Kullback–Leibler divergence

Not to be confused with divergence in calculus.

In probability theory and information theory, the **Kullback–Leibler divergence**,[1][2] also called **information divergence**, **information gain**, **relative entropy**, **KLIC**, or **KL divergence**, is a measure (but not a metric) of the non-symmetric difference between two probability distributions *P* and *Q*. The Kullback–Leibler divergence was originally introduced by Solomon Kullback and Richard Leibler in 1951 as the **directed divergence** between two distributions; Kullback himself preferred the name **discrimination information**.[3] It is discussed in Kullback's historic text, *Information Theory and Statistics*.[2]

Expressed in the language of Bayesian inference, the Kullback–Leibler divergence from *Q* to *P*, denoted *D*KL(*P*‖*Q*), is a measure of the information gained when one revises 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*.[4] 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*.

The Kullback–Leibler divergence is a special case of a broader class of 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*KL(*P*‖*Q*) does not equal *D*KL(*Q*‖*P*). However, its infinitesimal form, specifically its Hessian, gives a metric tensor known as the Fisher information metric.

In the context of machine learning, the Kullback–Leibler divergence is often called the information gain achieved if *P* is used instead of *Q*. By analogy with information theory, it is also called the **relative entropy** of *P* with respect to *Q*. In the context of coding theory, Kullback–Leibler divergence can be construed as measure 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*.

1 Definition

For discrete probability distributions *P* and *Q*, the Kullback–Leibler divergence from *Q* to *P* is defined[5] to be

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}.$$

In words, it is the expectation of the logarithmic difference between the probabilities *P* and *Q*, where the expectation is taken using the probabilities *P*. The Kullback–Leibler divergence is defined only if *Q*(*i*)=0 implies *P*(*i*)=0, for all *i* (absolute continuity). Whenever *P*(*i*) is zero the contribution of the *i*-th term is interpreted as zero because $\lim_{x\to 0} x\log(x) = 0$.

For distributions *P* and *Q* of a continuous random variable, the Kullback–Leibler divergence is defined to be the integral:[6]

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

where *p* and *q* denote the 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 Kullback–Leibler divergence from *Q* to *P* is defined as

$$D_{\mathrm{KL}}(P\|Q) = \int_X \log\frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}P,$$

where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon–Nikodym derivative of *P* with respect to *Q*, and provided the expression on the right-hand side exists. Equivalently, this can be written as

$$D_{\mathrm{KL}}(P\|Q) = \int_X \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q,$$

which we recognize as the entropy of *P* relative to *Q*. Continuing in this case, if μ is any measure on *X* for which $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ exist (meaning that *p* and *q* are absolutely continuous with respect to μ), then the Kullback–Leibler divergence from *Q* to *P* is given as

$$D_{\mathrm{KL}}(P\|Q) = \int_X p\,\log\frac{p}{q}\,\mathrm{d}\mu.$$

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 the Kullback–Leibler divergence hold regardless of the base of the logarithm.

Various conventions exist for referring to *D*KL(*P*‖*Q*) in words. Often it is referred to as the divergence *between* *P* and *Q*; however this fails to convey the fundamental asymmetry in the relation. Sometimes, as in this article, it may be found described as the divergence of *P* from, or with respect to *Q*. This reflects the asymmetry in Bayesian inference, which starts *from* a prior *Q* and updates *to* the posterior *P*.

2 Characterization

Arthur Hobson proved that the Kullback–Leibler divergence 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.[7] 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.

There is also a Bayesian characterization of the Kullback–Leibler divergence.[8]

3 Motivation

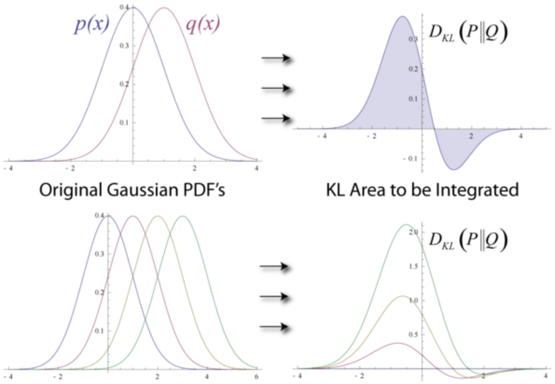


Illustration of the Kullback–Leibler (KL) divergence for two normal Gaussian distributions. Note the typical asymmetry for the Kullback–Leibler divergence is clearly visible.

