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

Frank Nielsen

Sony Computer Science Laboratories, Inc

Sony CSL

https://franknielsen.github.io/

Paper: https://www.mdpi.com/1099-4300/21/5/485

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

July 2020

Unbounded Kullback-Leibler divergence (KLD)

$\mathrm{KL}:\mathcal{P}\times\mathcal{P}\rightarrow[0,\infty]$

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

$P,Q\ll\mu$

Also called **relative entropy**:

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

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

Shannon's entropy: (self cross-entropy) $$h(p):=\int p\log\frac{1}{p}\mathrm{d}\mu=h_\times(p:p),$$

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

Symmetrizations of the KLD

Jeffreys' divergence (twice the arithmetic mean of oriented KLDs):

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

Resistor average divergence (harmonic mean of forward+reverse KLD)

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

Question: Role and extensions of the mean?

Bounded Jensen-Shannon divergence (JSD)

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

$$= \frac{1}{2}\int \left(p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q}\right) \mathrm{d}\mu.$$

$$\boxed{\mathrm{JS}(p;q) = h\left(\frac{p+q}{2}\right) - \frac{h(p)+h(q)}{2}}$$

(Shannon entropy h is strictly concave, JSD>=0)

JSD is **bounded**: $0 \leq \mathrm{JS}(p:q) \leq \log 2$

Proof: $\mathrm{KL}\left(p : \frac{p+q}{2}\right) = \int p \log \frac{2p}{p+q} \mathrm{d}\mu \leq \int p \log \frac{2p}{p} \mathrm{d}\mu = \log 2.$

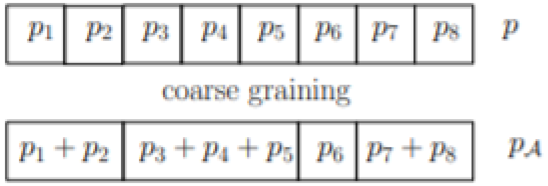
$\sqrt{\mathrm{JS}}$: Square root of the JSD is a **metric distance** (moreover Hilbertian)

Invariant f-divergences, symmetrized f-divergences

Convex generator f, strictly convex at 1
with f(1)=0   (standard when f'(1)=0, f''(1)=1)

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

f-divergences are said **invariant** in *information geometry* because they satisfy **coarse-graining** (data processing inequality)



$$\boxed{D(\theta_{\bar{A}} : \theta'_{\bar{A}}) \leq D(\theta : \theta')}$$

f-divergences can always be symmetrized: **Reverse f-divergence** for $f^*(x) = xf(\frac{1}{x})$

Jeffreys f-generator: $f_J(u) \quad := \quad (u-1)\log u,$

Jensen-Shannon f-generator: $f_{\mathrm{JS}}(u) \quad := \quad -(u+1)\log\frac{1+u}{2} + u\log u.$

Statistical distances vs parameter vector distances

A **statistical distance D** between two parametric distributions of a same family (eg., Gaussian family) amount to a **parameter distance P**:

$$\boxed{P(\theta : \theta') := D(p_\theta : p_{\theta'})}$$

For example, the KLD between two densities of a same exponential family amounts to a **reverse Bregman divergence** for the *Bregman cumulant generator*:

$$\boxed{\mathrm{KL}(p_\theta : p_{\theta'}) = B_F^*(\theta : \theta')} = B_F(\theta' : \theta).$$

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

From a smooth C3 parameter distance (=contrast function),
we can build a dualistic information-geometric structure

Skewed Jensen-Bregman divergences

JS-kind symmetrization of the *parameter Bregman divergence*:

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

Notation for the **linear interpolation**: $(\theta_p\theta_q)_\alpha := (1-\alpha)\theta_p+\alpha\theta_q$

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

J-Symmetrization and JS-Symmetrization

J-symmetrization of a statistical/parameter distance D:

$$\boxed{J_D^\alpha(p:q) := (1-\alpha)D(p:q) + \alpha D(q:p)} \qquad \alpha \in [0,1]$$

JS-symmetrization of a statistical/parameter distance D:

$$\begin{aligned} \mathrm{JS}_D^\alpha(p:q) &:= (1-\alpha)D(p:(1-\alpha)p+\alpha q) + \alpha D(q:(1-\alpha)p+\alpha q) \\ &= (1-\alpha)D(p:(pq)_\alpha) + \alpha D(q:(pq)_\alpha). \end{aligned}$$

