Jensen-Shannon divergence and diversity index: Origins and some extensions

Frank Nielsen
Sony Computer Science Laboratories Inc.
Tokyo, Japan

April 2021

Abstract

Lin coined the skewed Jensen-Shannon divergence between two distributions in 1991, and further extended it to the Jensen-Shannon diversity of a set of distributions. Sibson proposed the information radius based on Rényi α-entropies in 1969, and recovered for the special case of $\alpha = 1$ the Jensen-Shannon diversity index. In this note, we summarize how the Jensen-Shannon divergence and diversity index were extended by either considering skewing vectors or using mixtures induced by generic means.

1 Origins

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space, and $(w_1, P_1), \ldots, (w_n, P_n)$ be n weighted probability measures dominated by a measure μ (with $w_i > 0$ and $\sum w_i = 1$). Denote by $\mathcal{P} := \{(w_1, p_1), \ldots, (w_n, p_n)\}$ the set of their weighted Radon-Nikodym densities $p_i = \frac{\mathrm{d}P_i}{\mathrm{d}\mu}$ with respect to μ.

A *statistical divergence* $D[p : q]$ is a measure of dissimilarity between two densities p and q (i.e., a 2-point distance) such that $D[p : q] \geq 0$ with equality if and only if $p(x) = q(x)$ μ-almost everywhere. A *statistical diversity index* $D(\mathcal{P})$ is a measure of variation of the weighted densities in $\mathcal{P}$ related to a measure of centrality, i.e., a n-point distance which generalizes the notion of 2-point distance when $\mathcal{P}_2(p, q) := \{(\frac{1}{2}, p_1), (\frac{1}{2}, p_2)\}$:

$$D[p : q] := D(\mathcal{P}_2(p, q)).$$

The fundamental measure of dissimilarity in information theory is the *I-divergence* (also called the *Kullback-Leibler divergence*, KLD, see Equation (2.5) page 5 of [5]):

$$D_{\mathrm{KL}}[p : q] := \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x).$$

The KLD is asymmetric (hence the delimiter notation ":" instead of ',') but can be symmetrized by defining the Jeffreys *J-divergence* (Jeffreys divergence, denoted by I_2 in Equation (1) in 1946's paper [4]):

$$D_J[p, q] := D_{\mathrm{KL}}[p : q] + D_{\mathrm{KL}}[q : p] = \int_{\mathcal{X}} (p(x) - q(x)) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x).$$

Although symmetric, any positive power of Jeffreys divergence fails to satisfy the triangle inequality: That is, D_J^α is never a metric distance for any $\alpha > 0$, and furthermore D_J^α cannot be upper bounded.

In 1991, Lin proposed the asymmetric *K-divergence* (Equation (3.2) in [7]):

$$D_K[p:q] := D_{\mathrm{KL}}\left[p : \frac{p+q}{2}\right],$$

and defined the *L-divergence* by analogy to Jeffreys's symmetrization of the KLD (Equation (3.4) in [7]):

$$D_L[p,q] = D_K[p:q] + D_K[q:p].$$

By noticing that

$$D_L[p,q] = 2h\left[\frac{p+q}{2}\right] - (h[p]+h[q]),$$

where h denotes Shannon entropy (Equation (3.14) in [7]), Lin coined the (skewed) *Jensen-Shannon divergence* between two weighted densities $(1-\alpha, p)$ and (α, q) for $\alpha \in (0,1)$ as follows (Equation (4.1) in [7]):

$$D_{\mathrm{JS},\alpha}[p,q] = h[(1-\alpha)p + \alpha q] - (1-\alpha)h[p] - \alpha h[q]. \tag{1}$$

Finally, Lin defined the *generalized Jensen-Shannon divergence* (Equation (5.1) in [7]) for a finite weighted set of densities:

$$D_{\mathrm{JS}}[\mathcal{P}] = h\left[\sum_i w_i p_i\right] - \sum_i w_i h[p_i].$$

This generalized Jensen-Shannon divergence is nowadays called the *Jensen-Shannon diversity index*.

