On the Jensen-Shannon symmetrization of distances relying on abstract means

Entropy 2019, 21(5), 485 (also arxiv 1904.04017)
https://franknielsen.github.io/M-JS/

Frank Nielsen

Sony Computer Science Laboratories, Inc
Tokyo, Japan

April 2019

Outline

- Fundamental dissimilarity between distributions in information sciences [3] (Kullback-Leibler divergence and f-divergences) and their usual Jeffreys and Jensen-Shannon (JS) symmetrizations

- Jensen-Shannon divergence between Gaussian densities is not available in closed-form

- Definitions: JS-symmetrizations of *any* parameter distance and of *any* statistical distance using abstract means, and properties

- Three cases studies with reported closed-form expressions:
  - Geometric Jensen-Shannon divergence for multivariate Gaussians (or any exponential family)

  - Harmonic Jensen-Shannon divergence for Cauchy distributions

  - Arithmetic (=ordinary) Jensen-Shannon divergence for mixture distributions

- Conclusion

The Kullback-Leibler divergence (KLD)

- **Kullback-Leibler divergence** [3] is the relative entropy:

$$\mathrm{KL}[p:q] := \int p \log \frac{p}{q} \mathrm{d}\mu = h_\times[p:q] - h[p]$$

$$h_\times[p:q] := \int p \log \frac{1}{q} \mathrm{d}\mu, \quad h[p] = h_\times[p:p]$$

- KLD unbounded and potentially ∞ when the integral diverges
- Asymmetric (non-metric): Define the **reverse Kullback-Leibler divergence**

$$\mathrm{KL}^*[p:q] := \mathrm{KL}[q:p]$$

Statistical distance and parameter distance

- $\mathrm{KL}[p : q]$ is a **statistical distance** between probability densities (or measures), hence the bracket notation
- When $p = p_{\theta_1}$ and $q = p_{\theta_2}$ belong to the same parametric family $\mathcal{P}$ of distributions, the statistical distance D amount to a **parameter distance** $D_{\mathcal{P}}$:

$$D_{\mathcal{P}}(\theta_1 : \theta_2) := D[p_{\theta_1} : p_{\theta_2}]$$

- For example, when $p = p_{\theta_1}$ and $q = p_{\theta_2}$ belong to the same exponential family [11] $\mathcal{E}$, we have

$$D_{\mathcal{E}}(\theta_1 : \theta_2) := \mathrm{KL}[p_{\theta_1} : p_{\theta_2}]$$

with parameter divergence

$$D_{\mathcal{E}}(\theta_1 : \theta_2) = B_F^*(\theta_1 : \theta_2) = B_F(\theta_2 : \theta_1)$$

where B_F is the **Bregman divergence** [2] defined for a strictly convex and differentiable convex generator F

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - \langle \theta - \theta', \nabla F(\theta') \rangle$$

Renown symmetrizations of the Kullback-Leibler divergence

- **Jeffreys divergence** [8] symmetrizes KLD:

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

$\rightarrow$ unbounded

- **Jensen-Shannon divergence** [6] also symmetrizes KLD:

$$\begin{aligned}\mathrm{JS}[p;q] &:= \frac{1}{2}\left(\mathrm{KL}\left[p:\frac{p+q}{2}\right] + \mathrm{KL}\left[q:\frac{p+q}{2}\right]\right)\\ &= \frac{1}{2}\int\left(p\log\frac{2p}{p+q} + q\log\frac{2q}{p+q}\right)\mathrm{d}\mu.\end{aligned}$$

$\rightarrow$ always bounded:

$$0 \leq \mathrm{JS}[p:q] \leq \log 2$$

$\rightarrow$ $\sqrt{\mathrm{JS}}$ is metric distance [5]

Symmetrizations of statistical f-divergences

- Class of f-**divergences** [4] for a convex function f strictly convex at 1 (with $f(1) = f'(1) = 0$):

$$I_f[p:q] = \int pf\left(\frac{q}{p}\right)\mathrm{d}\mu.$$

- KLD belongs to the f-divergences for f-generator $f_{\mathrm{KL}}(u) = -\log u$

$$\mathrm{KL}[p:q] = I_{f_{\mathrm{KL}}}[q:p]$$

- The Jeffreys and Jensen-Shannon f-generators are

$$\begin{aligned} f_J(u) &:= (u-1)\log u, \\ f_{\mathrm{JS}}(u) &:= -(u+1)\log\frac{1+u}{2} + u\log u. \end{aligned}$$

