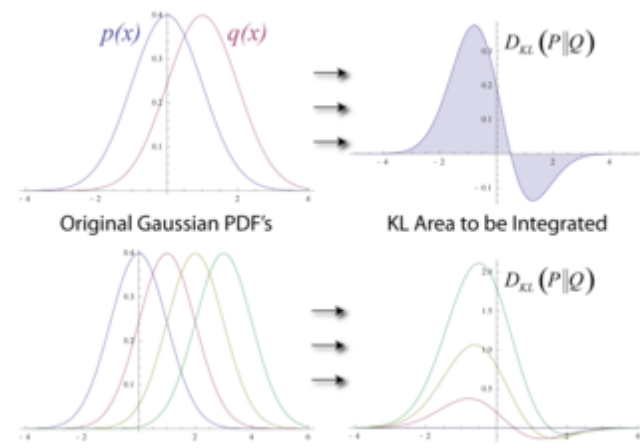


Illustration of the relative entropy for two normal distributions. The typical asymmetry is clearly visible.

$$\begin{aligned} D_{\text{KL}}(P \parallel Q) &= - \sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= H(P, Q) - H(P) \end{aligned}$$

where  $H(P, Q)$  is the cross entropy of  $P$  and  $Q$ , and  $H(P)$  is the entropy of  $P$  (which is the same as the cross-entropy of  $P$  with itself).

Relative entropy  $D_{\text{KL}}(P \parallel Q)$  can be thought of as something like a measurement of how far the distribution  $Q$  is from the distribution  $P$ . The cross-entropy  $H(P, Q)$  is itself such a measurement, but it has the defect that  $H(P, P) =: H(P)$  isn't zero, so we subtract  $H(P)$  to make  $D_{\text{KL}}(P \parallel Q)$  agree more closely with our notion of distance. (Unfortunately it still isn't symmetric.) Relative entropy relates to "rate function" in the theory of large deviations.<sup>[8][9]</sup>

## Properties

- Relative entropy is always non-negative,

$$D_{\text{KL}}(P \parallel Q) \geq 0,$$

a result known as Gibbs' inequality, with  $D_{\text{KL}}(P \parallel Q)$  equals zero if and only if  $P = Q$  almost everywhere. The entropy  $H(P)$  thus sets a minimum value for the cross-entropy  $H(P, Q)$ , the expected number of bits required when using a code based on  $Q$  rather than  $P$ ; and the Kullback–Leibler divergence therefore represents the expected number of extra bits that must be transmitted to identify a value  $x$  drawn from  $X$ , if a code is used corresponding to the probability distribution  $Q$ , rather than the "true" distribution  $P$ .

- Relative entropy remains well-defined for continuous distributions, and furthermore is invariant under parameter transformations. For example, if a transformation is made from variable  $x$  to variable  $y(x)$ , then, since  $P(x)dx = P(y)dy$  and  $Q(x)dx = Q(y)dy$  the relative entropy may be rewritten:

$$\begin{aligned} D_{\text{KL}}(P \parallel Q) &= \int_{x_a}^{x_b} P(x) \log \left( \frac{P(x)}{Q(x)} \right) dx \\ &= \int_{y_a}^{y_b} P(y) \log \left( \frac{P(y) \frac{dy}{dx}}{Q(y) \frac{dy}{dx}} \right) dy = \int_{y_a}^{y_b} P(y) \log \left( \frac{P(y)}{Q(y)} \right) dy \end{aligned}$$

where  $y_a = y(x_a)$  and  $y_b = y(x_b)$ . Although it was assumed that the transformation was continuous, this need not be the case. This also shows that the relative entropy produces a dimensionally consistent quantity, since if  $x$  is a dimensioned variable,  $P(x)$  and  $Q(x)$  are also dimensioned, since e.g.  $P(x)dx$  is

dimensionless. The argument of the logarithmic term is and remains dimensionless, as it must. It can therefore be seen as in some ways a more fundamental quantity than some other properties in information theory<sup>[10]</sup> (such as self-information or Shannon entropy), which can become undefined or negative for non-discrete probabilities.

- Relative entropy is additive for independent distributions in much the same way as Shannon entropy. If  $P_1, P_2$  are independent distributions, with the joint distribution  $P(x, y) = P_1(x)P_2(y)$ , and  $Q, Q_1, Q_2$  likewise, then

$$D_{\text{KL}}(P \parallel Q) = D_{\text{KL}}(P_1 \parallel Q_1) + D_{\text{KL}}(P_2 \parallel Q_2).$$

- Relative entropy  $D_{\text{KL}}(P \parallel Q)$  is convex in the pair of probability mass functions  $(p, q)$ , i.e. if  $(p_1, q_1)$  and  $(p_2, q_2)$  are two pairs of probability mass functions, then

$$D_{\text{KL}}(\lambda p_1 + (1 - \lambda)p_2 \parallel \lambda q_1 + (1 - \lambda)q_2) \leq \lambda D_{\text{KL}}(p_1 \parallel q_1) + (1 - \lambda)D_{\text{KL}}(p_2 \parallel q_2) \text{ for } 0 \leq \lambda \leq 1.$$

- The following theorem is referred to as a duality formula for variational inference. It explains some important properties of the variational distributions used in variational Bayesian methods<sup>[11]</sup>.

