Video abstract

# On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means

*Entropy* 2019, *21*(5), 485

arxiv:1009.4004

Frank Nielsen

Sony Computer Science Laboratories Inc.

Sony CSL

# Two historical roots of the celebrated Jensen-Shannon divergence (1/2)

- **Lin (1991)**: Jensen-Shannon divergence defined as a *generalization of the* ***L divergence*** (divergence when *n=2* and diversity index when *n>2*):

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

⬅ common definition of the JSD (= symmetrization of the K-divergence)

JS diversity ➡ $$D_{\mathrm{JS}}[p_1,\ldots,p_n;w_1,\ldots,w_n] := H\left[\sum_{i=1}^n w_i p_i\right] - \sum_{i=1}^n w_i H[p_i]$$

with entropy $H(p) = \int p \log \frac{1}{p} \mathrm{d}\mu$

"Entropy of the average distribution minus the average of the entropies"

where I (Kullback-Leibler divergence), J (Jeffreys divergence), K, L divergences are defined by

$$\begin{aligned}
I[p:q] &= \int p \log \frac{p}{q} \mathrm{d}\mu \quad \text{⬅ asymmetric}\\
J[p;q] &= I[p:q] + I[q:p] = J[q;p]\\
K[p:q] &= I\left[p : \frac{p+q}{2}\right] = \int p \log \frac{2p}{p+q} \mathrm{d}\mu \quad \text{⬅ asymmetric}\\
L[p;q] &= K[p:q] + K[q:p] = 2H\left[\frac{p+q}{2}\right] - H[p] - H[q]
\end{aligned}$$

see also **arxiv:1009.4004**

# Two historical roots of the celebrated Jensen-Shannon divergence (2/2)

- **Sibson (1969)**: Jensen-Shannon divergence as an **"information radius"** by studying a **variational problem** relying on Rényi α-divergences and Rényi α-means for aggregating these Rényi α-divergences **[*Entropy* 2021, *23*(4), 464]**
- When α=1, the **information radius of order 1** corresponds to the Jensen-Shannon divergence/diversity index (1-mean=arithmetic):

$$D_{\mathrm{JS}}[p_1,\ldots,p_n;w_1,\ldots,w_n] := \min_c \sum_{i=1}^n w_i D_{\mathrm{KL}}[p_i : c]$$

KLD right centroid:

$$c^* = \sum_{i=1}^n w_i p_i$$

$$D_{\mathrm{JS}}[p_1,\ldots,p_n;w_1,\ldots,w_n] := \sum_{i=1}^n w_i D_{\mathrm{KL}}\left[p_i : \sum_{i=1}^n w_i p_i\right]$$

JS diversity →

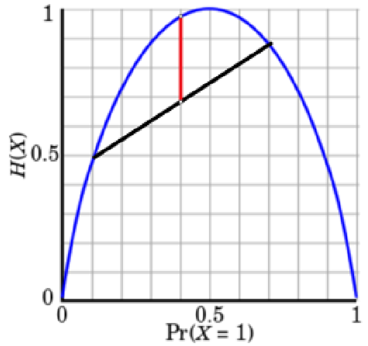
$$= H\left[\sum_{i=1}^n w_i p_i\right] - \sum_{i=1}^n w_i H[p_i]$$

# Key properties of Jensen-Shannon divergence and its computational limitation

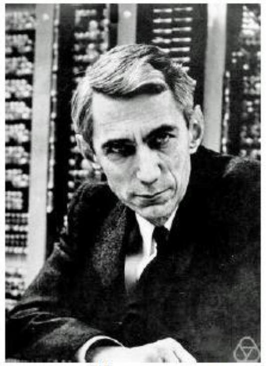
Jensen
1859-1925

- Jensen-Shannon divergence stems from **Jensen's inequality** on *minus the concave function* of **Shannon entropy**
- Jensen-Shannon divergence is **always** upper bounded by log 2 (important property when distributions have different supports)
- **The square root of the Jensen-Shannon divergence** is a metric distance. However, JSD usually not computationally tractable for continuous distributions because of the integral of the log mixture density:

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




Some remarkable exceptions: E.g., KLD between mixtures of two Cauchy distributions in closed form! However, KLD between Gaussian mixtures is provably not analytic.