Example: J-symmetrization and JS-symmetrization of f-divergences:

$$f_\alpha^J(u) = (1-\alpha)f(u) + \alpha f^\diamond(u),$$

$$I_{f^\diamond}(p:q) = I_f^*(p:q) = I_f(q:p)$$

$$I_f^\alpha(p:q) := (1-\alpha)I_f(p:(pq)_\alpha) + \alpha I_f(q:(pq)_\alpha)$$

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

Conjugate f-generator:

$$f^\diamond(u) = uf\left(\tfrac{1}{u}\right)$$

Generalized Jensen-Shannon divergences: Role of abstract weighted means, generalized mixtures

Quasi-arithmetic weighted means for a strictly increasing function h:

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

Definition 1 (*M*-mixture). *The* M_α-interpolation $(pq)_\alpha^M$ *(with* $\alpha \in [0,1]$*) of densities p and q with respect to a mean M is a α-weighted M-mixture defined by:*

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

where

$$Z_\alpha^M(p:q) = \int_{t\in\mathcal{X}} M_\alpha(p(t),q(t))\mathrm{d}\mu(t) =: \langle M_\alpha(p,q)\rangle. \tag{45}$$

is the normalizer function (or scaling factor) ensuring that $(pq)_\alpha^M \in \mathcal{P}$. *(The bracket notation* $\langle f \rangle$ *denotes the integral of f over* $\mathcal{X}$.*)*

When M=A Arithmetic mean, Normalizer Z is 1

Definitions: M-JSD and M-JS symmetrizations

Definition 2 (*M*-Jensen–Shannon divergence). *For a mean M, the skew M-Jensen–Shannon divergence (for $\alpha \in [0,1]$) is defined by*

$$\boxed{\mathrm{JS}^{M_\alpha}(p:q) := (1-\alpha)\mathrm{KL}\left(p:(pq)^M_\alpha\right) + \alpha \mathrm{KL}\left(q:(pq)^M_\alpha\right)} \tag{48}$$

When $M_\alpha = A_\alpha$, we recover the ordinary Jensen–Shannon divergence since $A_\alpha(p:q) = (pq)_\alpha$ (and $Z^A_\alpha(p:q) = 1$).

We can extend the definition to the JS-symmetrization of any distance:

For generic distance D (not necessarily KLD):

Definition 3 (*M*-JS symmetrization). *For a mean M and a distance D, the skew M-JS symmetrization of D (for $\alpha \in [0,1]$) is defined by*

$$\boxed{\mathrm{JS}^{M_\alpha}_D(p:q) := (1-\alpha)D\left(p:(pq)^M_\alpha\right) + \alpha D\left(q:(pq)^M_\alpha\right)} \tag{49}$$

Generic definition: (M,N)-JS symmetrization

Consider two **abstract means** M and N:

Definition 5 (Skew (M,N)-D divergence). *The skew (M,N)-divergence with respect to weighted means M_α and N_β as follows:*

$$\boxed{\mathrm{JS}_D^{M_\alpha,N_\beta}(p:q) := N_\beta\left(D\left(p:(pq)_\alpha^M\right), D\left(q:(pq)_\alpha^M\right)\right)} \tag{61}$$

The main advantage of (M,N)-JSD is to get closed-form formula
for distributions belonging to given parametric families
by carefully choosing the M-mean.
For example, *geometric mean for exponential families*,
or *harmonic mean for Cauchy or t-Student families*, etc.

(A,G)-Jensen-Shannon divergence for exponential families

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

Natural parameter space: $\Theta = \left\{ \theta : \int_{\mathcal{X}} \exp(\theta^\top x)\mathrm{d}\mu < \infty \right\}$

Geometric statistical mixture:

$$\boxed{\forall x \in \mathcal{X}, \quad (p_{\theta_1} p_{\theta_2})_\alpha^G(x) \quad := \quad \frac{G_\alpha(p_{\theta_1}(x), p_{\theta_2}(x))}{\int G_\alpha(p_{\theta_1}(t), p_{\theta_2}(t))\mathrm{d}\mu(t)} = \frac{p_{\theta_1}^{1-\alpha}(x) p_{\theta_2}^{\alpha}(x)}{Z_\alpha^G(p:q)},}$$