**Theorem** Consider two probability spaces  $(\Theta, \mathcal{F}, P)$  and  $(\Theta, \mathcal{F}, Q)$  with  $Q \ll P$ . Assume that there is a common dominating probability measure  $\lambda$  such that  $P \ll \lambda$  and  $Q \ll \lambda$ . Let  $h$  denote any real-valued random variable on  $(\Theta, \mathcal{F}, P)$  that satisfies  $h \in L_1(P)$ . Then the following equality holds

$$\log E_P[\exp h] = \sup_{Q \ll P} \{E_Q[h] - D_{\text{KL}}(Q \parallel P)\}.$$

Further, the supremum on the right-hand side is attained if and only if it holds

$$\frac{q(\theta)}{p(\theta)} = \frac{\exp h(\theta)}{E_P[\exp h]},$$

almost surely with respect to probability measure  $Q$ , where  $p(\theta) = dP/d\lambda$  and  $q(\theta) = dQ/d\lambda$  denote the Radon-Nikodym derivatives of the probability measures  $P$  and  $Q$  with respect to  $\lambda$ , respectively.

## Examples

### Multivariate normal distributions

Suppose that we have two multivariate normal distributions, with means  $\mu_0, \mu_1$  and with (non-singular) covariance matrices  $\Sigma_0, \Sigma_1$ . If the two distributions have the same dimension,  $k$ , then the relative entropy between the distributions is as follows:<sup>[12]:p. 13</sup>

$$D_{\text{KL}}(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2} \left( \text{tr}(\Sigma_1^{-1} \Sigma_0) + (\mu_1 - \mu_0)^\top \Sigma_1^{-1} (\mu_1 - \mu_0) - k + \ln \left( \frac{\det \Sigma_1}{\det \Sigma_0} \right) \right).$$

The logarithm in the last term must be taken to base  $e$  since all terms apart from the last are base- $e$  logarithms of expressions that are either factors of the density function or otherwise arise naturally. The equation therefore gives a result measured in nats. Dividing the entire expression above by  $\ln(2)$  yields the divergence in bits.

A special case, and a common quantity in variational inference, is the relative entropy between a diagonal multivariate normal, and a standard normal distribution (with zero mean and unit variance):

$$D_{\text{KL}} \left( \mathcal{N} \left( (\mu_1, \dots, \mu_k)^\top, \text{diag}(\sigma_1^2, \dots, \sigma_k^2) \right) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}) \right) = \frac{1}{2} \sum_{i=1}^k (\sigma_i^2 + \mu_i^2 - 1 - \ln(\sigma_i^2)).$$

# Relation to metrics

One might be tempted to call relative entropy a "distance metric" on the space of probability distributions, but this would not be correct as it is not symmetric – that is,  $D_{\text{KL}}(P \parallel Q) \neq D_{\text{KL}}(Q \parallel P)$  – nor does it satisfy the triangle inequality. It generates a topology on the space of probability distributions. More concretely, if  $\{P_1, P_2, \dots\}$  is a sequence of distributions such that

$$\lim_{n \rightarrow \infty} D_{\text{KL}}(P_n \parallel Q) = 0$$

then it is said that

$$P_n \xrightarrow{D} Q.$$

Pinsker's inequality entails that

$$P_n \xrightarrow{D} P \Rightarrow P_n \xrightarrow{TV} P,$$

where the latter stands for the usual convergence in total variation.

## Fisher information metric

Relative entropy is directly related to the Fisher information metric. This can be made explicit as follows. Assume that the probability distributions  $P$  and  $Q$  are both parameterized by some (possibly multi-dimensional) parameter  $\theta$ . Consider then two close by values of  $P = P(\theta)$  and  $Q = P(\theta_0)$  so that the parameter  $\theta$  differs by only a small amount from the parameter value  $\theta_0$ . Specifically, up to first order one has (using the Einstein summation convention)

$$P(\theta) = P(\theta_0) + \Delta\theta_j P_j(\theta_0) + \dots$$

with  $\Delta\theta_j = (\theta - \theta_0)_j$  a small change of  $\theta$  in the  $j$  direction, and  $P_j(\theta_0) = \frac{\partial P}{\partial \theta_j}(\theta_0)$  the corresponding rate of change in the probability distribution. Since relative entropy has an absolute minimum 0 for  $P = Q$ , i.e.  $\theta = \theta_0$ , it changes only to *second* order in the small parameters  $\Delta\theta_j$ . More formally, as for any minimum, the first derivatives of the divergence vanish

$$\left. \frac{\partial}{\partial \theta_j} D_{\text{KL}}(P(\theta) \parallel P(\theta_0)) \right|_{\theta=\theta_0} = 0,$$

and by the Taylor expansion one has up to second order

$$D_{\text{KL}}(P(\theta) \parallel P(\theta_0)) = \frac{1}{2} \Delta\theta_j \Delta\theta_k g_{jk}(\theta_0) + \dots$$

where the Hessian matrix of the divergence

$$g_{jk}(\theta_0) = \left. \frac{\partial^2}{\partial \theta_j \partial \theta_k} D_{\text{KL}}(P(\theta) \parallel P(\theta_0)) \right|_{\theta=\theta_0}$$

must be positive semidefinite. Letting  $\theta_0$  vary (and dropping the subindex 0) the Hessian  $g_{jk}(\theta)$  defines a (possibly degenerate) Riemannian metric on the  $\theta$  parameter space, called the Fisher information metric.

