Kullback–Leibler divergence

In mathematical statistics, the **Kullback–Leibler** (**KL**) **divergence** (also called **relative entropy** and **I-divergence**[1]), denoted $D_{\mathrm{KL}}(P \parallel Q)$, is a type of statistical distance: a measure of how much a model probability distribution Q is different from a true probability distribution P.[2][3] Mathematically, it is defined as

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

A simple interpretation of the KL divergence of P from Q is the expected excess surprise from using Q as a model instead of P when the actual distribution is P. While it is a measure of how different two distributions are, and in some sense is thus a "distance", it is not actually a metric, which is the most familiar and formal type of distance. In particular, it is not symmetric in the two distributions (in contrast to variation of information), and does not satisfy the triangle inequality. Instead, in terms of information geometry, it is a type of divergence,[4] a generalization of squared distance, and for certain classes of distributions (notably an exponential family), it satisfies a generalized Pythagorean theorem (which applies to squared distances).[5]

Relative entropy is always a non-negative real number, with value 0 if and only if the two distributions in question are identical. It has diverse applications, both theoretical, such as characterizing the relative (Shannon) entropy in information systems, randomness in continuous time-series, and information gain when comparing statistical models of inference; and practical, such as applied statistics, fluid mechanics, neuroscience, bioinformatics, and machine learning.

Introduction and context

Consider two probability distributions P and Q. Usually, P represents the data, the observations, or a measured probability distribution. Distribution Q represents instead a theory, a model, a description or an approximation of P. The Kullback–Leibler divergence $D_{\mathrm{KL}}(P \parallel Q)$ 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. Note that the roles of P and Q can be reversed in some situations where that is easier to compute, such as with the expectation–maximization algorithm (EM) and evidence lower bound (ELBO) computations.

Etymology

The relative entropy was introduced by Solomon Kullback and Richard Leibler in Kullback & Leibler (1951) as "the mean information for discrimination between H_1 and H_2 per observation from μ_1",[6] where one is comparing two probability measures μ_1, μ_2, and H_1, H_2 are the hypotheses that one is selecting from measure μ_1, μ_2 (respectively). They denoted this by $I(1:2)$, and defined the "'divergence' between μ_1 and μ_2" as the symmetrized quantity $J(1,2) = I(1:2) + I(2:1)$, which had already been defined and used by Harold Jeffreys in 1948.[7] In Kullback (1959), the symmetrized form is again referred to as the "divergence", and the relative entropies in each direction are referred to as a "directed divergences" between two distributions;[8] Kullback preferred the term **discrimination information**.[9] The term "divergence" is in contrast to a distance (metric), since the symmetrized divergence does not satisfy the triangle inequality.[10] Numerous references to earlier uses

of the symmetrized divergence and to other statistical distances are given in Kullback (1959, pp. 6–7, §1.3 Divergence). The asymmetric "directed divergence" has come to be known as the Kullback–Leibler divergence, while the symmetrized "divergence" is now referred to as the **Jeffreys divergence**.

Definition

For discrete probability distributions P and Q defined on the same sample space, $\mathcal{X}$, the relative entropy from Q to P is defined[11] to be

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

which is equivalent to

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

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 only defined in this way if, for all x, $Q(x) = 0$ implies $P(x) = 0$ (absolute continuity). Otherwise, it is often defined as $+\infty$,[1] but the value $+\infty$ is possible even if $Q(x) \neq 0$ everywhere,[12][13] provided that $\mathcal{X}$ is infinite in extent. Analogous comments apply to the continuous and general measure cases defined below.

Whenever $P(x)$ is zero the contribution of the corresponding term is interpreted as zero because

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

For distributions P and Q of a continuous random variable, relative entropy is defined to be the integral[14]

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

where p and q denote the probability densities of P and Q.

More generally, if P and Q are probability measures on a measurable space $\mathcal{X}$, and P is absolutely continuous with respect to Q, then the relative entropy from Q to P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{x \in \mathcal{X}} \log\left(\frac{P(\mathrm{d}x)}{Q(\mathrm{d}x)}\right) P(\mathrm{d}x),$$

where $\frac{P(\mathrm{d}x)}{Q(\mathrm{d}x)}$ is the Radon–Nikodym derivative of P with respect to Q, i.e. the unique Q almost everywhere defined function r on $\mathcal{X}$ such that $P(\mathrm{d}x) = r(x)Q(\mathrm{d}x)$ which exists because P is absolutely continuous with respect to Q. Also we assume the expression on the right-hand side exists. Equivalently (by the chain rule), this can be written as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{x \in \mathcal{X}} \frac{P(\mathrm{d}x)}{Q(\mathrm{d}x)} \log\left(\frac{P(\mathrm{d}x)}{Q(\mathrm{d}x)}\right) Q(\mathrm{d}x),$$

