A family of statistical symmetric divergences based on Jensen's inequality

Frank Nielsen
Ecole Polytechnique
Sony Computer Science Laboratories, Inc.

September 2010

Abstract

We introduce a novel parametric family of symmetric information-theoretic distances based on Jensen's inequality on a convex generator that unifies Jeffreys divergence with Jensen-Shannon divergence for the Shannon entropy generator. We then design a generic algorithm to compute the unique centroid defined as the minimum average divergence. This yields a smooth family of centroids linking the Jeffreys to the Jensen-Shannon centroid.

1 Introduction

The Shannon entropy [4] of a probability distribution p measures the amount of uncertainty:

$$H(p) = \int p(x) \log \frac{1}{p(x)} \mathrm{d}x = -\int p(x) \log p(x) \mathrm{d}x. \tag{1}$$

The cross-entropy [4] measures the amount of extra bits required to compute a code based on an observed empirical probability $\tilde{p}$ instead of the true probability (hidden by nature):

$$H(p : \tilde{p}) = \int p(x) \log \frac{1}{\tilde{p}(x)} \mathrm{d}x = -\int p(x) \log \tilde{p}(x) \mathrm{d}x. \tag{2}$$

The ":" notation emphasizes on the oriented aspect [4] of the functional: $H(p : q) \neq H(q : p)$. The Kullback-Leibler divergence [9, 4] is a statistical distance measure computing the *relative entropy* as follows:

$$\begin{aligned} \mathrm{KL}(p : q) &= \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x & (3)\\ &= H(p : q) - H(p) \geq 0, & (4) \end{aligned}$$

This last inequality is called Gibb's inequality [4], with equality if and only if $p = q$. We have $H(p : q) = H(p) + \mathrm{KL}(p : q)$. The Kullback-Leibler divergence can be extended to unnormalized positive distributions (or positive arrays) as follows:

$$\begin{aligned} \mathrm{eKL}(p : q) &= \int \left(p(x) \log \frac{p(x)}{q(x)} + q(x) - p(x) \right) \mathrm{d}x, & (5)\\ &= \mathrm{eH}(p : q) - \mathrm{eH}(p) \geq 0, & (6) \end{aligned}$$

with $\mathrm{eH}(p : q) = \int (p(x) \log \frac{1}{q(x)} + q(x)) \mathrm{d}x$ and $\mathrm{eH}(p) = \mathrm{eH}(p, p)$.

(Rényi based on an axiomatic approach [13] derived yet another expression for the Kullback-Leibler divergence of unnormalized generalized distributions.)

Many applications in information retrieval (IR) requires to deal with a symmetric distortion measure. Jeffreys divergence [7] (also called J-divergence) symmetrizes the oriented Kullback-Leibler divergence as follows:

$$
\begin{aligned}
J(p,q) &= \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = J(q,p) && (7)\\
&= H(p:q) + H(q:p) - (H(p)+H(q)), && (8)\\
&= \int (p(x)-q(x)) \log \frac{p(x)}{q(x)} \mathrm{d}x. && (9)
\end{aligned}
$$

Here, we replaced ":" by "," in the distortion measure to emphasize the symmetric property: $J(p,q) = J(q,p)$. Jeffreys divergence is interpreted as *twice the average of the cross-entropies minus the average of the entropies*. One of the drawbacks of Jeffreys divergence is that it may be unbounded and therefore numerically quite unstable to compute in practice: For example, let $p = (p_i)_{i=1}^d$ and $q = (q_i)_{i=1}^d$ be frequency histograms with d bins, then $J(p,q) \to \infty$ if there exists one bin $l \in \{1,...,d\}$ such that p_l is above some constant, and $q_l \to 0$. In that case, $p_l \log \frac{p_l}{q_l} \to \infty$. To circumvent this unboundedness problem, the Jensen-Shannon divergence was introduced in [10]. The Jensen-Shannon divergence symmetrizes the Kullback-Leibler divergence by taking the *average relative entropy of the source distributions to the average distribution* $\frac{p+q}{2}$:

$$
\begin{aligned}
\mathrm{JS}(p,q) &= \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right) = \mathrm{JS}(q,p) && (10)\\
&= \frac{1}{2}\left(H\left(p : \frac{p+q}{2}\right) - H(p) + H\left(q : \frac{p+q}{2}\right) - H(q)\right), && (11)\\
&= \frac{1}{2}\int \left(p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q}\right) \mathrm{d}x, && (12)\\
&= H\left(\frac{p+q}{2}\right) - \frac{H(p)+H(q)}{2} \geq 0. && (13)
\end{aligned}
$$