In information theory, the Kraft–McMillan theorem establishes that any directly decodable coding scheme for coding a message to identify one value *xi* out of a set of possibilities *X* can be seen as representing an implicit probability distribution *q*(*xi*)=2^{-li} over *X*, where *li* is the length of the code for *xi* in bits. Therefore, the Kullback–Leibler divergence 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*.

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &= -\sum_x p(x)\log q(x) &+& \sum_x p(x)\log p(x) \\ &= H(P,Q) &-& H(P) \end{aligned}$$

where *H*(*P*,*Q*) is the cross entropy of *P* and *Q*, and *H*(*P*) is the entropy of *P*.

Note also that there is a relation between the Kullback–Leibler divergence and the "rate function" in the theory of large deviations.[9][10]

4 Properties

- The Kullback–Leibler divergence is always non-negative,

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

a result known as Gibbs' inequality, with *D*KL(*P*‖*Q*) 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*.

- The Kullback–Leibler divergence 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 Kullback–Leibler divergence may be rewritten:

$$D_{\mathrm{KL}}(P\|Q) = \int_{x_a}^{x_b} P(x)\log\left(\frac{P(x)}{Q(x)}\right)\,dx = \int_{y_a}^{y_b} P(y)\log\left(\frac{P(y)dy/dx}{Q(y)dy/dx}\right.$$

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 Kullback–Leibler divergence 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[11] (such as self-information or Shannon entropy), which can become undefined or negative for non-discrete probabilities.

- The Kullback–Leibler divergence 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

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

5 Kullback–Leibler divergence for multivariate normal distributions

Suppose that we have two multivariate normal distributions, with means μ_0, μ_1 and with (nonsingular) covariance matrices Σ_0, Σ_1 . If the two distributions have the same dimension, *k*, then the Kullback–Leibler divergence between the distributions is as follows.[12]

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

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 log*e* 2 yields the divergence in bits.

6 Relation to metrics

One might be tempted to call the Kullback–Leibler divergence a "distance metric" on the space of probability distributions, but this would not be correct as it is not symmetric – that is, $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$, – nor does it satisfy the triangle inequality. Even so, being a premetric, it generates a topology on the space of probability distributions. More concretely, if $\{P_1, P_2, \cdots\}$ is a sequence of distributions such that

then it is said that

Pinsker's inequality entails that

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

Following Rényi (1970, 1961)[13][14]

6.1 Fisher information metric

However, the Kullback–Leibler divergence is rather directly related to a metric, specifically, 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 $P_j(\theta_0) = \frac{\partial P}{\partial \theta^j}(\theta_0)$ the corresponding rate of change in the probability distribution. Since the Kullback–Leibler divergence 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

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

and by the Taylor expansion one has up to second order

$$D_{\mathrm{KL}}(P(\theta)\|P(\theta_0)) = \frac{1}{2}\Delta\theta^j\Delta\theta^k g_{jk}(\theta_0) + \cdots$$

where the Hessian matrix of the divergence

$$g_{jk}(\theta_0) = \left.\frac{\partial^2}{\partial \theta^j \partial \theta^k}\right|_{\theta=\theta_0} D_{KL}(P(\theta)\|P(\theta_0))$$

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.

7 Relation to other quantities of information theory

Many of the other quantities of information theory can be interpreted as applications of the Kullback–Leibler divergence to specific cases.

The self-information,

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

is the Kullback–Leibler divergence 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*.

The mutual information,

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

is the Kullback–Leibler divergence 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.