To contrast with the Jeffreys' divergence, the Jensen-Shannon divergence (JSD) $D_{\mathrm{JS}} := D_{\mathrm{JS},\frac{1}{2}}$ is upper bounded by $\log 2$ (does not require the densities to have the same support), and $\sqrt{D_{\mathrm{JS}}}$ is a metric distance [2, 3]. Lin cited precursor work [17, 8] yielding definition of the Jensen-Shannon divergence: The Jensen-Shannon divergence of Eq. 1 is the so-called "increments of entropy" defined in (19) and (20) of [17].

The Jensen-Shannon diversity index was also obtained very differently by Sibson in 1969 when he defined the *information radius* [16] of order α using Rényi α-means and Rényi α-entropies [15]. In particular, the information radius IR_1 of order 1 of a weighted set $\mathcal{P}$ of densities is a diversity index obtained by solving the following variational optimization problem:

$$\mathrm{IR}_1[\mathcal{P}] := \min_c \sum_{i=1}^n w_i D_{\mathrm{KL}}[p_i : c]. \tag{2}$$

Sibson solved a more general optimization problem, and obtained the following expression (term K_1 in Corollary 2.3 [16]):

$$\mathrm{IR}_1[\mathcal{P}] = h\left[\sum_i w_i p_i\right] - \sum_i w_i h[p_i] := D_{\mathrm{JS}}[\mathcal{P}].$$

Thus Eq. 2 is a variational definition of the Jensen-Shannon divergence.

2 Some extensions

- **Skewing the JSD.**

  The K-divergence of Lin can be skewed with a scalar parameter $\alpha \in (0,1)$ to give

$$D_{K,\alpha}[p:q] := D_{\mathrm{KL}}\left[p:(1-\alpha)p+\alpha q\right]. \tag{3}$$

  Skewing parameter α was first studied in [6] (2001, see Table 2 of [6]). We proposed to unify the Jeffreys divergence with the Jensen-Shannon divergence as follows (Equation 19 in [9]):

$$D^J_{K,\alpha}[p:q] := \frac{D_{K,\alpha}[p:q]+D_{K,\alpha}[q:p]}{2}. \tag{4}$$

  When $\alpha=\frac{1}{2}$, we have $D^J_{K,\frac{1}{2}}=D_{\mathrm{JS}}$, and when $\alpha=1$, we get $D^J_{K,1}=\frac{1}{2}D_J$.

  Notice that

$$D_{\mathrm{JS}}^{\alpha,\beta}[p;q] := (1-\beta)D_{\mathrm{KL}}[p:(1-\alpha)p+\alpha q]+\beta D_{\mathrm{KL}}[q:(1-\alpha)p+\alpha q]$$

  amounts to calculate

$$h^\times[(1-\beta)p+\beta q:(1-\alpha)p+\alpha q]-((1-\beta)h[p]+\beta h[q])$$

  where

$$h^\times[p,q] := \int -p(x)\log q(x)\mathrm{d}\mu(x)$$

  denotes the *cross-entropy*. By choosing $\alpha=\beta$, we have $h^\times[(1-\beta)p+\beta q:(1-\alpha)p+\alpha q] = h[(1-\alpha)p+\alpha q]$, and thus recover the skewed Jensen-Shannon divergence of Eq. 1.

  In [11] (2020), we considered a positive *skewing vector* $\alpha\in[0,1]^k$ and a unit positive weight w belonging to the standard simplex Δ_k, and defined the following *vector-skewed Jensen-Shannon divergence*:

$$\begin{aligned} D_{\mathrm{JS}}^{\alpha,w}[p:q] &:= \sum_{i=1}^k D_{\mathrm{KL}}[(1-\alpha_i)p+\alpha_i q:(1-\bar\alpha)p+\bar\alpha q], && (5)\\ &= h[(1-\bar\alpha)p+\bar\alpha q]-\sum_{i=1}^k h[(1-\alpha_i)p+\alpha_i q], && (6)\end{aligned}$$

  where $\bar\alpha=\sum_{i=1}^k w_i\alpha_i$. The divergence $D_{\mathrm{JS}}^{\alpha,w}$ generalizes the (scalar) skew Jensen-Shannon divergence when $k=1$, and is a Ali-Silvey-Csiszár f-divergence upper bounded by $\log\frac{1}{\bar\alpha(1-\bar\alpha)}$ [11].