The Jensen-Shannon divergence has always finite value, and its square root yields a metric, satisfying the triangular inequality. Moreover, we have the following information-theoretic inequality [10]

$$
0 \leq \mathrm{JS}(p,q) \leq \frac{1}{4} J(p,q). \quad (14)
$$

By introducing the K-divergence [10] (see Eq. 7):

$$
K(p:q) = \int p(x) \log \frac{2p(x)}{p(x)+q(x)} \mathrm{d}x = \mathrm{KL}\left(p : \frac{p+q}{2}\right), \quad (15)
$$

we interpret the Jensen-Shannon divergence as the Jeffreys symmetrization of the K-divergence (see Eq. 7).

$$
\begin{aligned}
\mathrm{JS}(p,q) &= \frac{1}{2}(K(p:q) + K(q:p)), && (16)\\
&= H\left(\frac{p+q}{2}\right) - \frac{H(p)+H(q)}{2}. && (17)
\end{aligned}
$$

The Jensen-Shannon divergence is also widely used in earth sciences as a *diversity index*. Indeed, the basic two-point measure can further be generalized to a *population* as follows:

$$
\mathrm{JS}(p_1,...,p_n;w) = H\left(\sum_{i=1}^{n} w_i p_i\right) - \sum_{i=1}^{n} w_i H(p_i), \tag{18}
$$

for a given normalized unit positive weight vector w.

Let P be a random variable following density p with associated weight distribution w ($W \sim w$), then the Jensen-Shannon divergence can be defined as

$$
\begin{aligned}
\mathrm{JS}(P;W) &= H\left(\int w(x)p(x)\mathrm{d}x\right) - \int w(x)H(p(x))\mathrm{d}x, &(19)\\
&= H(E_W[P]) - E_W[H(P)], &(20)
\end{aligned}
$$

where $E_W[H(P)] = \int w(x)H(p(x))\mathrm{d}x$ denote the expectation of the entropy with respect to the weight distribution. Since $H(x)$ is a concave function, it follows from Jensen inequality that $\mathrm{JS}(P;W) \geq 0$.

Consider

$$
K_\alpha(p:q) = p\log\frac{p}{(1-\alpha)p+\alpha q}, \tag{21}
$$

and its symmetrized divergence

$$
\mathrm{JS}_\alpha(p,q) = \frac{K_\alpha(p:q)+K_\alpha(q:p)}{2} = \mathrm{JS}_\alpha(q,p). \tag{22}
$$

For $\alpha = \frac{1}{2}$, we find the Jensen-Shannon divergence: $\mathrm{JS}(p,q) = \mathrm{JS}_{\frac{1}{2}}(p,q)$. For $\alpha = 1$, we obtain half of Jeffreys divergence: $\mathrm{JS}_1(p,q) = \frac{1}{2}J(p,q)$. It turns out that this family of α-Jensen-Shannon divergence belongs to a broader family of information-theoretic measures, called Ali-Silvey-Csiszár divergences [5, 1]. A ϕ-divergence is defined for a strictly convex function ϕ such that $\phi(1) = 0$ as:

$$
I_\phi(p:q) = \int q(x)\phi\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x. \tag{23}
$$

We can always symmetrize ϕ-divergences by taking the *coupled* function $\phi^*(x) = x\phi(\frac{1}{x})$. Indeed, we get

$$
\begin{aligned}
I_{\phi^*}(p:q) &= \int q(x)\phi^*\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x, &(24)\\
&= \int q(x)\frac{p(x)}{q(x)}\phi\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x, &(25)\\
&= \int p(x)\phi\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x = I_\phi(q:p). &(26)
\end{aligned}
$$

Therefore, $I_{\phi+\phi^*}(p,q)$ is a symmetric divergence. Let $\phi^s = \phi + \phi^*$ denote the symmetrized generator. Jeffreys divergence is a ϕ-divergence for $\phi(u) = -\log u$ (and $\phi^s(u) = (u-1)\log u$). Similarly, Jensen-Shannon divergence is interpreted as $\mathrm{JS}(p,q) = \frac{1}{2}(K(p:q)+K(q:p))$, with $\frac{1}{2}K(p:q)$ a ϕ-divergence for $\phi(u) = \frac{u}{2}\log\frac{2u}{1+u}$, see [10]. It follows that Jensen-Shannon is also a ϕ-divergence. The α-Jensen-Shannon divergences are ϕ-divergences for the generators $\phi^s_\alpha = \phi^*_\alpha + \phi_\alpha$, with $\phi^*_\alpha(x) = -\log((1-\alpha)+\alpha x)$ and $\phi_\alpha(x) = -x\log((1-\alpha)+\frac{\alpha}{x})$. α-Jensen-Shannon divergences are convex in both arguments.