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

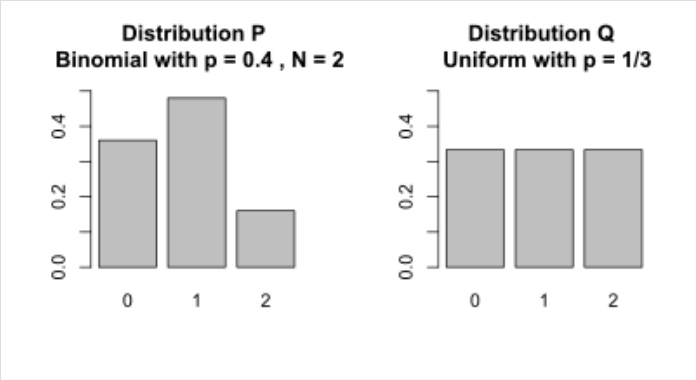
$$D_{\mathrm{KL}}(P \parallel Q) = \int_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mu(\mathrm{d}x) .$$

Note that such a measure μ for which densities can be defined always exists, since one can take $\mu = \frac{1}{2}(P+Q)$ although in practice it will usually be one that in the context like counting measure for discrete distributions, or Lebesgue measure or a convenient variant thereof like Gaussian measure or the uniform measure on the sphere, Haar measure on a Lie group etc. for continuous distributions. The logarithms in these formulae are usually 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 $D_{\mathrm{KL}}(P \parallel Q)$ 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 $D_{\mathrm{KL}}(P \parallel Q)$ is as the relative entropy of P *with respect to* Q or the information gain from P over Q.

Basic example

Kullback[3] 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 $x = 0, 1, 2$ (i.e. $\mathcal{X} = \{0, 1, 2\}$), each with probability $p = 1/3$.



Two distributions to illustrate relative entropy

x	0	1	2
Distribution $P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
Distribution $Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

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

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

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

Interpretations

Statistics

In the field of statistics, the Neyman–Pearson lemma states that the most powerful way to distinguish between the two distributions P and Q based on an observation Y (drawn from one of them) is through the log of the ratio of their likelihoods: $\log P(Y) - \log Q(Y)$. The KL divergence is the expected value of this statistic if Y is actually drawn from P. Kullback motivated the statistic as an expected log likelihood ratio.[15]

Coding

In the context of coding theory, $D_{\mathrm{KL}}(P \parallel Q)$ 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.

Inference

In the context of machine learning, $D_{\mathrm{KL}}(P \parallel Q)$ 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.

Expressed in the language of Bayesian inference, $D_{\mathrm{KL}}(P \parallel Q)$ 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.[16]

Information geometry

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 the KL divergence and compute an information projection.

While it is a statistical distance, it is not a metric, the most familiar type of distance, but instead it is a divergence.[4] While metrics are symmetric and generalize *linear* distance, satisfying the triangle inequality, divergences are asymmetric and generalize *squared* distance, in some cases satisfying a generalized Pythagorean theorem. In general $D_{\mathrm{KL}}(P \parallel Q)$ does not equal $D_{\mathrm{KL}}(Q \parallel P)$, and the asymmetry is an important part of the geometry.[4] The infinitesimal form of relative entropy, specifically its Hessian, gives a metric tensor that equals the Fisher information metric; see § Fisher information metric. Fisher information metric on the certain probability distribution let determine the natural gradient for information-geometric optimization algorithms.[17] Its quantum version is Fubini-study metric.[18] Relative entropy satisfies a generalized Pythagorean theorem for exponential families (geometrically interpreted as dually flat manifolds), and this allows one to minimize relative entropy by geometric means, for example by information projection and in maximum likelihood estimation.[5]

The relative entropy is the Bregman divergence generated by the negative entropy, but it is also of the form of an f-divergence. For probabilities over a finite alphabet, it is unique in being a member of both of these classes of statistical divergences. The application of Bregman divergence can be found in mirror descent.[19]

Finance (game theory)

Consider a growth-optimizing investor in a fair game with mutually exclusive outcomes (e.g. a “horse race” in which the official odds add up to one). The rate of return expected by such an investor is equal to the relative entropy between the investor's believed probabilities and the official odds.[20] This is a special case of a much more general connection between financial returns and divergence measures.[21]

Financial risks are connected to D_{KL} via information geometry.[22] Investors' views, the prevailing market view, and risky scenarios form triangles on the relevant manifold of probability distributions. The shape of the triangles determines key financial risks (both qualitatively and quantitatively). For instance, obtuse triangles in which investors' views and risk scenarios appear on “opposite sides” relative to the market describe negative risks, acute triangles describe positive exposure, and the right-angled situation in the middle corresponds to zero risk. Extending this concept, relative entropy can be hypothetically utilised to identify the behaviour of informed investors, if one takes this to be represented by the magnitude and deviations away from the prior expectations of fund flows, for example[23].

Motivation

In information theory, the Kraft–McMillan theorem establishes that any directly decodable coding scheme for coding a message to identify one value x_i out of a set of possibilities X can be seen as representing an implicit probability distribution $q(x_i) = 2^{-\ell_i}$ over X, where ℓ_i is the length of the code for 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: it is the *excess* entropy.