The Shannon entropy,

$$\begin{aligned} H(X) &= \text{(i) } \mathrm{E}_x\{I(x)\} \\ &= \text{(ii) } \log N - D_{\mathrm{KL}}(P(X)\|P_U(X)) \end{aligned}$$

is the number of bits which would have to be transmitted to identify *X* from *N* equally likely possibilities, *less* the Kullback–Leibler divergence of the uniform distribution *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*).

The conditional entropy,

$$\begin{aligned} H(X \mid Y) &= \log N - D_{\mathrm{KL}}(P(X,Y)\|P_U(X)P(Y)) \\ &= \text{(i) } \log N - D_{\mathrm{KL}}(P(X,Y)\|P(X)P(Y)) - D_{\mathrm{KL}}(P(X)\|P_U(X)) \\ &= H(X) - I(X;Y) \\ &= \text{(ii) } \log N - \mathrm{E}_Y\{D_{\mathrm{KL}}(P(X|Y)\|P_U(X))\} \end{aligned}$$

is the number of bits which would have to be transmitted to identify *X* from *N* equally likely possibilities, *less* the Kullback–Leibler divergence 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.

The cross entropy between two probability distributions 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:

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

8 Kullback–Leibler divergence and Bayesian updating

In Bayesian statistics the Kullback–Leibler divergence 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:

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

...which may be less than or greater than the original entropy $H(p(\cdot \mid I))$. 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:

$$D_{\mathrm{KL}}\big(p(\cdot \mid y, I) \mid p(\cdot \mid I)\big) = \sum_x p(x \mid y, I) \log \frac{p(x \mid y, I)}{p(x \mid I)}$$

to the message length. This therefore represents the amount of useful information, or information gain, about X , that we can estimate 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:

$\sum_x p(x \mid y_1, y_2, I) \log \frac{p(x|y_1,y_2,I)}{p(x|I)}$ may be ≤ or > than $\sum_x p(x \mid y_1, I) \log \dfrac{p(x \mid y_1, I)}{p(x \mid I)}$

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

$D_{\mathrm{KL}}\big(p(\cdot \mid y_1, y_2, I)\big\|p(\cdot \mid I)\big)$ may be <, = or > than $D_{\mathrm{KL}}\big(p(\cdot \mid y_1, y_2, I)\big\|p(\cdot|y_1, I)\big) +$ $D_{\mathrm{KL}}\big(p(\cdot \mid y_1, I)\big\|p(\cdot \mid I)\big)$

All one can say is that on *average*, averaging using $p(y_2 \mid y_1, x, I)$, the two sides will average out.

8.1 Bayesian experimental design

A common goal in Bayesian experimental design is to maximise the expected Kullback–Leibler divergence between the prior and the posterior.[15] When posteriors are approximated to be Gaussian distributions, a design maximising the expected Kullback–Leibler divergence is called Bayes d-optimal.

9 Discrimination information

The Kullback–Leibler divergence *D*KL(*p*(*x*|H_1) ∥ *p*(*x*|H_0)) 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.[16] 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,

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

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.

9.1 Principle of minimum discrimination information

The idea of Kullback–Leibler divergence 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*KL(f ∥ 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*), the Kullback–Leibler divergence between the new joint distribution for *x* and *a*, *q*(*x*|*a*) *u*(*a*), and the earlier prior distribution would be:

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

i.e. the sum of the Kullback–Leibler divergence 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 Kullback–Leibler divergence of the prior conditional distribution *p*(*x*|*a*) from the new conditional distribution *q*(*x*|*a*). (Note that often the later expected value is called the *conditional Kullback–Leibler divergence* (or *conditional relative entropy*) and denoted by *D*KL(*q*(*x*|*a*)∥*p*(*x*|*a*))[17]) This is minimized if *q*(*x*|*a*) = *p*(*x*|*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.