## Fisher information metric theorem

When  $p_{(x,\rho)}$  satisfies the following regularity conditions:

$$\begin{aligned} \frac{\partial \log(p)}{\partial \rho}, \frac{\partial^2 \log(p)}{\partial \rho^2}, \frac{\partial^3 \log(p)}{\partial \rho^3} \text{ exist,} \\ \left| \frac{\partial p}{\partial \rho} \right| < F(x) : \int_{x=0}^{\infty} F(x) dx < \infty, \\ \left| \frac{\partial^2 p}{\partial \rho^2} \right| < G(x) : \int_{x=0}^{\infty} G(x) dx < \infty \\ \left| \frac{\partial^3 \log(p)}{\partial \rho^3} \right| < H(x) : \int_{x=0}^{\infty} p(x, 0) H(x) dx < \xi < \infty \end{aligned}$$

where  $\xi$  is independent of  $\rho$

$$\int_{x=0}^{\infty} \frac{\partial p(x, \rho)}{\partial \rho} \bigg|_{\rho=0} dx = \int_{x=0}^{\infty} \frac{\partial^2 p(x, \rho)}{\partial \rho^2} \bigg|_{\rho=0} dx = 0$$

then:

$$\mathcal{D}(p(x, 0) \parallel p(x, \rho)) = \frac{c\rho^2}{2} + \mathcal{O}(\rho^3) \text{ as } \rho \rightarrow 0.$$

## Variation of information

Another information-theoretic metric is Variation of information, which is roughly a symmetrization of conditional entropy. It is a metric on the set of partitions of a discrete probability space.

## Relation to other quantities of information theory

Many of the other quantities of information theory can be interpreted as applications of relative entropy to specific cases.

## Self-information

The self-information, also known as the information content of a signal, random variable, or event is defined as the negative logarithm of the probability of the given outcome occurring.

When applied to a discrete random variable, the self-information can be represented as

$$I(m) = D_{\text{KL}}(\delta_{\text{im}} \parallel \{p_i\}),$$

is the relative entropy of the probability distribution  $P(i)$  from a Kronecker delta representing certainty that  $i = m$  – i.e. the number of extra bits that must be transmitted to identify  $i$  if only the probability distribution  $P(i)$  is available to the receiver, not the fact that  $i = m$ .

## Mutual information

The mutual information,

$$\begin{aligned} I(X; Y) &= D_{\text{KL}}(P(X, Y) \parallel P(X)P(Y)) \\ &= \mathbb{E}_X\{D_{\text{KL}}(P(Y \mid X) \parallel P(Y))\} \\ &= \mathbb{E}_Y\{D_{\text{KL}}(P(X \mid Y) \parallel P(X))\} \end{aligned}$$



is the relative entropy of the product  $P(X)P(Y)$  of the two marginal probability distributions from the joint probability distribution  $P(X, Y)$  – i.e. the expected number of extra bits that must be transmitted to identify  $X$  and  $Y$  if they are coded using only their marginal distributions instead of the joint distribution. Equivalently, if the joint probability  $P(X, Y)$  is known, it is the expected number of extra bits that must on average be sent to identify  $Y$  if the value of  $X$  is not already known to the receiver.

## Shannon entropy

The Shannon entropy,

$$\begin{aligned} H(X) &= \mathbf{E}[\mathbf{I}_X(x)] \\ &= \log(N) - D_{\text{KL}}(p_X(x) \parallel P_U(X)) \end{aligned}$$

is the number of bits which would have to be transmitted to identify  $X$  from  $N$  equally likely possibilities, *less* the relative entropy of the uniform distribution on the random variates of  $X$ ,  $P_U(X)$ , from the true distribution  $P(X)$  – i.e. *less* the expected number of bits saved, which would have had to be sent if the value of  $X$  were coded according to the uniform distribution  $P_U(X)$  rather than the true distribution  $P(X)$ .

## Conditional entropy

The conditional entropy<sup>[13]</sup>,

$$\begin{aligned} H(X | Y) &= \log(N) - D_{\text{KL}}(P(X, Y) \parallel P_U(X)P(Y)) \\ &= \log(N) - D_{\text{KL}}(P(X, Y) \parallel P(X)P(Y)) - D_{\text{KL}}(P(X) \parallel P_U(X)) \\ &= H(X) - \mathbf{I}(X; Y) \\ &= \log(N) - \mathbf{E}_Y[D_{\text{KL}}(P(X | Y) \parallel P_U(X))] \end{aligned}$$

is the number of bits which would have to be transmitted to identify  $X$  from  $N$  equally likely possibilities, *less* the relative entropy of the product distribution  $P_U(X)P(Y)$  from the true joint distribution  $P(X, Y)$  – i.e. *less* the expected number of bits saved which would have had to be sent if the value of  $X$  were coded according to the uniform distribution  $P_U(X)$  rather than the conditional distribution  $P(X|Y)$  of  $X$  given  $Y$ .

## Cross entropy

When we have a set of possible events, coming from the distribution  $p$ , we can encode them (with a lossless data compression) using entropy encoding. This compresses the data by replacing each fixed-length input symbol with a corresponding unique, variable-length, prefix-free code (e.g.: the events (A, B, C) with probabilities  $p = (1/2, 1/4, 1/4)$  can be encoded as the bits (0, 10, 11)). If we know the distribution  $p$  in advance, we can devise an encoding that would be optimal (e.g.: using Huffman coding). Meaning the messages we encode will have the shortest length on average (assuming the encoded events are sampled from  $p$ ), which will be equal to Shannon's Entropy of  $p$  (denoted as  $\mathbf{H}(p)$ ). However, if we use a different probability distribution ( $q$ ) when creating the entropy encoding scheme, then a larger number of bits will be used (on average) to identify an event from a set of possibilities. This new (larger) number is measured by the cross entropy between  $p$  and  $q$ .