One drawback for estimating α-JS divergences on *continuous* parametric densities (say, Gaussians), is that the mixture of two Gaussians is not a Gaussian, and therefore the average distribution falls outside the family of considered distributions. This explains the lack of closed-form solution for computing the Jensen-Shannon divergence on Gaussians.

Next, we introduce a novel family of symmetrized divergences which occur in the closed form equations of statistical distances of a large class of parametric distributions, called exponential families.

2 A novel parametric family of Jensen divergences

At the heart of many statistical distances lies the celebrated Jensen's convex inequality [8]. For a strictly convex function F and a parameter $\alpha \in \mathbb{R}\backslash\{0,1\}$, let us define the α-skew Jensen divergence as

$$J_F^{(\alpha)}(p:q) = \frac{1}{\alpha(1-\alpha)} \int ((1-\alpha)F(p(x)) + \alpha F(q(x)) - F((1-\alpha)p(x) + \alpha q(x))\mathrm{d}x. \tag{27}$$

In the limit cases, we find the oriented Kullback-Leibler divergences [11]:

$$\lim_{\alpha \to 0} J_F^{(\alpha)}(p:q) = \mathrm{KL}(p:q), \tag{28}$$
$$\lim_{\alpha \to 1} J_F^{(\alpha)}(p:q) = \mathrm{KL}(q:p). \tag{29}$$

Observe also that $J_F^{(\alpha)}(q:p) = J_F^{(1-\alpha)}(p:q)$, and that therefore α-skew Jensen divergences are asymmetric distortion measures (except for $\alpha = \frac{1}{2}$). Therefore, let us symmetrize those α-skew divergences by averaging the two orientations as follows:

$$\begin{aligned}
\mathrm{sJ}_F^{(\alpha)}(p,q) &= \frac{1}{2}(J_F^{(\alpha)}(p:q) + J_F^{(\alpha)}(q:p)) && (30)\\
&= \frac{1}{2}(J_F^{(\alpha)}(p:q) + J_F^{(1-\alpha)}(p:q)) && (31)\\
&= \frac{1}{2\alpha(1-\alpha)} \int (F(p(x)) + F(q(x)) \\
&\quad -F(\alpha p(x) + (1-\alpha)q(x)) - F((1-\alpha)p(x) + \alpha q(x))\,\mathrm{d}x && (32)\\
&= \mathrm{sJ}_F^{(\alpha)}(q,p) = \mathrm{sJ}_F^{(1-\alpha)}(p,q) \geq 0 && (33)
\end{aligned}$$

Figure 1 depicts this novel family of symmetric Jensen divergences (it is enough to consider $\alpha \in [0, \frac{1}{2}]$). Note that except for $\alpha \in \{0,1\}$, this family of divergences have the boundedness property: $\mathrm{sJ}_F^{(\alpha)}(p,q) < \infty, \forall \alpha \notin \{0,1\}$

Consider the strict convex generator $F(x) = x \log x$ (Shannon information). Rewriting the divergence for $F(x) = -H(x)$ (Shannon entropy is concave) the negative Shannon entropy we get a family of *symmetric Kullback-Leibler divergences*:

$$\mathrm{sKL}^{(\alpha)}(p,q) = \frac{1}{2\alpha(1-\alpha)} \left(H(\alpha p + (1-\alpha)q) + H((1-\alpha)p + \alpha q) - (H(p) + H(q))\right) \geq 0 \tag{34}$$

We have in the limit case:

$$\lim_{\alpha \to 0} \mathrm{sKL}^{(\alpha)}(p,q) = J(p,q) = \mathrm{sKL}^{(0)}(p,q). \tag{35}$$

That is, symmetrized α-Jensen divergences tend asymptotically to the Jeffreys divergence for the Shannon information generator. Furthermore, consider the case $\alpha = \frac{1}{2}$:

$$\mathrm{sKL}^{(\frac{1}{2})}(p,q) = 2\left(2H\left(\frac{p+q}{2}\right) - (H(p) + H(q))\right) = 4\mathrm{JS}(p,q). \tag{36}$$

Thus this family of symmetric Kullback-Leibler divergences unify both Jensen-Shannon divergence (up to a constant factor for $\alpha = \frac{1}{2}$) with Jeffreys divergence ($\alpha \to 0$).