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} - \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \mathrm{H}(P, Q) - \mathrm{H}(P) \end{aligned}$$

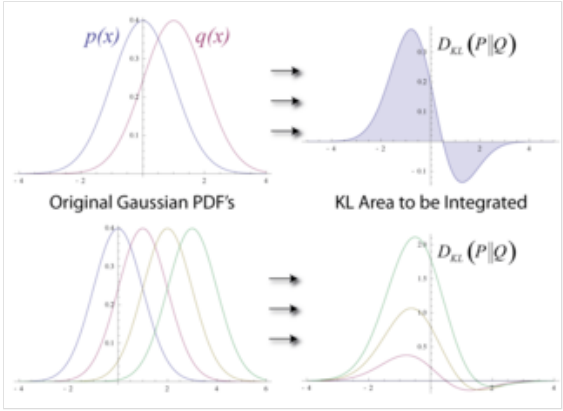


Illustration of the relative entropy for two normal distributions. The typical asymmetry is clearly visible.

where $\mathrm{H}(P, Q)$ is the cross entropy of Q relative to P and $\mathrm{H}(P)$ is the entropy of P (which is the same as the cross-entropy of P with itself).

The relative entropy $D_{\mathrm{KL}}(P \parallel Q)$ can be thought of geometrically as a statistical distance, a measure of how far the distribution Q is from the distribution P. Geometrically it is a divergence: an asymmetric, generalized form of squared distance. The cross-entropy $H(P, Q)$ is itself such a measurement (formally a loss function), but it cannot be thought of as a distance, since $H(P, P) =: H(P)$ is not zero. This can be fixed by subtracting $H(P)$ to make $D_{\mathrm{KL}}(P \parallel Q)$ agree more closely with our notion of distance, as the *excess* loss. The resulting function is asymmetric, and while this can be symmetrized (see § Symmetrised divergence), the asymmetric form is more useful. See § Interpretations for more on the geometric interpretation.

Relative entropy relates to "rate function" in the theory of large deviations.[24][25]

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.[26] 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.

Properties

- Relative entropy is always non-negative,

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

a result known as Gibbs' inequality, with $D_{\mathrm{KL}}(P \parallel Q)$ equals zero if and only if $P = Q$ as measures.

In particular, if $P(dx) = p(x)\mu(dx)$ and $Q(dx) = q(x)\mu(dx)$, then $p(x) = q(x)$ μ-almost everywhere. The entropy $\mathrm{H}(P)$ thus sets a minimum value for the cross-entropy $\mathrm{H}(P, Q)$, 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.

- No upper-bound exists for the general case. However, it is shown that if P and Q are two discrete probability distributions built by distributing the same discrete quantity, then the maximum value of $D_{\mathrm{KL}}(P \parallel Q)$ can be calculated.[27]
- 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 $P(dx) = p(x)\,dx = \tilde{p}(y)\,dy = \tilde{p}(y(x))\left|\frac{dy}{dx}(x)\right|dx$ and $Q(dx) = q(x)\,dx = \tilde{q}(y)\,dy = \tilde{q}(y)\left|\frac{dy}{dx}(x)\right|dx$ where $\left|\frac{dy}{dx}(x)\right|$ is the absolute value of the derivative or more generally of the Jacobian, the relative entropy may be rewritten:

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

  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. $P(dx) = p(x)\,dx$ 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[28] (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, and $P(dx, dy) = P_1(dx)P_2(dy)$, and likewise $Q(dx, dy) = Q_1(dx)Q_2(dy)$ for independent distributions Q_1, Q_2 then

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

- Relative entropy $D_{\mathrm{KL}}(P \parallel Q)$ is convex in the pair of probability measures (P, Q), i.e. if (P_1, Q_1) and (P_2, Q_2) are two pairs of probability measures then

$$D_{\mathrm{KL}}(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D_{\mathrm{KL}}(P_1 \parallel Q_1) + (1-\lambda) D_{\mathrm{KL}}(P_2 \parallel Q_2) \text{ for } 0 \leq \lambda \leq 1.$$

- $D_{\mathrm{KL}}(P \parallel Q)$ may be Taylor expanded about its minimum (i.e. $P = Q$) as

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \sum_{x \in \mathcal{X}} \frac{(Q(x) - P(x))^n}{Q(x)^{n-1}}$$

  which converges if and only if $P \leq 2Q$ almost surely w.r.t Q.

[Proof]

Denote $f(\alpha) := D_{\mathrm{KL}}((1-\alpha)Q + \alpha P \parallel Q)$ and note that $D_{\mathrm{KL}}(P \parallel Q) = f(1)$. The first derivative of f may be derived and evaluated as follows

$$\begin{aligned} f'(\alpha) &= \sum_{x \in \mathcal{X}} (P(x) - Q(x)) \left(\log\left(\frac{(1-\alpha)Q(x) + \alpha P(x)}{Q(x)}\right) + 1\right) \\ &= \sum_{x \in \mathcal{X}} (P(x) - Q(x)) \log\left(\frac{(1-\alpha)Q(x) + \alpha P(x)}{Q(x)}\right) \\ f'(0) &= 0 \end{aligned}$$

