Inequalities between the Jenson-Shannon and Jeffreys divergences

Tech. Note 004 v2
http://threeplusone.com/jensenshannon

Gavin E. Crooks

2008

Jeffreys' divergence and the Jensen-Shannon divergence are shown to be related by an inequality that involves a transcendental function of the Jeffreys divergence.

Jeffreys' divergence[1] [1–3]

$$\begin{aligned}\text{Jeffreys}(p;q) &= \tfrac{1}{2}D(p\|q) + \tfrac{1}{2}D(p\|q)\\ &= \sum_i p_i \tfrac{1}{2}\log\frac{p_i}{q_i} + \tfrac{1}{2}\sum_i q_i \log\frac{q_i}{p_i}\end{aligned}$$

and the Jensen-Shannon divergence [4–9, 3]

$$\begin{aligned}\text{JS}(p;q) &= \tfrac{1}{2}D(p\|\tfrac{1}{2}(p+q)) + \tfrac{1}{2}D(q\|\tfrac{1}{2}(p+q))\\ &= \tfrac{1}{2}\sum_i p_i \ln\frac{p_i}{\frac{1}{2}(p_i+q_i)} + \tfrac{1}{2}\sum_i q_i \ln\frac{q_i}{\frac{1}{2}(p_i+q_i)}\end{aligned}$$

are related by the inequality

$$\text{JS}(p;q) \leqslant \min\left\{\tfrac{1}{4}\text{Jeffreys}(p;q)\ ,\ \ln\frac{2}{1+e^{-\text{Jeffreys}(p;q)}}\right\} \tag{1}$$

This inequality is considerable sharper than the inequality $\text{JS}(p;q) \leqslant \frac{1}{2}\text{Jeffreys}(p;q)$ described by Lin [4].

The first part of inequality (1), $\text{JS}(p;q) \leqslant \frac{1}{4}\text{Jeffreys}(p;q)$, is described by Taneja [10]. We note that many interesting measures between probability distributions can be written as an f-divergence [11, 12, 3]

$$C_f(p;q) = \sum_i p_i f(\frac{q_i}{p_i}) = \left\langle f(\frac{q_i}{p_i})\right\rangle \geqslant 0$$

where the function f is convex and normalized such that $f(1) = 0$. For example, if $f_D(x) \equiv -\ln x$ then C_f is the relative entropy $D(p\|q)$. The relation $C_f(p;q) \geqslant 0$ follows from an application of Jensen's inequality [13] for convex functions $\langle f(x)\rangle \geqslant f(\langle x\rangle)$.

Now suppose that we can write

$$f_c(x) = f_b(x) - c f_a(x)$$

where f_a, f_b and f_c are all convex and normalized, and c is a constant. Then $\langle f_c\rangle = \langle f_b\rangle - c\langle f_a\rangle \geqslant 0$ or equivalently

$$\langle f_b\rangle \geqslant c\langle f_a\rangle$$

[1] Note that definitions of Jeffreys divergence often omit the factors of $\frac{1}{2}$.

The desired inequality follows given

$$\begin{aligned}f_{\text{Jeffreys}}(x) &= \tfrac{1}{2}x\ln(x) - \tfrac{1}{2}\ln(x)\\ f_{\text{JS}}(x) &= \tfrac{1}{2}\ln\frac{2}{1+x} + \tfrac{1}{2}x\ln\frac{2}{1+x^{-1}}\\ f_c(x) &= f_{\text{Jeffreys}}(x) - 4f_{\text{JS}}(x)\end{aligned}$$

This inequality has the same form as the asymptotic scaling between Jensen-Shannon and symeterized KL divergence for infinitesimally different distributions. We use the expansion $\log(1+x) = x - x^2/2 + O(x^3)$ [9] and find that

$$\text{JS}(p;p+dp) = \frac{1}{4}\text{Jeffreys}(p;p+dp) + O(dp^3)\,. \tag{2}$$

The second part of inequality (1) follows from the convexity of the function $f(x) = \ln(1+e^x)$ ($f''(x) \geqslant 0$). Jensen's inequality [13] $\langle f(x)\rangle \geqslant f(\langle x\rangle)$ implies that

$$\left\langle \ln\frac{2}{1+e^x}\right\rangle \leqslant \ln\frac{2}{1+e^{\langle x\rangle}}$$