Normalization coefficient: $Z_\alpha^G(p:q) = \exp(-J_F^\alpha(\theta_1 : \theta_2)),$

Jensen parameter divergence: $J_F^\alpha(\theta_1 : \theta_2) := (F(\theta_1)F(\theta_2))_\alpha - F((\theta_1\theta_2)_\alpha).$

(A,G)-Jensen-Shannon divergence for exponential families

Closed-form formula the KLD between two geometric mixtures in term of a Bregman divergence between interpolated parameters:

$$
\begin{aligned}
\mathrm{KL}\left(p_\theta : (p_{\theta_1} p_{\theta_2})_\alpha^G\right) &= \mathrm{KL}\left(p_\theta : p_{(\theta_1\theta_2)_\alpha}\right), \\
&= B_F((\theta_1\theta_2)_\alpha : \theta).
\end{aligned}
$$

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

Theorem 2 (G-JSD and its dual JS-symmetrization in exponential families). *The α-skew G-Jensen–Shannon divergence JS^{G_α} between two distributions p_{θ_1} and p_{θ_2} of the same exponential family $\mathcal{E}_F$ is expressed in closed form for $\alpha \in (0,1)$ as:*

$$
\mathrm{JS}^{G_\alpha}(p_{\theta_1} : p_{\theta_2}) = (1-\alpha) B_F\left((\theta_1\theta_2)_\alpha : \theta_1\right) + \alpha B_F\left((\theta_1\theta_2)_\alpha : \theta_2\right), \tag{80}
$$

$$
\mathrm{JS}^{G_\alpha}_{\mathrm{KL}^*}(p_{\theta_1} : p_{\theta_2}) = \mathrm{JB}_F^\alpha(\theta_1 : \theta_2) = J_F^\alpha(\theta_1 : \theta_2). \tag{81}
$$

Example: Multivariate Gaussian exponential family

Family of Normal distributions: $\{N(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \succ 0\}$. $\lambda := (\lambda_v, \lambda_M) = (\mu, \Sigma)$

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

Canonical factorization: $p_\theta(x;\theta) := \exp\left(\langle t(x), \theta\rangle - F_\theta(\theta)\right) = p_\lambda(x;\lambda(\theta)),$

$$\theta = (\theta_v, \theta_M) = \left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right) = \theta(\lambda) = \left(\lambda_M^{-1}\lambda_v, -\frac{1}{2}\lambda_M^{-1}\right)$$

Sufficient statistics: $t(x) = (x, -xx^\top)$

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

$$F_\lambda(\lambda) = \frac{1}{2}\left(\lambda_v^\top\lambda_M^{-1}\lambda_v + \log|\lambda_M| + d\log 2\pi\right) = \frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu + \log|\Sigma| + d\log 2\pi\right).$$

Example: Multivariate Gaussian exponential family

Dual moment parameterization: $\eta = (\eta_v, \eta_M) = E[t(x)] = \nabla F(\theta)$

Conversions between ordinary/natural/expectation parameters:

$$\left\{\begin{array}{l} \theta_v(\lambda) = \lambda_M^{-1}\lambda_v = \Sigma^{-1}\mu \\ \theta_M(\lambda) = \frac{1}{2}\lambda_M^{-1} = \frac{1}{2}\Sigma^{-1} \end{array}\right. \Leftrightarrow \left\{\begin{array}{l} \lambda_v(\theta) = \frac{1}{2}\theta_M^{-1}\theta_v = \mu \\ \lambda_M(\theta) = \frac{1}{2}\theta_M^{-1} = \Sigma \end{array}\right.$$

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

$$\left\{\begin{array}{l} \lambda_v(\eta) = \eta_v = \mu \\ \lambda_M(\eta) = -\eta_M - \eta_v\eta_v^\top = \Sigma \end{array}\right. \Leftrightarrow \left\{\begin{array}{l} \eta_v(\lambda) = \lambda_v = \mu \\ \eta_M(\lambda) = -\lambda_M - \lambda_v\lambda_v^\top = -\Sigma - \mu\mu^\top \end{array}\right.$$

Dual potential function (=negative differential Shannon entropy):