- **A priori mid-density**. The JSD can be interpreted as the total divergence of the densities to the *mid-density* $\bar p=\sum_{i=1}^n w_i p_i$, a statistical mixture:

$$D_{\mathrm{JS}}[\mathcal{P}] = \sum_{i=1}^n w_i D_{\mathrm{KL}}[p_i:\bar p] = h[\bar p]-\sum_{i=1}^n w_i h[p_i].$$

  Unfortunately, the JSD between two Gaussian densities is not known in closed form because of the definite integral of a log-sum term (i.e., K-divergence between a density and a mixture density $\bar{p}$). For the special case of the Cauchy family, a closed-form formula [14] for the JSD between two Cauchy densities was obtained. Thus we may choose a *geometric mixture distribution* [10] instead of the ordinary arithmetic mixture $\bar{p}$. More generally, we can choose any weighted mean M_α (say, the geometric mean, or the harmonic mean, or any other power mean) and define a generalization of the K-divergence of Equation 3:

$$D_K^{M_\alpha}[p:q] := D_K[p:(pq)_{M_\alpha}], \tag{7}$$

where

$$(pq)_{M_\alpha}(x) := \frac{M_\alpha(p(x), q(x))}{Z_{M_\alpha}(p:q)}$$

is a statistical M-mixture with $Z_{M_\alpha}(p,q)$ denoting the normalizing coefficient:

$$Z_{M_\alpha}(p:q) = \int M_\alpha(p(x), q(x))\mathrm{d}\mu(x)$$

so that $\int (pq)_{M_\alpha}(x)\mathrm{d}\mu(x) = 1$. These M-mixtures are well-defined provided the convergence of the definite integrals.

Then we define a generalization of the JSD [10] termed (M_α, N_β)*-Jensen-Shannon divergence* as follows:

$$D_{\mathrm{JS}}^{M_\alpha,N_\beta}[p:q] := N_\beta\left(D_K[p:(pq)_{M_\alpha}], D_K[q:(pq)_{M_\alpha}]\right), \tag{8}$$

where N_β is yet another weighted mean to average the two M_α-K-divergences. We have $D_{\mathrm{JS}} = D_{\mathrm{JS}}^{A,A}$ where $A(a,b) = \frac{a+b}{2}$ is the arithmetic mean. The geometric JSD yields a closed-form formula between two multivariate Gaussians, and has been used in deep learning [1]. More generally, we may consider the Jensen-Shannon symmetrization of an arbitrary distance D as

$$D_{M_\alpha,N_\beta}^{\mathrm{JS}}[p:q] := N_\beta\left(D[p:(pq)_{M_\alpha}], D[q:(pq)_{M_\alpha}]\right). \tag{9}$$

- **A posteriori mid-density**. We consider a generalization of Sibson's information radius [16]. Let $S_w(a_1, \ldots, a_n)$ denote a generic weighted mean of n positive scalars $a_1, \ldots, a_n$, with weight vector $w \in \Delta_n$. Then we define the *S-variational Jensen-Shannon diversity index* [12] as

$$D_{\mathrm{vJS}}^{S_w}(\mathcal{P}) := \min_c S_w\left(D_{\mathrm{KL}}[p_1:c], D_{\mathrm{KL}}[p_n:c]\right). \tag{10}$$

When $S_w = A_w$ (with $A_w(a_1, \ldots, a_n) = \sum_{i=1}^n w_i a_i$ the arithmetic weighted mean), we recover the ordinary Jensen-Shannon diversity index. More generally, we define the *S-Jensen-Shannon index of an arbitrary distance D* as

$$D_{S_w}^{\mathrm{vJS}}(\mathcal{P}) := \min_c S_w\left(D[p_1:c], \ldots, D[p_n:c]\right). \tag{11}$$

When $n = 2$, this yields a Jensen-Shannon-symmetrization of distance D.