Theorem 1 *There exists a parametric family of symmetric information-theoretic divergences* $\{\mathrm{sKL}^{(\alpha)}\}_\alpha$ *that unifies Jeffreys J-divergence with Jensen-Shannon divergence.*

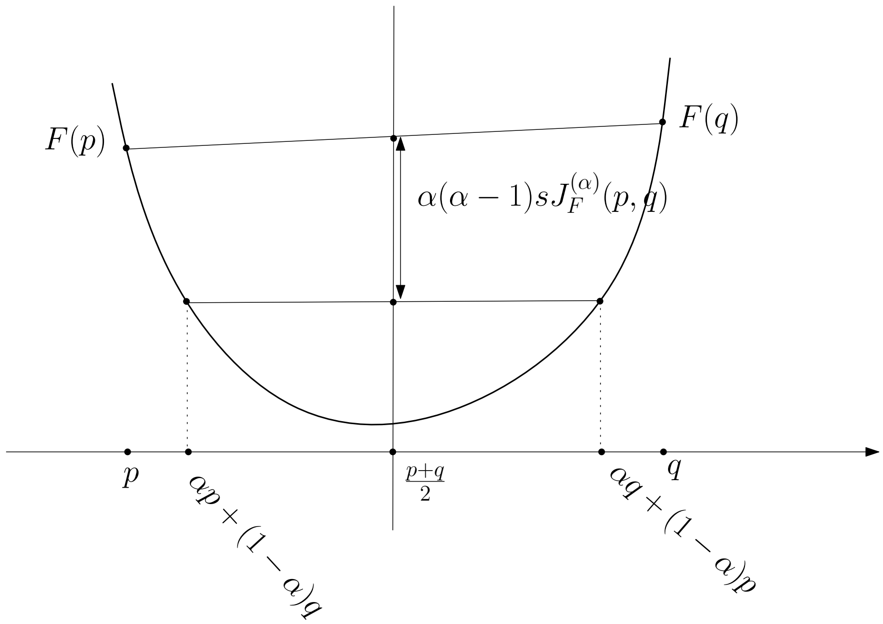

Figure 1: A family of symmetric Jensen divergences $\{\mathrm{sJ}_F^{(\alpha)}\}_\alpha$ for $\alpha \in 0, \frac{1}{2}]$ that includes both Jeffreys divergence in the limit case $\alpha = 0$ and Jensen-Shannon divergence for $\alpha = \frac{1}{2}$, for the Shannon information generator $F(x) = x \log x$.

This result can be obtained by considering skew average of distributions instead of the one-half of Eq. 15:

$$L_\alpha(p : q) \quad = \quad \frac{H((1-\alpha)p + \alpha q) - H(p)}{\alpha(1-\alpha)} \geq 0 \tag{37}$$

Then it comes out that (see Eq. 7)

$$\mathrm{sKL}^{(\alpha)}(p, q) = \frac{1}{2\alpha(1-\alpha)}(L_\alpha(p : q) + L_\alpha(q : p)). \tag{38}$$

Note that $L_{\frac{1}{2}}(p : q) = 4K(p : q)$. The scaling factor is due to historical convention. However L_α is in general not a ϕ-divergence (excepts for $\alpha \in \{0, 1\}$).

An alternative description of the symmetric family is given by

$$S_F^{(\alpha)}(p, q) = \frac{2}{1-\alpha^2}\left(F(p) + F(q) - F\left(\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q\right) - F\left(\frac{1+\alpha}{2}p + \frac{1-\alpha}{2}q\right)\right). \tag{39}$$

It can be checked that $\mathrm{sJ}_F^{(\alpha)}(p, q) = S_F^{(\alpha')}(p, q)$ for $\alpha' = 1 - 2\alpha$.

Many parametric distributions follow a regular structure called exponential families. We shall link next that class of symmetric sJ^α-divergences to equivalent symmetric α-Bhattacharrya divergences computed on the parameter space.

3 Case of exponential families

Many common statistical distributions are handled in the unified framework of exponential families [12, 11]. A distribution is said to belong to an exponential family E_F, if its *parametric* density can be canonically rewritten as

$$p_F(x;\theta) = \exp(\langle t(x), \theta\rangle - F(\theta) + k(x)), \tag{40}$$

where θ describes the member of the exponential family $E_F = \{p_F(x;\theta) \; | \theta \in \Theta\}$, characterized by the log-normalizer $F(\theta)$, a convex differentiable function. $\langle x, y\rangle$ denotes the inner-product (e.g., $x^T y$ for vectors, etc. – see [12, 11]). $t(x)$ is the sufficient statistic.

Discrete d-dimensional distributions (corresponding to frequency histograms with d bins in visual applications) are multinomials, an exponential family with the dimension of the natural space Θ being $d-1$ (the order of the family). In information retrieval, one often needs to perform clustering on frequency histograms for building a codebook to perform efficiently retrieval queries (eg., bag of words method [6]).