Therefore,

$$\begin{aligned}\text{JS}(p;q) &= \tfrac{1}{2}\sum_i p_i\ln\frac{p_i}{\frac{1}{2}(p_i+q_i)} + \tfrac{1}{2}\sum_i q_i\ln\frac{q_i}{\frac{1}{2}(p_i+q_i)}\\ &= \tfrac{1}{2}\sum_i p_i\ln\frac{2}{1+e^{\ln\frac{q_i}{p_i}}} + \tfrac{1}{2}\sum_i q_i\ln\frac{2}{1+e^{\ln\frac{p_i}{q_i}}}\\ &\leqslant \tfrac{1}{2}\ln\frac{2}{1+\exp\big(-D(p\|q)\big)} + \tfrac{1}{2}\ln\frac{2}{1+\exp\big(-D(q\|p)\big)}\\ &\leqslant \ln\frac{2}{1+\exp\big(-\text{Jeffreys}(p;q)\big)}\end{aligned}$$

The last line follows from the previous line by a second application of the same Jensen inequality. Since the J-divergence ranges between zero and positive infinity, whereas the Jensen-Shannon divergence ranges between zero and $\ln 2$ [i.e. 1 bit], this inequality has the correct limits for identical ($p_i = q_i$, $\text{JS}(p;q) = \text{Jeffreys}(p;q) = 0$) and orthogonal ($p_i q_i = 0$, $\text{JS}(p;q) = \ln 2$, $\text{Jeffreys}(p;q) = +\infty$) distributions.

Note that we can split the Jensen-Shannon divergence into two directed parts [5], and write the equivalent of inequality (1) with respect to each part separately.

$$\tfrac{1}{2}\mathrm{D}\left(p\|\tfrac{1}{2}(p+q)\right) \leqslant \tfrac{1}{2}\ln\frac{2}{1+\exp\left(-\mathrm{D}(p\|q)\right)} \qquad (3)$$

There is no corresponding lower bound; for any value of the Jensen-Shannon divergence the Jeffreys divergence can be arbitrarily large. Consider a pair of binary distributions $(a, 1-a)$ and $(b, 1-b)$. Fix a and let b limit to zero. As b decreases the Jensen-Shannon divergence will limit to a fixed value between 0 and $\ln 2$, but the Jeffreys divergence will increase to infinity.

References

[1] H. Jeffreys. *Theory of probability*. Clarendon Press, Oxford, 2nd edition, 1948.

[2] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[3] G. E. Crooks. On measures of entropy and information, 2016. Tech. Note 009 v5 beta http://threeplusone.com/info.

[4] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, 37:145–151, 1991. doi:10.1109/18.61115.

[5] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory*, 46(4):1602–1609, 2000.

[6] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860, 2003. doi:10.1109/TIT.2003.813506.

[7] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and and its statistical applications. *Ann. Inst. Statist. Math.*, 55(3):639–653, 2003. doi:10.1007/BF02517812.

[8] B. Fuglede and F. Topsøe. Jensen-Shannon divergence and Hilbert space embedding. In *IEEE Int. Sym. on Information Theory*, page 31. 2004. doi:10.1109/ISIT.2004.1365067.

[9] A. Majtey, P. W. Lamberti, M. T. Martin, and A. Plastino. Wootters' distance revisited: a new distinguishability criterium. *Eur. Phys. J. D*, 32:413–419, 2005. doi:10.1140/epjd/e2005-00005-1.

[10] I. J. Taneja. Refinement inequalities among symmetric divergence measure. *Aust. J. Math. Anal. Appl.*, 2(1):8, 2005.

[11] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. B*, 28(1):131–142, 1966.

[12] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[13] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906. doi:10.1007/BF02418571.