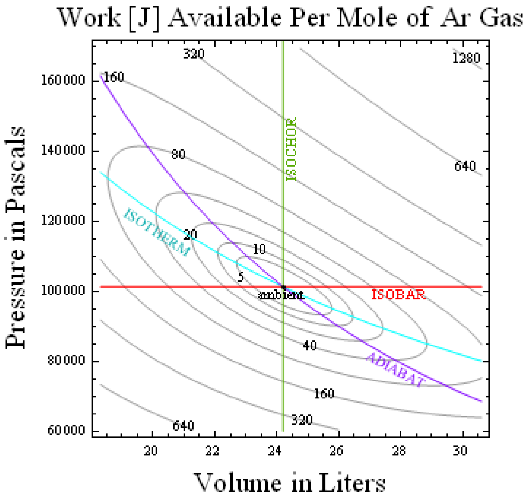
MDI can be seen as an extension of Laplace's Principle of Insufficient Reason, and the Principle of Maximum Entropy of E.T. Jaynes. In particular, it is the natural extension of the principle of maximum entropy from discrete to continuous distributions, for which Shannon entropy ceases to be so useful (see *differential entropy*), but the Kullback–Leibler divergence 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 the Kullback–Leibler divergence from *m* to *p* with respect to *m* is equivalent to minimizing the cross-entropy of *p* and *m*, since

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

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 $D\mathrm{KL}(p\|m)$ 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 $D\mathrm{KL}(p\|m)$, rather than $H(p,m)$.

10 Relationship to available work



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.

Surprisals[18] 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.

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[19] and quantum mechanically,[20] minimizes Gibbs availability in entropy units[21] $A \equiv -k\ln Z$ 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[22] the work available relative to some ambient is obtained by multiplying ambient temperature T_o by Kullback–Leibler divergence 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 Kullback–Leibler divergence $\Delta I = Nk[\Theta(V/V_o) + \frac{3}{2}\Theta(T/T_o)]$. The resulting contours of constant Kullback–Leibler divergence, 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.[23] Thus Kullback–Leibler divergence measures thermodynamic availability in bits.

11 Quantum information theory

For density matrices P and Q on a Hilbert space, the K–L divergence (or quantum relative entropy as it is often called in this case) from Q to P is defined to be

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

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

12 Relationship between models and reality

Just as Kullback–Leibler divergence of "actual from ambient" measures thermodynamic availability, Kullback–Leibler divergence 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 Kullback–Leibler divergence 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[24] and a book[25] by Burnham and Anderson. In a nutshell the Kullback–Leibler divergence of reality from a model may be estimated, to within a constant additive term, by a function (like the squares summed) of the deviations observed between data and the model's predictions. 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 Kullback–Leibler divergence, such as maximum likelihood and maximum spacing estimators.

13 Symmetrised divergence

Kullback and Leibler themselves actually defined the divergence as:

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

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.

An alternative is given via the λ divergence,

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

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 − λ) respectively.

The value λ = 0.5 gives the Jensen–Shannon divergence, defined by

$$D_{\mathrm{JS}}=\tfrac{1}{2}D_{\mathrm{KL}}\left(P\|M\right)+\tfrac{1}{2}D_{\mathrm{KL}}\left(Q\|M\right)$$

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

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

*D*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 induces the same affine connection on a statistical manifold), and equal to one-half the so-called *Jeffreys divergence*.[26][27]

14 Relationship to other probability-distance measures

There are many other important measures of probability distance. Some of these are particularly connected with the Kullback–Leibler divergence. For example:

- The total variation distance, $\delta(p,q)$ This is connected to the divergence through Pinsker's inequality: $\delta(P,Q)\le\sqrt{\tfrac{1}{2}D_{\mathrm{KL}}(P\|Q)}$
- The family of Rényi divergences provide generalizations of the Kullback–Leibler divergence. 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*.[26]

15 Data differencing

Main article: 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).

16 See also

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

17 References

[1] Kullback, S.; Leibler, R.A. (1951). "On information and sufficiency". *Annals of Mathematical Statistics*. **22** (1): 79–86. doi:10.1214/aoms/1177729694. MR 39968.

[2] Kullback S. (1959), *Information Theory and Statistics* (John Wiley & Sons).