The cross entropy between two probability distributions ( $p$  and  $q$ ) measures the average number of bits needed to identify an event from a set of possibilities, if a coding scheme is used based on a given probability distribution  $q$ , rather than the "true" distribution  $p$ . The cross entropy for two distributions  $p$  and  $q$  over the same probability space is thus defined as follows:

$$\mathbf{H}(p, q) = \mathbf{E}_p[-\log(q)] = \mathbf{H}(p) + D_{\text{KL}}(p \parallel q).$$

Under this scenario, relative entropies can be interpreted as the extra number of bits, on average, that are needed (beyond  $\mathbf{H}(p)$ ) for encoding the events because of using  $q$  for constructing the encoding scheme instead of  $p$ .

# Bayesian updating

In Bayesian statistics, relative entropy can be used as a measure of the information gain in moving from a prior distribution to a posterior distribution:  $p(\mathbf{x}) \rightarrow p(\mathbf{x} \mid I)$ . If some new fact  $\mathbf{Y} = \mathbf{y}$  is discovered, it can be used to update the posterior distribution for  $\mathbf{X}$  from  $p(\mathbf{x} \mid I)$  to a new posterior distribution  $p(\mathbf{x} \mid \mathbf{y}, I)$  using Bayes' theorem:

$$p(\mathbf{x} \mid \mathbf{y}, I) = \frac{p(\mathbf{y} \mid \mathbf{x}, I)p(\mathbf{x} \mid I)}{p(\mathbf{y} \mid I)}$$

This distribution has a new entropy:

$$\mathbf{H}(p(\mathbf{x} \mid \mathbf{y}, I)) = - \sum_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}, I) \log p(\mathbf{x} \mid \mathbf{y}, I),$$

which may be less than or greater than the original entropy  $\mathbf{H}(p(\mathbf{x} \mid I))$ . However, from the standpoint of the new probability distribution one can estimate that to have used the original code based on  $p(\mathbf{x} \mid I)$  instead of a new code based on  $p(\mathbf{x} \mid \mathbf{y}, I)$  would have added an expected number of bits:

$$D_{\text{KL}}(p(\mathbf{x} \mid \mathbf{y}, I) \parallel p(\mathbf{x} \mid I)) = \sum_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}, I) \log \left( \frac{p(\mathbf{x} \mid \mathbf{y}, I)}{p(\mathbf{x} \mid I)} \right)$$

to the message length. This therefore represents the amount of useful information, or information gain, about  $\mathbf{X}$ , that has been learned by discovering  $\mathbf{Y} = \mathbf{y}$ .

If a further piece of data,  $\mathbf{Y}_2 = \mathbf{y}_2$ , subsequently comes in, the probability distribution for  $\mathbf{x}$  can be updated further, to give a new best guess  $p(\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2, I)$ . If one reinvestigates the information gain for using  $p(\mathbf{x} \mid \mathbf{y}_1, I)$  rather than  $p(\mathbf{x} \mid I)$ , it turns out that it may be either greater or less than previously estimated:

$$\sum_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2, I) \log \left( \frac{p(\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2, I)}{p(\mathbf{x} \mid I)} \right) \text{ may be } \le \text{ or } > \text{ than } \sum_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y}_1, I) \log \left( \frac{p(\mathbf{x} \mid \mathbf{y}_1, I)}{p(\mathbf{x} \mid I)} \right)$$

and so the combined information gain does *not* obey the triangle inequality:

$$D_{\text{KL}}(p(\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2, I) \parallel p(\mathbf{x} \mid I)) \text{ may be } <, = \text{ or } > \text{ than } \\ D_{\text{KL}}(p(\mathbf{x} \mid \mathbf{y}_1, \mathbf{y}_2, I) \parallel p(\mathbf{x} \mid \mathbf{y}_1, I)) + D_{\text{KL}}(p(\mathbf{x} \mid \mathbf{y}_1, I) \parallel p(\mathbf{x} \mid I))$$

All one can say is that on *average*, averaging using  $p(\mathbf{y}_2 \mid \mathbf{y}_1, \mathbf{x}, I)$ , the two sides will average out.

## Bayesian experimental design

A common goal in Bayesian experimental design is to maximise the expected relative entropy between the prior and the posterior.<sup>[14]</sup> When posteriors are approximated to be Gaussian distributions, a design maximising the expected relative entropy is called Bayes d-optimal.

## Discrimination information

Relative entropy  $D_{\text{KL}}(p(\mathbf{x} \mid \mathbf{H}_1) \parallel p(\mathbf{x} \mid \mathbf{H}_0))$  can also be interpreted as the expected **discrimination information** for  $\mathbf{H}_1$  over  $\mathbf{H}_0$ : the mean information per sample for discriminating in favor of a hypothesis  $\mathbf{H}_1$  against a hypothesis  $\mathbf{H}_0$ , when hypothesis  $\mathbf{H}_1$  is true.<sup>[15]</sup> Another name for this quantity, given to it by I. J. Good, is the expected weight of evidence for  $\mathbf{H}_1$  over  $\mathbf{H}_0$  to be expected from each sample.

The expected weight of evidence for  $\mathbf{H}_1$  over  $\mathbf{H}_0$  is **not** the same as the information gain expected per sample about the probability distribution  $p(\mathbf{H})$  of the hypotheses,

$$D_{\text{KL}}(p(x | H_1) \parallel p(x | H_0)) \neq IG = D_{\text{KL}}(p(H | x) \parallel p(H | I)).$$

Either of the two quantities can be used as a utility function in Bayesian experimental design, to choose an optimal next question to investigate: but they will in general lead to rather different experimental strategies.