JS-symmetrization of parameter distances

- For any arbitrary parameter distance $D(\theta_1 : \theta_2)$ and $\alpha \in [0,1]$:

$$\begin{aligned}
\mathrm{JS}_D^\alpha(\theta_1 : \theta_2) &:= (1-\alpha) D\left(\theta_1 : (1-\alpha)\theta_1 + \alpha\theta_2\right) \\
&\quad + \alpha D\left(\theta_2 : (1-\alpha)\theta_1 + \alpha\theta_2\right) \\
&= (1-\alpha) D\left(\theta_1 : (\theta_1 : \theta_2)_\alpha\right) + \alpha D\left(\theta_2 : (\theta_1 : \theta_2)_\alpha\right),
\end{aligned}$$

  where $(\theta_p \theta_q)_\alpha := (1-\alpha)\theta_p + \alpha\theta_q$ to denote the *linear interpolation* (LERP) of the parameters.

- For example, **Jensen-Bregman divergence** [10] JB_F amounts to a **Jensen (gap) divergence** J_F (for a strictly convex generator $F : \Theta \to \mathbb{R}$)

$$\begin{aligned}
\mathrm{JB}_F(\theta : \theta') &:= \frac{1}{2}\left(B_F\left(\theta : \frac{\theta + \theta'}{2}\right) + B_F\left(\theta' : \frac{\theta + \theta'}{2}\right)\right), \\
&= \frac{F(\theta) + F(\theta')}{2} - F\left(\frac{\theta + \theta'}{2}\right) =: J_F(\theta : \theta')
\end{aligned}$$

JS-symmetrization of distances and f-divergences

- In particular, the JS-symmetrization of a f-divergence

$$I_f^\alpha[p:q]{:=}(1-\alpha)I_f[p:(pq)_\alpha]+\alpha I_f[q:(pq)_\alpha],$$

with $(pq)_\alpha=(1-\alpha)p+\alpha q$ is obtained by taking the f-generator

$$f_\alpha^{\mathrm{JS}}(u){:=}(1-\alpha)f(\alpha u+1-\alpha)+\alpha f\left(\alpha+\frac{1-\alpha}{u}\right).$$

- $(pq)_\alpha(x)=(1-\alpha)p(x)+\alpha q(x)$ is a **statistical mixture**

Jensen-Shannon divergence between Gaussians

- Jensen-Shannon divergence interpreted as a statistical Jensen gap divergence for the negative entropy $F = -h$:

$$
\begin{aligned}
\mathrm{JS}[p;q] &:= \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right)\\
&= \frac{1}{2}\int \left(p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q}\right) \mathrm{d}\mu\\
&= h\left[\frac{p+q}{2}\right] - \frac{h[p]+h[q]}{2} = J_{-h}[p;q]
\end{aligned}
$$

- $\frac{p+q}{2}$ is a statistical mixture
- Kullback-Leibler divergence between Gaussian mixtures is provably **not analytic** [14, 13]
  $\rightarrow$ no closed-form formula for the JSD between Gaussians
- Goal is to bypass this computational tractability issue by defining novel kinds of Jensen-Shannon divergences

Abstract means and generalized statistical mixtures

- **Abstract mean** [7] M: continuous bivariate function $M(\cdot, \cdot) : I \times I \to I$ on an interval $I \subset \mathbb{R}$ satifying the *in-betweenness* property:

$$\inf\{x, y\} \leq M(x, y) \leq \sup\{x, y\}, \quad \forall x, y \in I.$$

- **Weighted mean** $M_\alpha(p, q)$ (with $\alpha \in [0, 1]$) using the unique *dyadic expansion* [7] such that $M_0(p, q) = p$ and $M_1(p, q) = q$.