Further derivatives may be derived and evaluated as follows

$$f''(\alpha) = \sum_{x\in\mathcal{X}} \frac{(P(x)-Q(x))^2}{(1-\alpha)Q(x)+\alpha P(x)}$$
$$f''(0) = \sum_{x\in\mathcal{X}} \frac{(P(x)-Q(x))^2}{Q(x)}$$
$$f^{(n)}(\alpha) = (-1)^n(n-2)!\sum_{x\in\mathcal{X}} \frac{(P(x)-Q(x))^n}{((1-\alpha)Q(x)+\alpha P(x))^{n-1}}$$
$$f^{(n)}(0) = (-1)^n(n-2)!\sum_{x\in\mathcal{X}} \frac{(P(x)-Q(x))^n}{Q(x)^{n-1}}$$

Hence solving for $D_{\mathrm{KL}}(P \parallel Q)$ via the Taylor expansion of f about 0 evaluated at $\alpha = 1$ yields

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} \\ &= \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \sum_{x\in\mathcal{X}} \frac{(Q(x)-P(x))^n}{Q(x)^{n-1}} \end{aligned}$$

$P \leq 2Q$ a.s. is a sufficient condition for convergence of the series by the following absolute convergence argument

$$\begin{aligned} \sum_{n=2}^{\infty} \left| \frac{1}{n(n-1)} \sum_{x\in\mathcal{X}} \frac{(Q(x)-P(x))^n}{Q(x)^{n-1}} \right| &= \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \sum_{x\in\mathcal{X}} |Q(x)-P(x)| \left|1-\frac{P(x)}{Q(x)}\right|^{n-1} \\ &\leq \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \sum_{x\in\mathcal{X}} |Q(x)-P(x)| \\ &\leq \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \\ &= 1 \end{aligned}$$

$P \leq 2Q$ a.s. is also a necessary condition for convergence of the series by the following proof by contradiction. Assume that $P > 2Q$ with measure strictly greater than 0. It then follows that there must exist some values $\epsilon > 0$, $\rho > 0$, and $U < \infty$ such that $P \geq 2Q + \epsilon$ and $Q \leq U$ with measure ρ. The previous proof of sufficiency demonstrated that the measure $1-\rho$ component of the series where $P \leq 2Q$ is bounded, so we need only concern ourselves with the behavior of the measure ρ component of the series where $P \geq 2Q + \epsilon$. The absolute value of the nth term of this component of the series is then lower bounded by $\frac{1}{n(n-1)}\rho\left(1+\frac{\epsilon}{U}\right)^n$, which is unbounded as $n \to \infty$, so the series diverges.

Duality formula for variational inference

The following result, due to Donsker and Varadhan,[29] is known as **Donsker and Varadhan's variational formula**.

Theorem [Duality Formula for Variational Inference] — Let Θ be a set endowed with an appropriate σ-field $\mathcal{F}$, and two probability measures P and Q, which formulate two probability spaces $(\Theta, \mathcal{F}, P)$ and $(\Theta, \mathcal{F}, Q)$, with $Q \ll P$. ($Q \ll P$ indicates that Q is absolutely continuous with respect to P.) Let h be a real-valued integrable random variable on $(\Theta, \mathcal{F}, P)$. Then the following equality holds

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

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

$$\frac{Q(d\theta)}{P(d\theta)} = \frac{\exp h(\theta)}{E_P[\exp h]},$$

almost surely with respect to probability measure P, where $\frac{Q(d\theta)}{P(d\theta)}$ denotes the Radon-Nikodym derivative of Q with respect to P.

Proof

For a short proof assuming integrability of $\exp(h)$ with respect to P, let Q^* have P-density $\frac{\exp h(\theta)}{E_P[\exp h]}$, i.e. $Q^*(d\theta) = \frac{\exp h(\theta)}{E_P[\exp h]} P(d\theta)$ Then

$$D_{\mathrm{KL}}(Q \parallel Q^*) - D_{\mathrm{KL}}(Q \parallel P) = -E_Q[h] + \log E_P[\exp h].$$

Therefore,

$$E_Q[h] - D_{\mathrm{KL}}(Q \parallel P) = \log E_P[\exp h] - D_{\mathrm{KL}}(Q \parallel Q^*) \leq \log E_P[\exp h],$$

where the last inequality follows from $D_{\mathrm{KL}}(Q \parallel Q^*) \geq 0$, for which equality occurs if and only if $Q = Q^*$. The conclusion follows.

Examples

Multivariate normal distributions

Suppose that we have two multivariate normal distributions, with means μ_0, μ_1 and with (non-singular) covariance matrices Σ_0, Σ_1. If the two distributions have the same dimension, k, then the relative entropy between the distributions is as follows:[30]

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

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.