[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. JSTOR 2684769.

[4] Burnham K.P., Anderson D.R. (2002), *Model Selection and Multi-Model Inference* (Springer). (2nd edition), p.51

[5] MacKay, David J.C. (2003). *Information Theory, Inference, and Learning Algorithms* (First ed.). Cambridge University Press. p. 34.

[6] Bishop C. (2006). *Pattern Recognition and Machine Learning* p. 55.

[7] Hobson, Arthur (1971). *Concepts in statistical mechanics*. New York: Gordon and Breach. ISBN 0677032404.

[8] Baez, John; Fritz, Tobias (2014). “A Bayesian characterization of relative entropy”. *Theory and Application of Categories*. **29**: 421–456. arXiv:1402.3067.

[9] Sanov, I.N. (1957). “On the probability of large deviations of random magnitudes”. *Matem. Sbornik*. **42** (84): 11–44.

[10] Novak S.Y. (2011), *Extreme Value Methods with Applications to Finance* ch. 14.5 (Chapman & Hall). ISBN 978-1-4398-3574-6.

[11] See the section “differential entropy - 4” in Relative Entropy video lecture by Sergio Verdú NIPS 2009

[12] Duchi J., "Derivations for Linear Algebra and Optimization", p. 13.

[13] Rényi A. (1970). *Probability Theory*. Elsevier. Appendix, Sec.4. ISBN 0-486-45867-9.

[14] Rényi, A. (1961), “On measures of entropy and information” (PDF), *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*, pp. 547–561

[15] Chaloner, K.; Verdinelli, I. (1995). “Bayesian experimental design: a review”. *Statistical Science*. **10** (3): 273–304. doi:10.1214/ss/1177009939.

[16] Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. (2007). “Section 14.7.2. Kullback–Leibler Distance”. *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). Cambridge University Press. ISBN 978-0-521-88068-8

[17] Thomas M. Cover, Joy A. Thomas (1991) *Elements of Information Theory* (John Wiley & Sons), p.22

[18] Myron Tribus (1961), *Thermodynamics and Thermostatics* (D. Van Nostrand, New York)

[19] Jaynes, E. T. (1957). “Information theory and statistical mechanics” (PDF). *Physical Review*. **106**: 620–630. Bibcode:1957PhRv..106..620J. doi:10.1103/physrev.106.620.

[20] Jaynes, E. T. (1957). “Information theory and statistical mechanics II” (PDF). *Physical Review*. **108**: 171–190. Bibcode:1957PhRv..108..171J. doi:10.1103/physrev.108.171.

[21] 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.

[22] Tribus, M.; McIrvine, E. C. (1971). “Energy and information”. *Scientific American*. **224**: 179–186. doi:10.1038/scientificamerican0971-179.

[23] Fraundorf, P. (2007). “Thermal roots of correlation-based complexity”. *Complexity*. **13** (3): 18–26. doi:10.1002/cplx.20195.

[24] Burnham, K.P.; Anderson, D.R. (2001). “Kullback–Leibler information as a basis for strong inference in ecological studies”. *Wildlife Research*. **28**: 111–119. doi:10.1071/WR99107.

[25] 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.

[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.

[27] Jeffreys, H. (1946). “An invariant form for the prior probability in estimation problems”. *Proceedings of the Royal Society of London, Series A*. **186**: 453–461. Bibcode:1946RSPSA.186..453J. doi:10.1098/rspa.1946.0056. JSTOR 97883.