- α-weighted M-**mixture** $(pq)_\alpha^M$ (with $\alpha \in [0, 1]$) of densities p and q defined by:

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

$$Z_\alpha^M(p : q) = \int_{t \in \mathcal{X}} M_\alpha(p(t), q(t)) \mathrm{d}\mu(t)$$

Examples of means M and M-mixtures

- For $x, y > 0$,
  - **arithmetic mean** $A_\alpha(x,y) = (1-\alpha)x + \alpha_y$, $(h(u) = u)$
  - **geometric mean** $G_\alpha(x,y) = x^{1-\alpha} y^\alpha$, $(h(u) = \log u)$
  - **harmonic mean** $H_\alpha(x,y) = \frac{xy}{(1-\alpha)y+\alpha x}$, $(h(u) = \frac{1}{u})$
  - **quasi-arithmetic means** [9] for h is a strictly monotonous function h

$$M_\alpha^h(x,y) := h^{-1}\left((1-\alpha)h(x) + \alpha h(y)\right)$$

- Statistical M-mixtures and their normalization coefficients:

$$\begin{aligned}
(pq)_\alpha^A(x) &:= (1-\alpha)p(x) + \alpha q(x), \quad Z_\alpha^M(p:q) = 1\\
(pq)_\alpha^G(x) &:= \frac{p(x)^{(1-\alpha)} q(x)^\alpha}{Z_\alpha^G(p:q)}, \quad Z_\alpha^G(p:q) = \int p(x)^{(1-\alpha)} q(x)^\alpha \mathrm{d}\mu(x)\\
(pq)_\alpha^H(x) &:= \frac{1}{Z_\alpha^H(p:q)} \frac{p(x)q(x)}{(1-\alpha)q(x) + \alpha p(x)}, \quad Z_\alpha^H(p:q) = \int \frac{p(x)q(x)}{(1-\alpha)q(x)+\alpha p(x)} \mathrm{d}\mu(x)
\end{aligned}$$

Statistical M-Jensen-Shannon divergences

- Definitions of M-JS D-symmetrizations

$$\begin{aligned}
\mathrm{JS}_D^{M_\alpha}[p:q] &:= (1-\alpha)D\left[p:(pq)_\alpha^M\right]+\alpha D\left[q:(pq)_\alpha^M\right]\\
\mathrm{JS}^{M_\alpha}[p:q] &:= (1-\alpha)\mathrm{KL}\left[p:(pq)_\alpha^M\right]+\alpha\mathrm{KL}\left[q:(pq)_\alpha^M\right]
\end{aligned}$$

- **Key property**: The M-JSD is upper bounded by $\log \frac{Z_\alpha^M(p,q)}{1-\alpha}$ when $M \geq A$.

- Arithmetic mean-Geometric mean-Harmonic mean inequality (AGH):

$$A \geq G \geq H$$

M-JS symmetrizations of D for parametric family: A recipe to get closed-form formula

- Let $\mathcal{P}:=\{p_\theta(x) \ : \ \theta \in \Theta\}$ denote a **parametric family** of densities with convex parameter domain Θ

- **Parameter distance** $D_\mathcal{P}$ from statistical distance D between members of a family:

$$D_\mathcal{P}(\theta_1 : \theta_2):=D[p_{\theta_1} : p_{\theta_2}]$$

- Find abstract mean M such that $(p_{\theta_1} p_{\theta_2})^M_\alpha = p_{(\theta_1 \theta_2)_\alpha}$

- Then the M-JS symmetrization of D amount to the following **parameter divergence**:

$$\mathrm{JS}_D^{M_\alpha}[p_{\theta_1} : p_{\theta_2}] = (1-\alpha) D_\mathcal{P}(\theta_1 : (\theta_1 \theta_2)_\alpha) + \alpha D_\mathcal{P}(\theta_2 : (\theta_1 \theta_2)_\alpha) = \mathrm{JS}^\alpha_{D_\mathcal{P}}(\theta_1 : \theta_2)$$

Example 1: G-JS symmetrizations of KL for exponential families

- **Exponential family** [1] $\mathcal{E}_F$ with log-normalizer F:

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

- Geometric mixture, G-mixture, of exponential families:

$$\begin{aligned}
(p_{\theta_1} p_{\theta_2})^G_\alpha(x) &:= \frac{G_\alpha(p_{\theta_1}(x), p_{\theta_2}(x))}{\int G_\alpha(p_{\theta_1}(t), p_{\theta_2}(t))\mathrm{d}\mu(t)} = \frac{p_{\theta_1}^{1-\alpha}(x) p_{\theta_2}^\alpha(x)}{Z^G_\alpha(p:q)}, \\
&= p_{(\theta_1\theta_2)_\alpha}(x), \\
Z^G_\alpha(p:q) &= \exp(-J_F^\alpha(\theta_1 : \theta_2)), \\
J_F^\alpha(\theta_1 : \theta_2) &:= (F(\theta_1)F(\theta_2))_\alpha - F((\theta_1\theta_2)_\alpha)
\end{aligned}$$

- KLD between Gaussians amount to a reverse Bregman divergence [1] ${B_F}^*$

$$\mathrm{KL}_{\mathcal{P}}(\theta_1 : \theta_2) = \mathrm{KL}(p_{\theta_1} : p_{\theta_2}) = B^*_F(\theta_1 : \theta_2) := B_F(\theta_2 : \theta_1)$$

- G-Jensen-Shannon divergence (for KL):

$$\begin{aligned}\mathrm{JS}^G_\alpha[p_{\theta_1}:p_{\theta_2}] &:= (1-\alpha)\mathrm{KL}[p_{\theta_1}:(p_{\theta_1}p_{\theta_2})^G_\alpha]+\alpha\mathrm{KL}[p_{\theta_2}:(p_{\theta_1}p_{\theta_2})^G_\alpha]\\ &= (1-\alpha)B_F((\theta_1\theta_2)_\alpha:\theta_1)+\alpha B_F((\theta_1\theta_2)_\alpha:\theta_2).\end{aligned}$$

- G-Jensen-Shannon symmetrization for reverse KL:

$$\begin{aligned}\mathrm{JS}_{\mathrm{KL}^*}(p:q) &:= \frac{1}{2}\left(\mathrm{KL}^*\left[p:\frac{p+q}{2}\right]+\mathrm{KL}^*\left[q:\frac{p+q}{2}\right]\right),\\ &= \frac{1}{2}\left(\mathrm{KL}\left[\frac{p+q}{2}:p\right]+\mathrm{KL}\left[\frac{p+q}{2}:q\right]\right)\\ \mathrm{JS}^{G_\alpha}_{\mathrm{KL}^*}[p_{\theta_1}:p_{\theta_2}] &:= (1-\alpha)\mathrm{KL}[(p_{\theta_1}p_{\theta_2})^G_\alpha:p_{\theta_1}]+\alpha\mathrm{KL}[(p_{\theta_1}p_{\theta_2})^G_\alpha:p_{\theta_2}],\\ &= (1-\alpha)B_F(\theta_1:(\theta_1\theta_2)_\alpha)+\alpha B_F(\theta_2:(\theta_1\theta_2)_\alpha)=\mathrm{JB}^\alpha_F(\theta_1:\theta_2),\\ &= (1-\alpha)F(\theta_1)+\alpha F(\theta_2)-F((\theta_1\theta_2)_\alpha),\\ &= J^\alpha_F(\theta_1:\theta_2).\end{aligned}$$

To summarize:

$$\begin{aligned}\mathrm{JS}^{G_\alpha}_{\mathrm{KL}}[p_{\theta_1}:p_{\theta_2}] &= (1-\alpha)B_F\left((\theta_1\theta_2)_\alpha:\theta_1\right)+\alpha B_F\left((\theta_1\theta_2)_\alpha:\theta_2\right),\\ \mathrm{JS}^{G_\alpha}_{\mathrm{KL}^*}[p_{\theta_1}:p_{\theta_2}] &= J^\alpha_F(\theta_1:\theta_2).\end{aligned}$$

$\rightarrow$ Interpretation of the Jensen gap divergence J^α_F as a reverse KL JS-symmetrization between members of the same exponential family

