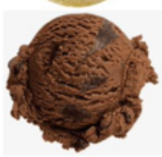
Jensen-Shannon Divergence

Three flavors of the Jensen-Shannon divergence

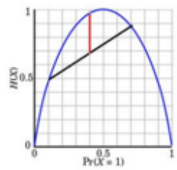
- **Symmetrization** of the relative entropy:

$$D_{\mathrm{JS}}[p,q] \;:=\; \frac{1}{2}\left(D_{\mathrm{KL}}\left[p:\frac{p+q}{2}\right]+D_{\mathrm{KL}}\left[q:\frac{p+q}{2}\right]\right)$$



- **Convexity gap** of Shannon negentropy:

$$D_{\mathrm{JS}}[p,q] \;:=\; h\left[\frac{p+q}{2}\right]-\frac{h[p]+h[q]}{2}$$



- **Diversity index** of two probability measures:

$$D_{\mathrm{JS}}[p,q] \;:=\; \min_{c\in\mathcal{D}}\frac{1}{2}\left(D_{\mathrm{KL}}[p:c]+D_{\mathrm{KL}}[q:c]\right)$$

arXiv:2102.09728

In probability theory and statistics, the **Jensen–Shannon divergence** is a method of measuring the similarity between two probability distributions. It is also known as **information radius** (**IRad**)[1][2] or **total divergence to the average**.[3] It is based on the Kullback–Leibler divergence, with some notable (and useful) differences, including that it is symmetric and it always has a finite value. The square root of the Jensen–Shannon divergence is a metric often referred to as Jensen–Shannon distance.[4][5][6]

Definition [edit]

Consider the set $M_+^1(A)$ of probability distributions where A is a set provided with some σ-algebra of measurable subsets. In particular we can take A to be a finite or countable set with all subsets being measurable.

The Jensen–Shannon divergence (JSD) is a symmetrized and smoothed version of the Kullback–Leibler divergence $D(P \parallel Q)$. It is defined by

$$\mathrm{JSD}(P \parallel Q)=\frac{1}{2}D(P \parallel M)+\frac{1}{2}D(Q \parallel M),$$

where $M=\frac{1}{2}(P+Q)$ is a mixture distribution of P and Q.

The geometric Jensen–Shannon divergence[7] (or G-Jensen–Shannon divergence) yields a closed-form formula for divergence between two Gaussian distributions by taking the geometric mean.

A more general definition, allowing for the comparison of more than two probability distributions, is:

$$\begin{aligned}\mathrm{JSD}_{\pi_1,\ldots,\pi_n}(P_1,P_2,\ldots,P_n) &= \sum_i \pi_i D(P_i \parallel M)\\ &= H(M)-\sum_{i=1}^{n}\pi_i H(P_i)\end{aligned}$$

How to define it: take half of P relative to Q, take half of Q relative to P, you get the metric that is a mixture distribution.

Relation to mutual information [edit]

The Jensen–Shannon divergence is the mutual information between a random variable X associated to a mixture distribution between P and Q and the binary indicator variable Z that is used to switch between P and Q to produce the mixture. Let X be some abstract function on the underlying set of events that discriminates well between events, and choose the value of X according to P if $Z=0$ and according to Q if $Z=1$, where Z is equiprobable. That is, we are choosing X according to the probability measure $M=(P+Q)/2$, and its distribution is the mixture distribution. We compute

$$\begin{aligned}
I(X;Z) &= H(X) - H(X|Z)\\
&= -\sum M\log M + \frac{1}{2}\left[\sum P\log P + \sum Q\log Q\right]\\
&= -\sum \frac{P}{2}\log M - \sum \frac{Q}{2}\log M + \frac{1}{2}\left[\sum P\log P + \sum Q\log Q\right]\\
&= \frac{1}{2}\sum P\left(\log P - \log M\right) + \frac{1}{2}\sum Q\left(\log Q - \log M\right)\\
&= \mathrm{JSD}(P \parallel Q)
\end{aligned}$$

It follows from the above result that the Jensen–Shannon divergence is bounded by 0 and 1 because mutual information is non-negative and bounded by $H(Z)=1$ in base 2 logarithm.

One can apply the same principle to a joint distribution and the product of its two marginal distribution (in analogy to Kullback–Leibler divergence and mutual information) and to measure how reliably one can decide if a given response comes from the joint distribution or the product distribution—subject to the assumption that these are the only two possibilities.[9]

Jensen–Shannon centroid [edit]

The centroid C* of a finite set of probability distributions can be defined as the minimizer of the average sum of the Jensen-Shannon divergences between a probability distribution and the prescribed set of distributions:

$$C^* = \arg\min_Q \sum_{i=1}^{n} \mathrm{JSD}(P_i \parallel Q)$$

An efficient algorithm[16] (CCCP) based on difference of convex functions is reported to calculate the Jensen-Shannon centroid of a set of discrete distributions (histograms).