18 External links

- Information Theoretical Estimators Toolbox
- Ruby gem for calculating Kullback–Leibler divergence
- Jon Shlens’ tutorial on Kullback–Leibler divergence and likelihood theory
- Matlab code for calculating Kullback–Leibler divergence for discrete distributions
- Sergio Verdú, Relative Entropy, NIPS 2009. One-hour video lecture.
- A modern summary of info-theoretic divergence measures

19 Text and image sources, contributors, and licenses

19.1 Text

- **Kullback–Leibler divergence** *Source:* https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence?oldid=763217976 *Contributors:* Edward, Michael Hardy, Den fjättrade ankan~enwiki, Kevin Baas, Cyan, BenKovitz, Charles Matthews, Jitse Niesen, Mpost89, Benwing, Mottzo, Wile E. Heresiarch, Giftlite, Inkling, Romanpoet, Dmb000006, Stern~enwiki, Schizoid, MarkSweep, AndrewKeenan-Richardson, Bender235, MisterSheik, 3mta3, PAR, Jheald, Oleg Alexandrov, MartinSpacek, Linas, Shreevatsa, BlaiseFEgan, Rjwilmsi, John Baez, Mathbot, Jmorgan, Spacepotato, Nothing1212, Mike Lin, Gzabers, Avraham, Cedar101, MDReid, Mebden, SmackBot, Mmernex, Adfernandes, Eskimbot, Object01, Mcld, Ignacioerrico, Nbarth, Tsca.bot, Memming, Jon Awbrey, Dnavarro, Cronholm144, Nijdam, Dfass, Kyellan, Yoderj, A. Pichler, JForget, Thermochap, Physic sox, Sir Vicious, Winterfors, Neonleonb, MaxEnt, Mct mht, FilipeS, Amit Moscovich, Headbomb, Rkrish67, Stangaa, Magioladitis, RogierBrussee, STBot, Wullj, Andre.holzner, Smite-Meister, LordAnubisBOT, Epistemenical, Punkstar89, TXiKiBoT, Miranda, Mundhenk, Jamelan, Loniousmonk, Forwardmeasure, Brech~enwiki, Melcombe, Rinconsoleao, Wittnate, Sun Creator, Kaba3, Edg2103, Qwfp, DumZiBoT, Addbot, Deepmath, Fyrael, Wikomidia, Baisemain, Lightbot, Luckas-bot, Yobot, Legendre17, AnomieBOT, Materialscientist, □□, Obersachsebot, Xqbot, GrouchoBot, Nathanielvirgo, Chjoaygame, X7q, Citation bot 1, Gaba p, Kiefer.Wolfowitz, Stpasha, Angelorf, Amkilpatrick, RjwilmsiBot, Ereiniona, Kastchei, Cstahlhut, Slawekb, Quondum, ClueBot NG, Helpful Pixie Bot, Leopd, Bibcode Bot, Epomqo, BG19bot, Marcocapelle, SciCompTeacher, Manoguru, Francis liberty, M.daryalal, Iturrate, SFK2, Lordbloth, Szzoli, Austrartsua, Sunil.log, Zoltan szabo, Opencooper, Leegrc, Engheta, Velvel2, SolidPhase, Samthemonad, Fmadd, ChrisNeuro, Bender the Bot, Drum book and Anonymous: 142

19.2 Images

- **File:ArgonKLdivergence.png** *Source:* https://upload.wikimedia.org/wikipedia/commons/c/c2/ArgonKLdivergence.png *License:* GFDL *Contributors:* Own work *Original artist:* P. Fraundorf
- **File:Edit-clear.svg** *Source:* https://upload.wikimedia.org/wikipedia/en/f/f2/Edit-clear.svg *License:* Public domain *Contributors:* The *Tango! Desktop Project*. *Original artist:*

  The people from the Tango! project. And according to the meta-data in the file, specifically: “Andreas Nilsson, and Jakub Steiner (although minimally).”
- **File:KL-Gauss-Example.png** *Source:* https://upload.wikimedia.org/wikipedia/en/a/a8/KL-Gauss-Example.png *License:* CC-BY-SA-3.0 *Contributors:*

  T. Nathan Mundhenk, Ph.D thesis appendix C.

  Original artist:

  Mundhenk (talk)
- **File:Lock-green.svg** *Source:* https://upload.wikimedia.org/wikipedia/commons/6/65/Lock-green.svg *License:* CC0 *Contributors:* en:File:Free-to-read_lock_75.svg *Original artist:* User:Trappist the monk

19.3 Content license