- Case study of G-JS: MultiVariate Gaussian/Normal density with $\lambda:=(\lambda_v,\lambda_M)=(\mu,\Sigma)$:

$$p_\lambda(x;\lambda) := \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\lambda_M|}}\exp\left(-\frac{1}{2}(x-\lambda_v)^\top\lambda_M^{-1}(x-\lambda_v)\right)$$
$$p_\theta(x;\theta) := \exp\left(\langle t(x),\theta\rangle - F_\theta(\theta)\right) = p_\lambda(x;\lambda(\theta))$$

with $\theta=(\theta_v,\theta_M)=\left(\Sigma^{-1}\mu,-\frac{1}{2}\Sigma^{-1}\right)=\theta(\lambda)=\left(\lambda_M^{-1}\lambda_v,-\frac{1}{2}\lambda_M^{-1}\right)$, $t(x)=(x,-xx^\top)$, and $\langle\theta,\theta'\rangle:=\theta_v^\top\theta_v'+\mathrm{tr}\left(\theta_M'^\top\theta_M\right)$

- Cumulant $F_\theta(\theta)=\frac{1}{2}\left(d\log\pi-\log|\theta_M|+\frac{1}{2}\theta_v^\top\theta_M^{-1}\theta_v\right)$

- Moment parameters $\eta=(\eta_v,\eta_M)=E[t(x)]=\nabla F(\theta)$:

$$\left\{\begin{array}{l}\eta_v(\theta)=\frac{1}{2}\theta_M^{-1}\theta_v\\ \eta_M(\theta)=-\frac{1}{2}\theta_M^{-1}-\frac{1}{4}(\theta_M^{-1}\theta_v)(\theta_M^{-1}\theta_v)^\top\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\theta_v(\eta)=-(\eta_M+\eta_v\eta_v^\top)^{-1}\eta_v\\ \theta_M(\eta)=-\frac{1}{2}(\eta_M+\eta_v\eta_v^\top)^{-1}\end{array}\right.$$

- Legendre convex conjugate $F_\eta^*(\eta)=-\frac{1}{2}\left(\log(1+\eta_v^\top\eta_M^{-1}\eta_v)+\log|-\eta_M|+d(1+\log 2\pi)\right)$

- The Kullback-Leibler between $p_{(\mu_1,\Sigma_1)}$ and $p_{(\mu_2,\Sigma_2)}$ (with $\Delta_\mu=\mu_2-\mu_1$) is

$$\mathrm{KL}[p_{(\mu_1,\Sigma_1)}:p_{(\mu_2,\Sigma_2)}] = \boxed{\frac{1}{2}\left\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1)+\Delta_\mu^\top\Sigma_2^{-1}\Delta_\mu+\log\frac{|\Sigma_2|}{|\Sigma_1|}-d\right\}} = \mathrm{KL}(p_{\lambda_1}:p_{\lambda_2}),$$
$$= B_F(\theta_2:\theta_1)=B_{F^*}(\eta_1:\eta_2)=A_F(\theta_2:\eta_1)=A_{F^*}(\eta_1:\theta_2)$$

- Bregman divergence B_F and canonical divergence A_F:

$$B_F(\theta:\theta') := F(\theta)-F(\theta')-\langle\theta-\theta',\nabla F(\theta')\rangle$$
$$A_F(\theta_1:\eta_2) := F(\theta_1)+F^*(\eta_2)-\langle\theta_1,\eta_2\rangle = A_{F^*}(\eta_2:\theta_1)$$

G-mixture of Gaussians: Normalization coefficient

- For the Gaussian family, we have

$$p_\theta(x;(\theta_1\theta_2)_\alpha) = \frac{p_\theta(x,\theta_1)^{1-\alpha} p_\theta(x,\theta_2)^\alpha}{Z_\alpha^G(p_{\theta_1} : p_{\theta_2})},$$

with the scaling normalization factor:

$$Z_\alpha^G(p_{\theta_1} : p_{\theta_2}) = \exp(-J_F^\alpha(\theta_1 : \theta_2)) = \frac{p_\theta(0;\theta_1)^{1-\alpha} p_\theta(0;\theta_2)^\alpha}{p_\theta(0;(\theta_1\theta_2)_\alpha)}.$$

- ... since $p_\theta(0;\theta) = \exp(-F(\theta))$ provided that $\langle t(0), \theta\rangle = 0$. Holds for Gaussians, $t(x) = (x, -xx^\top)$ (i.e., $t(0) = 0$)