On the entropy scale of *information gain* there is very little difference between near certainty and absolute certainty—coding according to a near certainty requires hardly any more bits than coding according to an absolute certainty. On the other hand, on the logit scale implied by weight of evidence, the difference between the two is enormous – infinite perhaps; this might reflect the difference between being almost sure (on a probabilistic level) that, say, the Riemann hypothesis is correct, compared to being certain that it is correct because one has a mathematical proof. These two different scales of loss function for uncertainty are *both* useful, according to how well each reflects the particular circumstances of the problem in question.

## Principle of minimum discrimination information

The idea of relative entropy as discrimination information led Kullback to propose the Principle of **Minimum Discrimination Information** (MDI): given new facts, a new distribution  $\mathbf{f}$  should be chosen which is as hard to discriminate from the original distribution  $\mathbf{f}_0$  as possible; so that the new data produces as small an information gain  $D_{\text{KL}}(\mathbf{f} \parallel \mathbf{f}_0)$  as possible.

For example, if one had a prior distribution  $p(\mathbf{x}, \mathbf{a})$  over  $\mathbf{x}$  and  $\mathbf{a}$ , and subsequently learnt the true distribution of  $\mathbf{a}$  was  $\mathbf{u}(\mathbf{a})$ , then the relative entropy between the new joint distribution for  $\mathbf{x}$  and  $\mathbf{a}$ ,  $q(\mathbf{x} | \mathbf{a})\mathbf{u}(\mathbf{a})$ , and the earlier prior distribution would be:

$$D_{\text{KL}}(q(\mathbf{x} | \mathbf{a})\mathbf{u}(\mathbf{a}) \parallel p(\mathbf{x}, \mathbf{a})) = \mathbf{E}_{\mathbf{u}(\mathbf{a})}\{D_{\text{KL}}(q(\mathbf{x} | \mathbf{a}) \parallel p(\mathbf{x} | \mathbf{a}))\} + D_{\text{KL}}(\mathbf{u}(\mathbf{a}) \parallel p(\mathbf{a})),$$

i.e. the sum of the relative entropy of  $p(\mathbf{a})$  the prior distribution for  $\mathbf{a}$  from the updated distribution  $\mathbf{u}(\mathbf{a})$ , plus the expected value (using the probability distribution  $\mathbf{u}(\mathbf{a})$ ) of the relative entropy of the prior conditional distribution  $p(\mathbf{x} | \mathbf{a})$  from the new conditional distribution  $q(\mathbf{x} | \mathbf{a})$ . (Note that often the later expected value is called the *conditional relative entropy* (or *conditional Kullback-Leibler divergence*) and denoted by  $D_{\text{KL}}(q(\mathbf{x} | \mathbf{a}) \parallel p(\mathbf{x} | \mathbf{a}))$  <sup>[2][13]:p. 22]</sup> This is minimized if  $q(\mathbf{x} | \mathbf{a}) = p(\mathbf{x} | \mathbf{a})$  over the whole support of  $\mathbf{u}(\mathbf{a})$ ; and we note that this result incorporates Bayes' theorem, if the new distribution  $\mathbf{u}(\mathbf{a})$  is in fact a  $\delta$  function representing certainty that  $\mathbf{a}$  has one particular value.

MDI can be seen as an extension of Laplace's Principle of Insufficient Reason, and the Principle of Maximum Entropy of E.T. Jaynes. In particular, it is the natural extension of the principle of maximum entropy from discrete to continuous distributions, for which Shannon entropy ceases to be so useful (see differential entropy), but the relative entropy continues to be just as relevant.

In the engineering literature, MDI is sometimes called the **Principle of Minimum Cross-Entropy** (MCE) or **Minxent** for short. Minimising relative entropy from  $\mathbf{m}$  to  $\mathbf{p}$  with respect to  $\mathbf{m}$  is equivalent to minimizing the cross-entropy of  $\mathbf{p}$  and  $\mathbf{m}$ , since

$$\mathbf{H}(\mathbf{p}, \mathbf{m}) = \mathbf{H}(\mathbf{p}) + D_{\text{KL}}(\mathbf{p} \parallel \mathbf{m}),$$

which is appropriate if one is trying to choose an adequate approximation to  $\mathbf{p}$ . However, this is just as often *not* the task one is trying to achieve. Instead, just as often it is  $\mathbf{m}$  that is some fixed prior reference measure, and  $\mathbf{p}$  that one is attempting to optimise by minimising  $D_{\text{KL}}(\mathbf{p} \parallel \mathbf{m})$  subject to some constraint. This has led to some ambiguity in the literature, with some authors attempting to resolve the inconsistency by redefining cross-entropy to be  $D_{\text{KL}}(\mathbf{p} \parallel \mathbf{m})$ , rather than  $\mathbf{H}(\mathbf{p}, \mathbf{m})$ .

## Relationship to available work

Surprisals<sup>[16]</sup> add where probabilities multiply. The surprisal for an event of probability  $\mathbf{p}$  is defined as  $\mathbf{s} = \mathbf{k} \ln(1/\mathbf{p})$ . If  $\mathbf{k}$  is  $\{1, 1/\ln 2, 1.38 \times 10^{-23}\}$  then surprisal is in  $\{\text{nats, bits, or } J/K\}$  so that, for instance, there are  $N$  bits of surprisal for landing all "heads" on a toss of  $N$  coins.