In a numerical implementation, it is helpful to express the result in terms of the Cholesky decompositions L_0, L_1 such that $\Sigma_0 = L_0 L_0^T$ and $\Sigma_1 = L_1 L_1^T$. Then with M and y solutions to the triangular linear systems $L_1 M = L_0$, and $L_1 y = \mu_1 - \mu_0$,

$$D_{\mathrm{KL}}\left(\mathcal{N}_0 \parallel \mathcal{N}_1\right) = \frac{1}{2}\left(\sum_{i,j=1}^{k} (M_{ij})^2 - k + |y|^2 + 2\sum_{i=1}^{k} \ln \frac{(L_1)_{ii}}{(L_0)_{ii}}\right).$$

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):

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

For two univariate normal distributions p and q the above simplifies to[31]

$$D_{\mathrm{KL}}(p \parallel q) = \log \frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_0 - \mu_1)^2}{2\sigma_1^2} - \frac{1}{2}$$

In the case of co-centered normal distributions with $k = \sigma_1/\sigma_0$, this simplifies[32] to:

$$D_{\mathrm{KL}}(p \parallel q) = \log_2 k + (k^{-2} - 1)/2/\ln(2)\mathrm{bits}$$

Uniform distributions

Consider two uniform distributions, with the support of $p = [A, B]$ enclosed within $q = [C, D]$ ($C \leq A < B \leq D$). Then the information gain is:

$$D_{\mathrm{KL}}(p \parallel q) = \log \frac{D - C}{B - A}$$

Intuitively,[32] the information gain to a k times narrower uniform distribution contains $\log_2 k$ bits. This connects with the use of bits in computing, where $\log_2 k$ bits would be needed to identify one element of a k long stream.

Relation to metrics

While relative entropy is a statistical distance, it is not a metric on the space of probability distributions, but instead it is a divergence.[4] While metrics are symmetric and generalize *linear* distance, satisfying the triangle inequality, divergences are asymmetric in general and generalize *squared* distance, in some cases satisfying a generalized Pythagorean theorem. In general $D_{\mathrm{KL}}(P \parallel Q)$ does not equal $D_{\mathrm{KL}}(Q \parallel P)$, and while this can be symmetrized (see § Symmetrised divergence), the asymmetry is an important part of the geometry.[4]

It generates a topology on the space of probability distributions. More concretely, if $\{P_1, P_2, \ldots\}$ is a sequence of distributions such that

$$\lim_{n \to \infty} D_{\mathrm{KL}}(P_n \parallel Q) = 0,$$

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 θ. Consider then two close by values of $P = P(\theta)$ and $Q = P(\theta_0)$ so that the parameter θ differs by only a small amount from the parameter value θ_0. Specifically, up to first order one has (using the Einstein summation convention)

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

with $\Delta\theta_j = (\theta - \theta_0)_j$ a small change of θ in the j direction, and $P_j(\theta_0) = \dfrac{\partial P}{\partial \theta_j}(\theta_0)$ 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

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

and by the Taylor expansion one has up to second order

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

where the Hessian matrix of the divergence

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

must be positive semidefinite. Letting θ_0 vary (and dropping the subindex 0) the Hessian $g_{jk}(\theta)$ defines a (possibly degenerate) Riemannian metric on the θ 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 ξ is independent of ρ

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

then:

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

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.

MAUVE Metric

MAUVE is a measure of the statistical gap between two text distributions, such as the difference between text generated by a model and human-written text. This measure is computed using Kullback-Leibler divergences between the two distributions in a quantized embedding space of a foundation model.

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

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

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,

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

is the relative entropy of the joint probability distribution $P(X,Y)$ from the product $P(X)P(Y)$ of the two marginal probability distributions — 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,

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

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 $P(X)$. This definition of Shannon entropy forms the basis of E.T. Jaynes's alternative generalization to continuous distributions, the limiting density of discrete points (as opposed to the usual differential entropy), which defines the continuous entropy as

$$\lim_{N\to\infty} H_N(X) = \log(N) - \int p(x)\log\frac{p(x)}{m(x)}\,dx,$$

which is equivalent to:

$$\log(N) - D_{\mathrm{KL}}(p(x)||m(x))$$

Conditional entropy

The conditional entropy[33],

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

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 $\mathrm{H}(p)$). 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.

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

For explicit derivation of this, see the Motivation section above.

Under this scenario, relative entropies (kl-divergence) can be interpreted as the extra number of bits, on average, that are needed (beyond $\mathrm{H}(p)$) 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: $p(x) \to p(x \mid I)$. If some new fact $Y = y$ is discovered, it can be used to update the posterior distribution for X from $p(x \mid I)$ to a new posterior distribution $p(x \mid y, I)$ using Bayes' theorem:

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

This distribution has a new entropy:

$$\mathrm{H}\big(p(x \mid y, I)\big) = -\sum_x p(x \mid y, I) \log p(x \mid y, I),$$

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

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

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

If a further piece of data, $Y_2 = y_2$, subsequently comes in, the probability distribution for x can be updated further, to give a new best guess $p(x \mid y_1, y_2, I)$. If one reinvestigates the information gain for using $p(x \mid y_1, I)$ rather than $p(x \mid I)$, it turns out that it may be either greater or less than previously estimated:

$$\sum_{x} p(x \mid y_1, y_2, I) \log\left(\frac{p(x \mid y_1, y_2, I)}{p(x \mid I)}\right) \text{ may be } \leq \text{ or } > \text{ than } \sum_{x} p(x \mid y_1, I) \log\left(\frac{p(x \mid y_1, I)}{p(x \mid I)}\right)$$

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

$$D_{\mathrm{KL}}\left(p(x \mid y_1, y_2, I) \parallel p(x \mid I)\right) \text{ may be } <, = \text{ or } > \text{ than } D_{\mathrm{KL}}\left(p(x \mid y_1, y_2, I) \parallel p(x \mid y_1, I)\right) + D_{\mathrm{KL}}\left(p(x \mid y_1, I) \parallel p(x \mid I)\right)$$

All one can say is that on *average*, averaging using $p(y_2 \mid y_1, x, I)$, 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.[34] When posteriors are approximated to be Gaussian distributions, a design maximising the expected relative entropy is called Bayes d-optimal.