G-Jensen-Shannon divergences between Gaussians

Given two multivariate Gaussians $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$:

$$\begin{aligned}
\mathrm{JS}^{G_\alpha}[p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}] &= (1-\alpha)\mathrm{KL}[p_{(\mu_1,\Sigma_1)} : p_{(\mu_\alpha,\Sigma_\alpha)}] + \alpha\mathrm{KL}[p_{(\mu_2,\Sigma_2)} : p_{(\mu_\alpha,\Sigma_\alpha)}] \\
&= (1-\alpha)B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2), \\
&= \frac{1}{2}\Bigg(\mathrm{tr}\left(\Sigma_\alpha^{-1}((1-\alpha)\Sigma_1 + \alpha\Sigma_2)\right) + \log\frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha} + \\
&\quad (1-\alpha)(\mu_\alpha-\mu_1)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_1) + \alpha(\mu_\alpha-\mu_2)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_2) - d\Bigg) \\
\mathrm{JS}_*^{G_\alpha}[p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}] &= (1-\alpha)\mathrm{KL}[p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_1,\Sigma_1)}] + \alpha\mathrm{KL}[p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_2,\Sigma_2)}], \\
&= (1-\alpha)B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha), \\
&= J_F(\theta_1 : \theta_2), \\
&= \frac{1}{2}\left((1-\alpha)\mu_1^\top\Sigma_1^{-1}\mu_1 + \alpha\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}{|\Sigma_\alpha|}\right) \\
\Sigma_\alpha &= (\Sigma_1\Sigma_2)_\alpha^\Sigma = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1} \\
\mu_\alpha &= (\mu_1\mu_2)_\alpha^\mu = \Sigma_\alpha\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right)
\end{aligned}$$

The JS-symmetrization of the reverse Kullback-Leibler divergence between densities of the same exponential family amount to calculate a Jensen/Burbea-Rao divergence between the corresponding natural parameters ($\to$ Bhattacharyya distance).

Example 2: Harmonic Jensen-Shannon divergence between scale Cauchy densities

- Well-suited for the *scale family* $\mathcal{C}$ of Cauchy probability distributions [9]:

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

  where γ denotes the scale and $p_{\mathrm{std}}(x) = \frac{1}{\pi(1+x^2)}$ the *standard Cauchy distribution*.

- H-mixture of Cauchy densities:

$$(p_{\gamma_1} p_{\gamma_2})^H_{\frac{1}{2}}(x) = \frac{H_\alpha(p_{\gamma_1}(x) : p_{\gamma_2}(x))}{Z^H_\alpha(\gamma_1, \gamma_2)} = p_{(\gamma_1\gamma_2)_\alpha}$$

  where the normalizing coefficient is

$$Z^H_\alpha(\gamma_1, \gamma_2) := \sqrt{\frac{\gamma_1\gamma_2}{(\gamma_1\gamma_2)_\alpha (\gamma_1\gamma_2)_{1-\alpha}}} = \sqrt{\frac{\gamma_1\gamma_2}{(\gamma_1\gamma_2)_\alpha (\gamma_2\gamma_1)_\alpha}},$$

  since we have $(\gamma_1\gamma_2)_{1-\alpha} = (\gamma_2\gamma_1)_\alpha$.

H-Jensen-Shannon divergence between scale Cauchy densities

- KLD between scale Cauchy densities:

$$\mathrm{KL}[p_{\gamma_1} : p_{\gamma_2}] = 2 \log \frac{A(\gamma_1, \gamma_2)}{G(\gamma_1, \gamma_2)} = 2 \log \frac{\gamma_1 + \gamma_2}{2\sqrt{\gamma_1 \gamma_2}}$$

- KLD is symmetric between Cauchy densities

- The harmonic Jensen-Shannon divergence between two scale Cauchy distributions p_{γ_1} and p_{γ_2} is

$$\mathrm{JS}^H[p_{\gamma_1} : p_{\gamma_2}] = \log \frac{(3\gamma_1 + \gamma_2)(3\gamma_2 + \gamma_1)}{8\sqrt{\gamma_1 \gamma_2}(\gamma_1 + \gamma_2)}$$