Applications [edit]

The Jensen–Shannon divergence has been applied in bioinformatics and genome comparison,[17][18] in protein surface comparison,[19] in the social sciences,[20] in the quantitative study of history,[21] in fire experiments,[22] and in machine learning.[23]

1.1. Kullback–Leibler Divergence and Its Symmetrizations

Let $(\mathscr{X}, \mathscr{A})$ be a measurable space [1] where $\mathscr{X}$ denotes the sample space and $\mathscr{A}$ the σ-algebra of measurable events. Consider a positive measure μ (usually the Lebesgue measure μ_L with Borel σ-algebra $\mathscr{B}(\mathbb{R}^d)$ or the counting measure μ_c with power set σ-algebra $2^{\mathscr{X}}$). Denote by $\mathscr{P}$ the set of probability distributions.

The Kullback–Leibler Divergence [2] (KLD) $\mathrm{KL}: \mathscr{P} \times \mathscr{P} \rightarrow [0, \infty]$ is the most fundamental distance [2] between probability distributions, defined by:

$$\mathrm{KL}(P:Q) := \int p \log \frac{p}{q} \mathrm{d}\mu, \tag{1}$$

The KLD is also called the relative entropy [2] because it can be written as the difference of the cross-entropy minus the entropy:

$$\mathrm{KL}(p:q) = h_{\times}(p:q) - h(p), \tag{2}$$

where $h_{\times}$ denotes the cross-entropy [2]:

$$h_{\times}(p:q) := \int p \log \frac{1}{q} \mathrm{d}\mu, \tag{3}$$

and

$$h(p) := \int p \log \frac{1}{p} \mathrm{d}\mu = h_{\times}(p:p),$$

$$\mathrm{KL}^{*}(P:Q) := \mathrm{KL}(Q:P) = \int q \log \frac{q}{p} \mathrm{d}\mu. \tag{5}$$

In general, the *reverse distance* or *dual distance* for a distance D is written as:

$$D^{*}(p:q) := D(q:p). \tag{6}$$

One way to symmetrize the KLD is to consider the *Jeffreys Divergence* [4] (JD, Sir Harold Jeffreys (1891–1989) was a British statistician.):

$$J(p;q) := \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = \int (p-q) \log \frac{p}{q} \mathrm{d}\mu = J(q;p). \tag{7}$$

Jefferey's divergence do not take half.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7514974/

Summary of Distances and Their Notations.

Weighted mean	$M_{\alpha}, \alpha \in (0,1)$
Arithmetic mean	$A_{\alpha}(x,y) = (1-\alpha)x + \alpha y$
Geometric mean	$G_{\alpha}(x,y) = x^{1-\alpha} y^{\alpha}$
Harmonic mean	$H_{\alpha}(x,y) = \frac{xy}{(1-\alpha)y + \alpha x}$
Power mean	$P_{\alpha}^{p}(x,y) = ((1-\alpha)x^{p} + \alpha y^{p})^{\frac{1}{p}}, \quad p \in \mathbb{R} \setminus \{0\}$, $\lim_{p\to 0} P_{\alpha}^{p} = G$
Quasi-arithmetic mean	$M_{\alpha}^{f}(x,y) = f^{-1}((1-\alpha)f(x) + \alpha f(y))$, f strictly monotonous
	$Z_{\alpha}^{M}(p,q) = \int_{t\in\mathcal{X}} M_{\alpha}(p(t), q(t))\,\mathrm{d}\mu(t)$

M-mixture	with $Z_{\alpha}^{M}(p,q) = \int_{t \in \mathscr{X}} M_{\alpha}(p(t), q(t))\,\mathrm{d}\mu(t)$

Statistical distance	$D(p:q)$
Dual/reverse distance D^*	$D^*(p:q) := D(q:p)$
Kullback-Leibler divergence	$\mathrm{KL}(p:q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\mu(x)$
reverse Kullback-Leibler divergence	$\mathrm{KL}^*(p:q) = \mathrm{KL}(q:p) = \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}\mu(x)$
Jeffreys divergence	$J(p;q) = \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = \int (p(x) - q(x)) \log \frac{p(x)}{q(x)} \mathrm{d}\mu(x)$
Resistor divergence	$\frac{1}{R(p;q)} = \frac{1}{2}\left(\frac{1}{\mathrm{KL}(p:q)} + \frac{1}{\mathrm{KL}(q:p)}\right)$. $R(p;q) = \frac{2J(p;q)}{\mathrm{KL}(p:q)\,\mathrm{KL}(q:p)}$
skew K-divergence	$K_{\alpha}(p:q) = \int p(x) \log \frac{p(x)}{(1-\alpha)p(x) + \alpha q(x)} \mathrm{d}\mu(x)$
	$\mathrm{JS}(p,q)$