Shannon
1916-2001

[arXiv:2104.13801]

# Generalizing the Jensen-Shannon divergence using *abstract means* to define *generic statistical mixtures*

- Consider a **generic mean** M(a,b) and define the **statistical $(pq)^M$ mixture**:

$$(pq)^M(x) = \frac{M(p(x), q(x))}{\int M(p(x), q(x)) \mathrm{d}\mu(x)}$$

(when M(a,b) is the arithmetic mean, the normalization constant is one)

- **M-Jensen-Shannon divergence** upper bounds the ordinary JSD:

$$\begin{aligned} D_{\mathrm{JS}}^M[p,q] &= \frac{1}{2}\left(D_{\mathrm{KL}}\left[p:(pq)^M\right] + D_{\mathrm{KL}}\left[q:(pq)^M\right)\right] \\ &= h_\times\left[\frac{p+q}{2}:(pq)^M\right] - \frac{h[p]+h[q]}{2} \\ &= D_{\mathrm{KL}}\left[\frac{p+q}{2}:(pq)^M\right] + D_{\mathrm{JS}}[p,q] \geq D_{\mathrm{JS}}[p,q] \end{aligned}$$

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

Cross-entropy:

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

# The geometric Jensen-Shannon divergence

- JS obtained for the geometric mixture instead of the arithmetic mixture:

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

$$= \frac{1}{4}D_J[p,q] - D_B[p,q] \geq 0$$

where $D_B[p,q] = -\log \int \sqrt{p(x)q(x)}\mathrm{d}\mu(x)$

where J is the Jeffreys divergence and $D_B$ is the Bhattacharyya distance

- When densities p and q belong to a same **exponential family** $E_F$,

$$\mathcal{E}_F = \left\{p_\theta(x)\mathrm{d}\mu = \exp(\theta^\top x - F(\theta))\mathrm{d}\mu : \theta \in \Theta\right\}$$

we get a closed-form formula for the G-JSD:

$$\mathrm{JS}^G(p_{\theta_1}, p_{\theta_2}) = \underbrace{\frac{1}{4}(\theta_2-\theta_1)^\top(\nabla F(\theta_2) - \nabla F(\theta_1))}_{\frac{1}{4}J(p_{\theta_1},p_{\theta_2})} - \underbrace{\left(\frac{F(\theta_1)+F(\theta_2)}{2} - F\left(\frac{\theta_1+\theta_2}{2}\right)\right)}_{B(p_{\theta_1},p_{\theta_2})=J_F(\theta_1,\theta_2)}.$$

# The harmonic Jensen-Shannon divergence

- H-JSD is well-suited for getting closed-form between Cauchy distributions

$$\mathcal{C}_\Gamma := \left\{ p_\gamma(x) = \frac{1}{\gamma} p_{\mathrm{std}}\left(\frac{x}{\gamma}\right) = \frac{\gamma}{\pi(\gamma^2 + x^2)} : \gamma \in \Gamma = (0, \infty) \right\}$$

- Weighted harmonic mean:

$$H_\alpha(x, y) := \frac{1}{(1-\alpha)\frac{1}{x} + \alpha \frac{1}{y}} = \frac{xy}{(1-\alpha)y + \alpha x} = \frac{xy}{(xy)_{1-\alpha}}, \quad \alpha \in [0, 1].$$

- Get closed-form formula for H-JSB between Cauchy densities:

$$\begin{aligned} \mathrm{JS}^H(p:q) &= \frac{1}{2}\left(\mathrm{KL}\left(p : (pq)^H_{\frac{1}{2}}\right) + \mathrm{KL}\left(q : (pq)^H_{\frac{1}{2}}\right)\right), \\ \mathrm{JS}^H(p_{\gamma_1} : p_{\gamma_2}) &= \frac{1}{2}\left(\mathrm{KL}\left(p_{\gamma_1} : p_{\frac{\gamma_1+\gamma_2}{2}}\right) + \mathrm{KL}\left(p_{\gamma_2} : p_{\frac{\gamma_1+\gamma_2}{2}}\right)\right) \\ &= \log\left(\frac{(3\gamma_1 + \gamma_2)(3\gamma_2 + \gamma_1)}{8\sqrt{\gamma_1\gamma_2}(\gamma_1 + \gamma_2)}\right). \end{aligned}$$

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

# JS-symmetrization of any arbitrary distances D

- **N-Jeffreys symmetrization**: Use a generic weighted mean N(a,b) to average the sided divergences:

$$J_D^{N_\beta}(p_1 : p_2) = N_\beta(D(p_1 : p_2), D(p_2 : p_1))$$

⇒ recover resistor average div. of **Johnson and Sinanovic (2001)**

$$\frac{1}{R(p;q)} = \frac{1}{2}\left(\frac{1}{\mathrm{KL}(p:q)} + \frac{1}{\mathrm{KL}(q:p)}\right)$$

- **M-K symmetrization**:

$$K_{D,\alpha}^M(p : q) = D\left(p : (pq)_\alpha^M\right)$$

- **(M,N) JS-symmetrization**: use two generic means M(a,b) and N(a,b) to define the of an arbitrary distance D:

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

$$\mathrm{JS}_D^{M_\alpha, N_\beta}(p_1 : p_2) = N_\beta(D(p_1, (p_1p_2)_\alpha^M), D(p_2, (p_1p_2)_\alpha^M))$$

M-K sided divergences