Discrimination information

Relative entropy $D_{\mathrm{KL}}\left(p(x \mid H_1) \parallel p(x \mid H_0)\right)$ can also be interpreted as the expected **discrimination information** for H_1 over H_0: the mean information per sample for discriminating in favor of a hypothesis H_1 against a hypothesis H_0, when hypothesis H_1 is true.[35] Another name for this quantity, given to it by I. J. Good, is the expected weight of evidence for H_1 over H_0 to be expected from each sample.

The expected weight of evidence for H_1 over H_0 is **not** the same as the information gain expected per sample about the probability distribution $p(H)$ of the hypotheses,

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

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 f should be chosen which is as hard to discriminate from the original distribution f_0 as possible; so that the new data produces as small an information gain $D_{\mathrm{KL}}(f \parallel f_0)$ as possible.

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

$$D_{\mathrm{KL}}(q(x \mid a)u(a) \parallel p(x,a)) = \mathrm{E}_{u(a)}\{D_{\mathrm{KL}}(q(x \mid a) \parallel p(x \mid a))\} + D_{\mathrm{KL}}(u(a) \parallel p(a)),$$

i.e. the sum of the relative entropy of $p(a)$ the prior distribution for a from the updated distribution $u(a)$, plus the expected value (using the probability distribution $u(a)$) of the relative entropy of the prior conditional distribution $p(x \mid a)$ from the new conditional distribution $q(x \mid a)$. (Note that often the later expected value is called the *conditional relative entropy* (or *conditional Kullback–Leibler divergence*) and denoted by $D_{\mathrm{KL}}(q(x \mid a) \parallel p(x \mid a))$[3][33]) This is minimized if $q(x \mid a) = p(x \mid a)$ over the whole support of $u(a)$; and we note that this result incorporates Bayes' theorem, if the new distribution $u(a)$ is in fact a δ function representing certainty that 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.

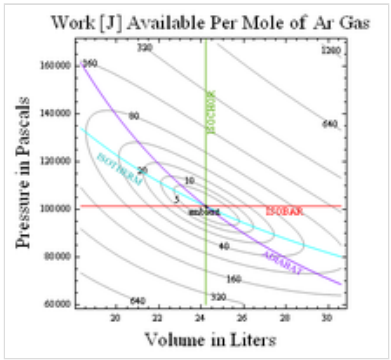
In the engineering literature, MDI is sometimes called the **Principle of Minimum Cross-Entropy** (MCE) or **Minxent** for short. Minimising relative entropy from m to p with respect to m is equivalent to minimizing the cross-entropy of p and m, since

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

which is appropriate if one is trying to choose an adequate approximation to p. However, this is just as often *not* the task one is trying to achieve. Instead, just as often it is m that is some fixed prior reference measure, and p that one is attempting to optimise by minimising $D_{\mathrm{KL}}(p \parallel m)$ 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 $D_{\mathrm{KL}}(p \parallel m)$, rather than $\mathrm{H}(p,m)$.

Relationship to available work

Surprisals[36] add where probabilities multiply. The surprisal for an event of probability p is defined as $s = k\ln(1/p)$. If k is $\{1, 1/\ln 2, 1.38 \times 10^{-23}\}$ then surprisal is in {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.



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

Best-guess states (e.g. for atoms in a gas) are inferred by maximizing the *average surprisal* S (entropy) for a given set of control parameters (like pressure P or volume V). This constrained entropy maximization, both classically[37] and quantum mechanically,[38] minimizes Gibbs availability in entropy units[39] $A \equiv -k\ln(Z)$ 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 and S is entropy) 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[40] 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.[41] 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

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

In quantum information science the minimum of $D_{\mathrm{KL}}(P \parallel Q)$ 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[42] and a book[43] 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.

Symmetrised divergence

Kullback & Leibler (1951) also considered the symmetrized function:[6]

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

which they referred to as the "divergence", though today the "KL divergence" refers to the asymmetric function (see § Etymology for the evolution of the term). This function is symmetric and nonnegative, and had already been defined and used by Harold Jeffreys in 1948;[7] it is accordingly called the **Jeffreys divergence**.

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 λ-divergence,

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

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 λ and $1-\lambda$ respectively.