Best-guess states (e.g. for atoms in a gas) are inferred by maximizing the *average surprisal*  $\mathcal{S}$  (entropy) for a given set of control parameters (like pressure  $P$  or volume  $V$ ). This constrained entropy maximization, both classically<sup>[17]</sup> and quantum mechanically,<sup>[18]</sup> minimizes Gibbs availability in entropy units<sup>[19]</sup>  $A \equiv -k \ln(Z)$  where  $Z$  is a constrained multiplicity or partition function.

When temperature  $T$  is fixed, free energy ( $T \times A$ ) is also minimized. Thus if  $T, V$  and number of molecules  $N$  are constant, the Helmholtz free energy  $F \equiv U - TS$  (where  $U$  is energy) is minimized as a system "equilibrates." If  $T$  and  $P$  are held constant (say during processes in your body), the Gibbs free energy  $G = U + PV - TS$  is minimized instead. The change in free energy under these conditions is a measure of available work that might be done in the process. Thus available work for an ideal gas at constant temperature  $T_o$  and pressure  $P_o$  is  $W = \Delta G = NkT_o \Theta(V/V_o)$  where  $V_o = NkT_o/P_o$  and  $\Theta(x) = x - 1 - \ln x \geq 0$  (see also Gibbs inequality).

More generally<sup>[20]</sup> the work available relative to some ambient is obtained by multiplying ambient temperature  $T_o$  by relative entropy or *net surprisal*  $\Delta I \geq 0$ , defined as the average value of  $k \ln(p/p_o)$  where  $p_o$  is the probability of a given state under ambient conditions. For instance, the work available in equilibrating a monatomic ideal gas to ambient values of  $V_o$  and  $T_o$  is thus  $W = T_o \Delta I$ , where relative entropy

$$\Delta I = Nk \left[ \Theta \left( \frac{V}{V_o} \right) + \frac{3}{2} \Theta \left( \frac{T}{T_o} \right) \right].$$

The resulting contours of constant relative entropy, shown at right for a mole of Argon at standard temperature and pressure, for example put limits on the conversion of hot to cold as in flame-powered air-conditioning or in the unpowered device to convert boiling-water to ice-water discussed here.<sup>[21]</sup> Thus relative entropy measures thermodynamic availability in bits.

## Quantum information theory

For density matrices  $P$  and  $Q$  on a Hilbert space, the quantum relative entropy from  $Q$  to  $P$  is defined to be

$$D_{\text{KL}}(P \parallel Q) = \text{Tr}(P(\log(P) - \log(Q))).$$

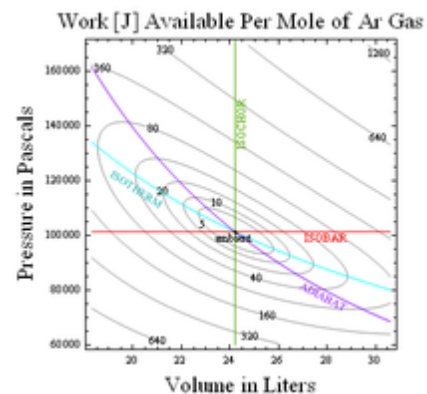
In quantum information science the minimum of  $D_{\text{KL}}(P \parallel Q)$  over all separable states  $Q$  can also be used as a measure of entanglement in the state  $P$ .

## Relationship between models and reality

Just as relative entropy of "actual from ambient" measures thermodynamic availability, relative entropy of "reality from a model" is also useful even if the only clues we have about reality are some experimental measurements. In the former case relative entropy describes *distance to equilibrium* or (when multiplied by ambient temperature) the amount of *available work*, while in the latter case it tells you about surprises that reality has up its sleeve or, in other words, *how much the model has yet to learn*.

Although this tool for evaluating models against systems that are accessible experimentally may be applied in any field, its application to selecting a statistical model via Akaike information criterion are particularly well described in papers<sup>[22]</sup> and a book<sup>[23]</sup> by Burnham and Anderson. In a nutshell the relative entropy of reality from a model may be estimated, to within a constant additive term, by a function of the deviations observed between data and the model's predictions (like the mean squared deviation) . Estimates of such divergence for models that share the same additive term can in turn be used to select among models.

When trying to fit parametrized models to data there are various estimators which attempt to minimize relative entropy, such as maximum likelihood and maximum spacing estimators.



Pressure versus volume plot of available work from a mole of argon gas relative to ambient, calculated as  $T_o$  times the Kullback–Leibler divergence.

# Symmetrised divergence

Kullback and Leibler themselves actually defined the divergence as:

$$D_{\text{KL}}(P \parallel Q) + D_{\text{KL}}(Q \parallel P)$$

which is symmetric and nonnegative. This quantity has sometimes been used for feature selection in classification problems, where  $P$  and  $Q$  are the conditional pdfs of a feature under two different classes. In the Banking and Finance industries, this quantity is referred to as Population Stability Index (PSI), and is used to assess distributional shifts in model features through time.

An alternative is given via the  $\lambda$  divergence,

$$D_{\lambda}(P \parallel Q) = \lambda D_{\text{KL}}(P \parallel \lambda P + (1 - \lambda)Q) + (1 - \lambda)D_{\text{KL}}(Q \parallel \lambda P + (1 - \lambda)Q),$$

which can be interpreted as the expected information gain about  $X$  from discovering which probability distribution  $X$  is drawn from,  $P$  or  $Q$ , if they currently have probabilities  $\lambda$  and  $1 - \lambda$  respectively.