Example 3: A-mixture of mixture families

- **Mixture family** [12] in information geometry [1]:

$$\mathcal{M} := \left\{ m_\theta(x) = \left(1 - \sum_{i=1}^{D} \theta_i p_i(x)\right) p_0(x) + \sum_{i=1}^{D} \theta_i p_i(x) \ :\ \theta_i > 0, \sum_i \theta_i < 1 \right\},$$

- Mixture manifold is dually flat with canonical Bregman divergence [12] for generator $F(\theta) = -h(m_\theta)$

$$\mathrm{KL}[m_{\theta_p} : m_{\theta_q}] = B_F(\theta_p : \theta_q)$$

- A-mixture belongs to $\mathcal{M}$ since $\frac{m_{\theta_p} + m_{\theta_q}}{2} = m_{\frac{\theta_p + \theta_q}{2}}$

- A-Jensen-Shannon divergence between mixture members:

$$\mathrm{JS}[m_{\theta_p}, m_{\theta_q})] = \frac{1}{2}\left(B_F\left(\theta_p : \frac{\theta_p + \theta_q}{2}\right) + B_F\left(\theta_q : \frac{\theta_p + \theta_q}{2}\right)\right).$$

This amounts to calculate the **Jensen divergence** (from JBD):

$$\mathrm{JS}(m_{\theta_p}, m_{\theta_q}) = J_F(\theta_1; \theta_2) = (F(\theta_1)F(\theta_2))_{\frac{1}{2}} - F((\theta_1\theta_2)_{\frac{1}{2}})$$

Summary: Motivations and contributions

- Jensen-Shannon divergence (JSD) is a symmetrization of the Kullback-Leibler divergence always upper bounded by $\log 2$

- However, JSD does not admit a closed-form between Gaussian densities

- Introduce abstract means M to define statistical M-mixtures and statistical M-Jensen-Shannon divergences

$$\mathrm{JS}_D^{M_\alpha}[p_1 : p_2] = (1-\alpha)(D(p_1 : (p_1 p_2)_\alpha^M) + \alpha D(p_2 : (p_1 p_2)_\alpha^M))$$

- Report closed-form expressions for (i) the G-JSD between multivariate Gaussians, (ii) the H-JSD between scale Cauchy densities, and (iii) the A-JSD between mixture densities.

- $\mathrm{JS}_D^{M_\alpha}$ is upper bounded by $\log \frac{Z_\alpha^M(p,q)}{1-\alpha}$ when $M \geq A$ (and we have $A \geq G \geq H$). Thus this fails for G and H.

Thank you!

https://franknielsen.github.io/M-JS/

References I

Shun-ichi Amari.
Information geometry and its applications.
Springer, 2016.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh.
Clustering with Bregman divergences.
Journal of machine learning research, 6(Oct):1705–1749, 2005.

Thomas M. Cover and Joy A. Thomas.
Elements of information theory.
John Wiley & Sons, 2012.

Imre Csiszár.
Information-type measures of difference of probability distributions and indirect observation.
studia scientiarum Mathematicarum Hungarica, 2:229–318, 1967.

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.

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

Constantin Niculescu and Lars-Erik Persson.
Convex functions and their applications.
Springer, 2018.
Second Edition.

References II

Frank Nielsen.
Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms.
IEEE Signal Processing Letters, 20(7):657–660, 2013.

Frank Nielsen.
Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means.
Pattern Recognition Letters, 42:25–34, 2014.

Frank Nielsen and Sylvain Boltz.
The Burbea-Rao and Bhattacharyya centroids.
IEEE Transactions on Information Theory, 57(8):5455–5466, 2011.

Frank Nielsen and Vincent Garcia.
Statistical exponential families: A digest with flash cards.
arXiv preprint arXiv:0911.4863, 2009.

Frank Nielsen and Richard Nock.
On the geometry of mixtures of prescribed distributions.
In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2861–2865. IEEE, 2018.

Frank Nielsen and Ke Sun.
Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities.
Entropy, 18(12):442, 2016.

Sumio Watanabe, Keisuke Yamazaki, and Miki Aoyagi.
Kullback information of normal mixture is not an analytic function.
IEICE technical report. Neurocomputing, 104(225):41–46, 2004.