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

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

where M is the average of the two distributions,

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

We can also interpret D_{JS} 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.[44][45]

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:

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

  Pinsker's inequality is vacuous for any distributions where $D_{\mathrm{KL}}(P \parallel Q) > 2$, since the total variation distance is at most 1. For such distributions, an alternative bound can be used, due to Bretagnolle and Huber[46] (see, also, Tsybakov[47]):

  $$\delta(P, Q) \leq \sqrt{1 - e^{-D_{\mathrm{KL}}(P \| Q)}}.$$

- The family of Rényi divergences generalize relative entropy. Depending on the value of a certain parameter, α, 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*.[48]

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
- Bhattacharyya distance

References

1. Csiszar, I (February 1975). "I-Divergence Geometry of Probability Distributions and Minimization Problems" (https://doi.org/10.1214%2Faop%2F1176996454). *Ann. Probab*. **3** (1): 146–158. doi:10.1214/aop/1176996454 (https://doi.org/10.1214%2Faop%2F1176996454).
2. 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://mathscinet.ams.org/mathscinet-getitem?mr=0039968).
3. Kullback 1959.
4. Amari 2016, p. 11.
5. Amari 2016, p. 28.
6. Kullback & Leibler 1951, p. 80.
7. Jeffreys 1948, p. 158.
8. Kullback 1959, p. 7.
9. 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).
10. Kullback 1959, p. 6.
11. MacKay, David J.C. (2003). *Information Theory, Inference, and Learning Algorithms* (https://books.google.com/books?id=AKuMj4PN_EMC) (1st ed.). Cambridge University Press. p. 34. ISBN 9780521642989 – via Google Books.
12. "What's the maximum value of Kullback-Leibler (KL) divergence?" (https://stats.stackexchange.com/q/351947). Machine learning. *Statistics Stack Exchange (stats.stackexchange.com)*. Cross validated.
13. "In what situations is the integral equal to infinity?" (https://math.stackexchange.com/q/20961). Integration. *Mathematics Stack Exchange (math.stackexchange.com)*.
14. Bishop, Christopher M. *Pattern recognition and machine learning* (http://worldcat.org/oclc/1334664824). p. 55. OCLC 1334664824 (https://search.worldcat.org/oclc/1334664824).
15. Kullback 1959, p. 5.
16. 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.
17. Abdulkadirov, Ruslan; Lyakhov, Pavel; Nagornov, Nikolay (January 2023). "Survey of Optimization Algorithms in Modern Neural Networks" (https://doi.org/10.3390%2Fmath11112466). *Mathematics*. **11** (11): 2466. doi:10.3390/math11112466 (https://doi.org/10.3390%2Fmath11112466). ISSN 2227-7390 (https://search.worldcat.org/issn/2227-7390).
18. Matassa, Marco (December 2021). "Fubini-Study metrics and Levi-Civita connections on quantum projective spaces" (https://linkinghub.elsevier.com/retrieve/pii/S0001870821005405). *Advances in Mathematics*. **393**: 108101. arXiv:2010.03291 (https://arxiv.org/abs/2010.03291). doi:10.1016/j.aim.2021.108101 (https://doi.org/10.1016%2Fj.aim.2021.108101). ISSN 0001-8708 (https://search.worldcat.org/issn/0001-8708).
19. Lan, Guanghui (March 2023). "Policy mirror descent for reinforcement learning: linear convergence, new sampling complexity, and generalized problem classes" (https://link.springer.com/article/10.1007/s10107-022-01816-5). *Mathematical Programming*. **198** (1): 1059–1106. doi:10.1007/s10107-022-01816-5 (https://doi.org/10.1007%2Fs10107-022-01816-5). ISSN 1436-4646 (https://search.worldcat.org/issn/1436-4646).
20. Kelly, J. L. Jr. (1956). "A New Interpretation of Information Rate". *Bell Syst. Tech. J*. **2** (4): 917–926. doi:10.1002/j.1538-7305.1956.tb03809.x (https://doi.org/10.1002%2Fj.1538-7305.1956.tb03809.x).
21. Soklakov, A. N. (2020). "Economics of Disagreement—Financial Intuition for the Rényi Divergence" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7517462). *Entropy*. **22** (8): 860. arXiv:1811.08308 (https://arxiv.org/abs/1811.08308). Bibcode:2020Entrp..22..860S (https://ui.adsabs.harvard.edu/abs/2020Entrp..22..860S). doi:10.3390/e22080860 (https://doi.org/10.3390%2Fe22080860). PMC 7517462 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7517462). PMID 33286632 (https://pubmed.ncbi.nlm.nih.gov/33286632).