The value  $\lambda = 0.5$  gives the Jensen–Shannon divergence, defined by

$$D_{\text{JS}} = \frac{1}{2}D_{\text{KL}}(P \parallel M) + \frac{1}{2}D_{\text{KL}}(Q \parallel M)$$

where  $M$  is the average of the two distributions,

$$M = \frac{1}{2}(P + Q).$$

$D_{\text{JS}}$  can also be interpreted as the capacity of a noisy information channel with two inputs giving the output distributions  $P$  and  $Q$ . The Jensen–Shannon divergence, like all  $f$ -divergences, is *locally* proportional to the Fisher information metric. It is similar to the Hellinger metric (in the sense that it induces the same affine connection on a statistical manifold).

Furthermore The Jensen-Shannon divergence can be generalized using abstract statistical M-mixtures relying on an abstract mean  $M$ . <sup>[24]</sup> <sup>[25]</sup>

## Relationship to other probability-distance measures

There are many other important measures of probability distance. Some of these are particularly connected with relative entropy. For example:

- The total variation distance,  $\delta(p, q)$ . This is connected to the divergence through Pinsker's inequality:

$$\delta(P, Q) \leq \sqrt{\frac{1}{2}D_{\text{KL}}(P \parallel Q)}$$

- The family of Rényi divergences generalize relative entropy. Depending on the value of a certain parameter,  $\alpha$ , various inequalities may be deduced.

Other notable measures of distance include the Hellinger distance, histogram intersection, Chi-squared statistic, quadratic form distance, match distance, Kolmogorov–Smirnov distance, and earth mover's distance.<sup>[26]</sup>

## Data differencing

Just as *absolute* entropy serves as theoretical background for data *compression*, *relative* entropy serves as theoretical background for data differencing – the absolute entropy of a set of data in this sense being the data required to reconstruct it (minimum compressed size), while the relative entropy of a target set of data, given a source set of data, is the data required to reconstruct the target *given* the source (minimum size of a patch).

# See also

- [Akaike information criterion](#)
- [Bayesian information criterion](#)
- [Bregman divergence](#)
- [Cross-entropy](#)
- [Deviance information criterion](#)
- [Entropic value at risk](#)
- [Entropy power inequality](#)
- [Hellinger distance](#)
- [Information gain in decision trees](#)
- [Information gain ratio](#)
- [Information theory and measure theory](#)
- [Jensen–Shannon divergence](#)
- [Quantum relative entropy](#)
- [Solomon Kullback](#) and [Richard Leibler](#)

# References

1. Kullback, S.; Leibler, R.A. (1951). "On information and sufficiency" (<https://doi.org/10.1214%2Faoms%2F1177729694>). *Annals of Mathematical Statistics*. **22** (1): 79–86. doi:10.1214/aoms/1177729694 (<https://doi.org/10.1214%2Faoms%2F1177729694>). JSTOR 2236703 (<https://www.jstor.org/stable/2236703>). MR 0039968 (<https://www.ams.org/mathscinet-getitem?mr=0039968>).
2. Kullback, S. (1959), *Information Theory and Statistics*, John Wiley & Sons. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.
3. Kullback, S. (1987). "Letter to the Editor: The Kullback–Leibler distance". *The American Statistician*. **41** (4): 340–341. doi:10.1080/00031305.1987.10475510 (<https://doi.org/10.1080%2F00031305.1987.10475510>). JSTOR 2684769 (<https://www.jstor.org/stable/2684769>).
4. MacKay, David J.C. (2003). *Information Theory, Inference, and Learning Algorithms* ([https://books.google.com/books?id=AKuMj4PN\\_EMC&q=%22Kullback%E2%80%93Leibler+divergence%22](https://books.google.com/books?id=AKuMj4PN_EMC&q=%22Kullback%E2%80%93Leibler+divergence%22)) (First ed.). Cambridge University Press. p. 34. ISBN 9780521642989.
5. Bishop C. (2006). *Pattern Recognition and Machine Learning*
6. Burnham, K. P.; Anderson, D. R. (2002). *Model Selection and Multi-Model Inference* (<https://archive.org/details/modelselectionmu0000burn>) (2nd ed.). Springer. p. 51 (<https://archive.org/details/modelselectionmu0000burn/page/51>). ISBN 9780387953649.
7. Hobson, Arthur (1971). *Concepts in statistical mechanics*. New York: Gordon and Breach. ISBN 978-0677032405.
8. Sanov, I.N. (1957). "On the probability of large deviations of random magnitudes". *Mat. Sbornik*. **42** (84): 11–44.
9. Novak S.Y. (2011), *Extreme Value Methods with Applications to Finance* ch. 14.5 (Chapman & Hall). ISBN 978-1-4398-3574-6.
10. See the section "differential entropy – 4" in [Relative Entropy](http://videlectures.net/nips09_verdu_re/) ([http://videlectures.net/nips09\\_verdu\\_re/](http://videlectures.net/nips09_verdu_re/)) video lecture by Sergio Verdú NIPS 2009
11. Lee, Se Yoon (2021). "Gibbs sampler and coordinate ascent variational inference: A set-theoretical review". *Communications in Statistics - Theory and Methods*. arXiv:2008.01006 (<https://arxiv.org/abs/2008.01006>). doi:10.1080/03610926.2021.1921214 (<https://doi.org/10.1080%2F03610926.2021.1921214>).
12. Duchi J., "[Derivations for Linear Algebra and Optimization](http://web.stanford.edu/~jduchi/projects/general_notes.pdf) ([http://web.stanford.edu/~jduchi/projects/general\\_notes.pdf](http://web.stanford.edu/~jduchi/projects/general_notes.pdf))".
13. Cover, Thomas M.; Thomas, Joy A. (1991), *Elements of Information Theory*, John Wiley & Sons
14. Chaloner, K.; Verdinelli, I. (1995). "Bayesian experimental design: a review" (<https://doi.org/10.1214%2Fss%2F1177009939>). *Statistical Science*. **10** (3): 273–304. doi:10.1214/ss/1177009939 (<https://doi.org/10.1214%2Fss%2F1177009939>).
15. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. (2007). "Section 14.7.2. Kullback–Leibler Distance" (<http://apps.nrbook.com/empanel/index.html#pg=756>). *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). Cambridge University Press. ISBN 978-0-521-88068-8.
16. Myron Tribus (1961), *Thermodynamics and Thermostatistics* (D. Van Nostrand, New York)
17. Jaynes, E. T. (1957). "Information theory and statistical mechanics" (<http://bayes.wustl.edu/etj/articles/theory.1.pdf>) (PDF). *Physical Review*. **106** (4): 620–630. Bibcode:1957PhRv..106..620J (<https://ui.adsabs.harvard.edu/abs/1957PhRv..106..620J>). doi:10.1103/physrev.106.620 (<https://doi.org/10.1103%2Fphysrev.106.620>).
18. Jaynes, E. T. (1957). "Information theory and statistical mechanics II" (<http://bayes.wustl.edu/etj/articles/theory.2.pdf>) (PDF). *Physical Review*. **108** (2): 171–190. Bibcode:1957PhRv..108..171J (<https://ui.adsabs.harvard.edu/abs/1957PhRv..108..171J>). doi:10.1103/physrev.108.171 (<https://doi.org/10.1103%2Fphysrev.108.171>).