The variational optimization defining the JSD can also be constrained to a (parametric) family of densities $\mathcal{D}$, thus defining the *$(S,\mathcal{D})$-relative Jensen-Shannon diversity index*:

$$D_{\mathrm{vJS}}^{S_w,\mathcal{D}}(\mathcal{P}) := \min_{c\in\mathcal{D}} S_w\left(D_{\mathrm{KL}}[p_1:c], \ldots, D_{\mathrm{KL}}[p_n:c]\right). \tag{12}$$

The relative Jensen-Shannon divergences are useful for clustering applications: Let p_{θ_1} and p_{θ_2} be two densities of an exponential family $\mathcal{E}$ with cumulant function $F(\theta)$. Then the $\mathcal{E}$-relative Jensen-Shannon divergence is the Bregman information of $\mathcal{P}_2(p,q)$ for the conjugate function $F^*(\eta) = -h[p_\theta]$ (with $\eta = \nabla F(\theta)$). The $\mathcal{E}$-relative JSD amounts to a *Jensen divergence* for F^*:

$$
\begin{aligned}
D_{\mathrm{vJS}}[p_{\theta_1}, p_{\theta_2}] &= \min_\theta \frac{1}{2}\left\{D_{\mathrm{KL}}[p_{\theta_1} : p_\theta] + D_{\mathrm{KL}}[p_{\theta_2} : p_\theta]\right\}, && (13)\\
&= \min_\theta \frac{1}{2}\left\{B_F[\theta : \theta_1] + B_F[\theta : \theta_2]\right\}, && (14)\\
&= \min_\eta \frac{1}{2}\left\{B_{F^*}[\eta_1 : \eta] + B_{F^*}[\eta_2 : \eta]\right\}, && (15)\\
&= \frac{F^*(\eta_1) + F^*(\eta_2)}{2} - F^*(\eta^*), && (16)\\
&=: J_{F^*}(\eta_1, \eta_2), && (17)
\end{aligned}
$$

since $\eta^* := \frac{\eta_1+\eta_2}{2}$ (a right-sided *Bregman centroid* [13]).

References

[1] Jacob Deasy, Nikola Simidjievski, and Pietro Liò. Constraining Variational Inference with Geometric Jensen-Shannon Divergence. In *Advances in Neural Information Processing Systems*, 2020.

[2] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[3] Bent Fuglede and Flemming Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

[4] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[5] Solomon Kullback. *Information theory and statistics.* Courier Corporation, 1997.

[6] Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics (AISTATS)*, page 65–72, 2001.

[7] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[8] Jianhua Lin and SKM Wong. Approximation of discrete probability distributions based on a new divergence measure. *Congressus Numerantium (Winnipeg)*, 61:75–80, 1988.

[9] Frank Nielsen. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010. URL https://arxiv.org/abs/1009.4004.

[10] Frank Nielsen. On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy*, 21(5), 2019. ISSN 1099-4300. doi: 10.3390/e21050485. URL https://www.mdpi.com/1099-4300/21/5/485.

[11] Frank Nielsen. On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid. *Entropy*, 22(2), 2020. ISSN 1099-4300. doi: 10.3390/e22020221. URL https://www.mdpi.com/1099-4300/22/2/221.

[12] Frank Nielsen. On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius. *Entropy*, 23(4), 2021. ISSN 1099-4300. doi: 10.3390/e23040464. URL https://www.mdpi.com/1099-4300/23/4/464.

[13] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.

[14] Frank Nielsen and Kazuki Okamura. On f-divergences between cauchy distributions. *arXiv:2101.12459*, 2021.

[15] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[16] Robin Sibson. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(2):149–160, 1969.

[17] Andrew KC Wong and Manlai You. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):599–609, 1985.