22. Soklakov, A. N. (2023). "Information Geometry of Risks and Returns". *Risk*. **June**. SSRN 4134885 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4134885).
23. Henide, Karim (30 September 2024). "Flow Rider: Tradable Ecosystems' Relative Entropy of Flows As a Determinant of Relative Value". *The Journal of Investing*. **33** (6): 34–58. doi:10.3905/joi.2024.1.321 (https://doi.org/10.3905%2Fjoi.2024.1.321).
24. Sanov, I.N. (1957). "On the probability of large deviations of random magnitudes". *Mat. Sbornik*. **42** (84): 11–44.
25. Novak S.Y. (2011), *Extreme Value Methods with Applications to Finance* ch. 14.5 (Chapman & Hall). ISBN 978-1-4398-3574-6.
26. Hobson, Arthur (1971). *Concepts in statistical mechanics*. New York: Gordon and Breach. ISBN 978-0677032405.
27. Bonnici, V. (2020). "Kullback-Leibler divergence between quantum distributions, and its upper-bound". arXiv:2008.05932 (https://arxiv.org/abs/2008.05932) [cs.LG (https://arxiv.org/archive/cs.LG)].
28. See the section "differential entropy – 4" in Relative Entropy (http://videolectures.net/nips09_verdu_re/) video lecture by Sergio Verdú NIPS 2009
29. Donsker, Monroe D.; Varadhan, SR Srinivasa (1983). "Asymptotic evaluation of certain Markov process expectations for large time. IV". *Communications on Pure and Applied Mathematics*. **36** (2): 183–212. doi:10.1002/cpa.3160360204 (https://doi.org/10.1002%2Fcpa.3160360204).
30. Duchi J. "Derivations for Linear Algebra and Optimization" (https://web.stanford.edu/~jduchi/projects/general_notes.pdf) (PDF). p. 13.
31. Belov, Dmitry I.; Armstrong, Ronald D. (2011-04-15). "Distributions of the Kullback-Leibler divergence with applications" (https://dx.doi.org/10.1348/000711010x522227). *British Journal of Mathematical and Statistical Psychology*. **64** (2): 291–309. doi:10.1348/000711010x522227 (https://doi.org/10.1348%2F000711010x522227). ISSN 0007-1102 (https://search.worldcat.org/issn/0007-1102). PMID 21492134 (https://pubmed.ncbi.nlm.nih.gov/21492134).
32. Buchner, Johannes (2022-04-29). *An intuition for physicists: information gain from experiments* (http://worldcat.org/oclc/1363563215). OCLC 1363563215 (https://search.worldcat.org/oclc/1363563215).
33. Cover, Thomas M.; Thomas, Joy A. (1991), *Elements of Information Theory*, John Wiley & Sons, p. 22
34. 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). hdl:11299/199630 (https://hdl.handle.net/11299%2F199630).
35. 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.
36. Tribus, Myron (1959). *Thermostatics and Thermodynamics: An Introduction to Energy, Information and States of Matter, with Engineering Applications* (https://books.google.com/books?id=eyrYrQEACAAJ). Van Nostrand.
37. 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). S2CID 17870175 (https://api.semanticscholar.org/CorpusID:17870175).
38. 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).
39. Gibbs, Josiah Willard (1871). *A Method of Geometrical Representation of the Thermodynamic Properties of Substances by Means of Surfaces* (https://books.google.com/books?id=6ijzXwAACAAJ). The Academy. footnote page 52.
40. 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).
41. 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.
42. 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).
43. Burnham, Kenneth P. (December 2010). *Model selection and multimodel inference : a practical information-theoretic approach* (http://worldcat.org/oclc/878132909). Springer. ISBN 978-1-4419-2973-0. OCLC 878132909 (https://search.worldcat.org/oclc/878132909).

44. Nielsen, Frank (2019). "On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7514974). *Entropy*. **21** (5): 485. arXiv:1904.04017 (https://arxiv.org/abs/1904.04017). Bibcode:2019Entrp..21..485N (https://ui.adsabs.harvard.edu/abs/2019Entrp..21..485N). doi:10.3390/e21050485 (https://doi.org/10.3390%2Fe21050485). PMC 7514974 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7514974). PMID 33267199 (https://pubmed.ncbi.nlm.nih.gov/33267199).
45. Nielsen, Frank (2020). "On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7516653). *Entropy*. **22** (2): 221. arXiv:1912.00610 (https://arxiv.org/abs/1912.00610). Bibcode:2020Entrp..22..221N (https://ui.adsabs.harvard.edu/abs/2020Entrp..22..221N). doi:10.3390/e22020221 (https://doi.org/10.3390%2Fe22020221). PMC 7516653 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7516653). PMID 33285995 (https://pubmed.ncbi.nlm.nih.gov/33285995).
46. Bretagnolle, J.; Huber, C. (1978), "Estimation des densités : Risque minimax", *Séminaire de Probabilités XII* (https://dx.doi.org/10.1007/bfb0064610), Lecture Notes in Mathematics (in French), vol. 649, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 342–363, doi:10.1007/bfb0064610 (https://doi.org/10.1007%2Fbfb0064610), ISBN 978-3-540-08761-8, S2CID 122597694 (https://api.semanticscholar.org/CorpusID:122597694), retrieved 2023-02-14 Lemma 2.1
47. B.), Tsybakov, A. B. (Alexandre (2010). *Introduction to nonparametric estimation* (http://worldcat.org/oclc/757859245). Springer. ISBN 978-1-4419-2709-5. OCLC 757859245 (https://search.worldcat.org/oclc/757859245). Equation 2.25.
48. 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).

- Amari, Shun-ichi (2016). *Information Geometry and Its Applications*. Applied Mathematical Sciences. Vol. 194. Springer Japan. pp. XIII, 374. doi:10.1007/978-4-431-55978-8 (https://doi.org/10.1007%2F978-4-431-55978-8). ISBN 978-4-431-55977-1.
- Kullback, Solomon (1959), *Information Theory and Statistics*, John Wiley & Sons. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.
- Jeffreys, Harold (1948). *Theory of Probability* (Second ed.). Oxford University Press.

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/), NIPS 2009. One-hour video lecture.
- A modern summary of info-theoretic divergence measures (https://arxiv.org/abs/math/0604246)