19. J.W. Gibbs (1873), "A method of geometrical representation of thermodynamic properties of substances by means of surfaces", reprinted in *The Collected Works of J. W. Gibbs, Volume I Thermodynamics*, ed. W. R. Longley and R. G. Van Name (New York: Longmans, Green, 1931) footnote page 52.
20. Tribus, M.; McIrvine, E. C. (1971). "Energy and information". *Scientific American*. **224** (3): 179–186. Bibcode:1971SciAm.225c.179T (<https://ui.adsabs.harvard.edu/abs/1971SciAm.225c.179T>). doi:10.1038/scientificamerican0971-179 (<https://doi.org/10.1038%2Fscientificamerican0971-179>).
21. Fraundorf, P. (2007). "Thermal roots of correlation-based complexity" (<https://archive.today/20110813083358/http://www3.interscience.wiley.com/cgi-bin/abstract/117861985/ABSTRACT>). *Complexity*. **13** (3): 18–26. arXiv:1103.2481 (<https://arxiv.org/abs/1103.2481>). Bibcode:2008Cmplx..13c..18F (<https://ui.adsabs.harvard.edu/abs/2008Cmplx..13c..18F>). doi:10.1002/cplx.20195 (<https://doi.org/10.1002%2Fcplx.20195>). S2CID 20794688 (<https://api.semanticscholar.org/CorpusID:20794688>). Archived from the original (<http://www3.interscience.wiley.com/cgi-bin/abstract/117861985/ABSTRACT>) on 2011-08-13.
22. Burnham, K.P.; Anderson, D.R. (2001). "Kullback–Leibler information as a basis for strong inference in ecological studies" (<https://doi.org/10.1071%2FWR99107>). *Wildlife Research*. **28** (2): 111–119. doi:10.1071/WR99107 (<https://doi.org/10.1071%2FWR99107>).
23. Burnham, K. P. and Anderson D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition* (Springer Science) ISBN 978-0-387-95364-9.
24. Nielsen, Frank (2019). "On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means" (<https://doi.org/10.3390%2Fe21050485>). *Entropy*. **21** (5): 485. doi:10.3390/e21050485 (<https://doi.org/10.3390%2Fe21050485>).
25. Nielsen, Frank (2020). "On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid" (<https://doi.org/10.3390%2Fe22020221>). *Entropy*. **22** (2): 221. doi:10.3390/e22020221 (<https://doi.org/10.3390%2Fe22020221>).
26. Rubner, Y.; Tomasi, C.; Guibas, L. J. (2000). "The earth mover's distance as a metric for image retrieval". *International Journal of Computer Vision*. **40** (2): 99–121. doi:10.1023/A:1026543900054 (<https://doi.org/10.1023%2FA%3A1026543900054>). S2CID 14106275 (<https://api.semanticscholar.org/CorpusID:14106275>).

## External links

- Information Theoretical Estimators Toolbox (<https://bitbucket.org/szzoli/ite/>)
- Ruby gem for calculating Kullback–Leibler divergence (<https://github.com/evansenter/diverge>)
- Jon Shlens' tutorial on Kullback–Leibler divergence and likelihood theory (<https://arxiv.org/abs/1404.2000>)
- Matlab code for calculating Kullback–Leibler divergence for discrete distributions (<http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=13089&objectType=file>)
- Sergio Verdú, [Relative Entropy](http://videolectures.net/nips09_verdu_re/) ([http://videolectures.net/nips09\\_verdu\\_re/](http://videolectures.net/nips09_verdu_re/)), NIPS 2009. One-hour video lecture.
- A modern summary of info-theoretic divergence measures (<https://arxiv.org/abs/math/0604246>)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Kullback–Leibler\_divergence&oldid=1024796164"

**This page was last edited on 24 May 2021, at 03:56 (UTC).**

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.