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

Corollary 1 (G-JSD between Gaussians). *The skew G-Jensen–Shannon divergence* JS^G_α *and the dual skew G-Jensen–Shannon divergence* JS^{*G}_α *between two multivariate Gaussians* $N(\mu_1, \Sigma_1)$ *and* $N(\mu_2, \Sigma_2)$ *is*

$$\begin{aligned}
\mathrm{JS}^{G_\alpha}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_\alpha,\Sigma_\alpha)}) + \alpha\mathrm{KL}(p_{(\mu_2,\Sigma_2)} : p_{(\mu_\alpha,\Sigma_\alpha)}), && (106)\\
&= (1-\alpha)B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2), && (107)\\
&= \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_\alpha^{-1}((1-\alpha)\Sigma_1 + \alpha\Sigma_2)\right) + \log\frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha} + \right.\\
&\quad \left.(1-\alpha)(\mu_\alpha-\mu_1)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_1) + \alpha(\mu_\alpha-\mu_2)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_2) - d\right) && (108)
\end{aligned}$$

$$\begin{aligned}
\mathrm{JS}^{G_\alpha}_*(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_1,\Sigma_1)}) + \alpha\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_2,\Sigma_2)}), && (109)\\
&= (1-\alpha)B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha), && (110)\\
&= J_F(\theta_1 : \theta_2), && (111)\\
&= \frac{1}{2}\left((1-\alpha)\mu_1^\top\Sigma_1^{-1}\mu_1 + \alpha\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}{|\Sigma_\alpha|}\right), && (112)
\end{aligned}$$

where

$$\Sigma_\alpha = (\Sigma_1\Sigma_2)^\Sigma_\alpha = \left((1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}\right)^{-1}, \tag{113}$$

(matrix harmonic barycenter) and

$$\mu_\alpha = (\mu_1\mu_2)^\mu_\alpha = \Sigma_\alpha\left((1-\alpha)\Sigma_1^{-1}\mu_1 + \alpha\Sigma_2^{-1}\mu_2\right). \tag{114}$$

More examples: Abstract means and M-mixtures

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	$P_\alpha^p(x,y) = ((1-\alpha)x^p + \alpha y^p)^{\frac{1}{p}}, \quad p \in \mathbb{R}\backslash\{0\}, \lim_{p\to 0} P_\alpha^p = G$
Quasi-arithmetic mean	$M_\alpha^f(x,y) = f^{-1}((1-\alpha)f(x) + \alpha f(y))$, f strictly monotonous
M-mixture	$Z_\alpha^M(p,q) = \int_{t\in\mathcal{X}} M_\alpha(p(t),q(t))\mathrm{d}\mu(t)$
	with $Z_\alpha^M(p,q) = \int_{t\in\mathcal{X}} M_\alpha(p(t),q(t))\mathrm{d}\mu(t)$

JS^{M_α}	**Mean M**	**Parametric Family**	$Z_\alpha^M(p:q)$
JS^{A_α}	arithmetic A	mixture family	$Z_\alpha^M(\theta_1:\theta_2) = 1$
JS^{G_α}	geometric G	exponential family	$Z_\alpha^G(\theta_1:\theta_2) = \exp(-J_F^\alpha(\theta_1:\theta_2))$
JS^{H_α}	harmonic H	Cauchy scale family	$Z_\alpha^H(\theta_1:\theta_2) = \sqrt{\frac{\theta_1\theta_2}{(\theta_1\theta_2)_\alpha(\theta_1\theta_2)_{1-\alpha}}}$

https://www.mdpi.com/1099-4300/21/5/485

Summary: Generalized Jensen-Shannon divergences

- Jensen-Shannon divergence (JSD) is a **bounded symmetrization** of the Kullback-Leibler divergence (KLD). Jeffreys divergence (JD) is an unbounded symmetrization of KLD. Both JSD and JD are invariant f-divergences.

- Although KLD and JD between Gaussians (or densities of a same exponential family) admits closed-form formulas, the JSD between Gaussians does not have a closed expression, and these distances need to be **approximated** in applications.

  (machine learning, eg., deep learning in GANs)

- The skewed Jensen-Shannon divergence is based on statistical arithmetic mixtures. We define generic **statistical M-mixtures** based on an abstract mean, and define accordingly the **M-Jensen-Shannon divergence**, and the (M,N)-JSD.

- When M=G is the **geometric weighted mean**, we obtain closed-form formula for the **G-Jensen-Shannon divergence** between **Gaussian distributions**. Applications to machine learning (eg, deep learning GANs) https://arxiv.org/abs/2006.10599

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