The Kullback-Leibler divergence of members $p \sim E_F(\theta_p)$ and $q \sim E_F(\theta_q)$ of the same exponential family E_F is equivalent to a Bregman divergence on the natural parameters [2]:

$$\mathrm{KL}(p_F(x;\theta_p) : p_F(x;\theta_q)) = B_F(\theta_q : \theta_p) \tag{41}$$

The Jeffreys J-divergence on members of the same exponential family can be computed as a symmetrized Bregman divergence, yielding a calculation on the natural parameter space:

$$J(p_F(x;\theta_p), p_F(x;\theta_q)) = (\theta_p - \theta_q)^T (\nabla F(\theta_p) - \nabla F(\theta_q)) \tag{42}$$

Note that although the product of two exponential families is an exponential family, it is *not* the case for the mixture of two exponential families. Indeed, the mixture $(1-\alpha)p + \alpha q$ does not in general belong to E_F. Therefore, the Jensen-Shannon divergence on members of the same exponential family *cannot* be computed directly from the natural parameters, since it requires to compute the entropy of the mixture distribution (with no known generic closed form):

$$\mathrm{JS}(p = p_F(x;\theta_p), q = p_F(x;\theta_q)) = H\left(\frac{p+q}{2}\right) - \frac{H(p)+H(q)}{2}, \tag{43}$$

In fact, Eq. 41 is the limit case of the property that α-skew Bhattacharrya divergence $B^{(\alpha)}$ of members $p = p_F(x;\theta_p)$ and $q = p_F(x;\theta_q)$ of the same exponential family E_F is equivalent to a α-Jensen divergence on the natural parameters [11]:

$$\begin{aligned} B^{(\alpha)}(p_F(x;\theta_p) : p_F(x;\theta_q)) &= -\log \int p_F(x;\theta_p)^\alpha p_F(x;\theta_q)^{1-\alpha} \mathrm{d}x, & (44)\\ &= J_F^{(\alpha)}(\theta_p : \theta_q) & (45) \end{aligned}$$

We can therefore symmetrize α-skew Bhattacharrya divergences:

$$\begin{aligned} \mathrm{sB}^{(\alpha)}(p_F(x;\theta_p), p_F(x;\theta_q)) &= \frac{1}{2}(B^{(\alpha)}(p_F(x;\theta_p) : p_F(x;\theta_q)) + B^{(\alpha)}(p_F(x;\theta_q) : p_F(x;\theta_p))), & (46)\\ &= -\frac{1}{2}\log\left(\int p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x\right)\left(\int p^{1-\alpha}(x) q^\alpha(x)\mathrm{d}x\right) & (47)\\ &= \alpha(1-\alpha)\mathrm{sJ}_F^{(\alpha)}(\theta_p, \theta_q), & (48) \end{aligned}$$

and obtain equivalently a symmetrized skew Jensen divergence on the natural parameters.

Theorem 2 *The symmetrized skew α-Bhattacharyya divergence on members of the same exponential family is equivalent to a symmetrized skew α-Jensen divergence defined for the log-normalizer and computed in the natural parameter space.*

Let us now consider computing centers (say, for k-means clustering applications [2]).

4 Symmetrized skew α-Jensen centroids

Consider the discrete symmetrized α-Jensen divergences (not any more on distributions but on d-dimensional parameter points). In particular, we get for separable divergences:

$$\mathrm{sJ}_F^{(\alpha)}(x,y) = \frac{1}{2\alpha(1-\alpha)} \sum_{i=1}^{d} \left(F(x_i) + F(y_i) - F(\alpha x_i + (1-\alpha)y_i) - F((1-\alpha)x_i + \alpha y_i\right). \tag{49}$$

This family of discrete measures includes the *extended Kullback-Leibler divergence* for unnormalized distributions by setting $F(x) = x \log x$. The *barycenter* b of n points $p_1, ..., p_n$ is defined as the (unique) point that minimizes the weighted average distance:

$$b = \arg\min_c \sum_{i=1}^{n} w_i \times \mathrm{sJ}_F^{(\alpha)}(p_i, c), \tag{50}$$

for $w = (w_1, ..., w_n)$ a normalized weight vector ($\forall i, w_i > 0$ and $\sum_i w_i = 1$). In particular, choosing $w_i = \frac{1}{n}$ for all i yields the *centroid*. Note that the multiplicative factor in the energy function of Eq. 50 does not impact the minimum. Thus we need to minimize:

$$\min_c E(c) = \min_c \sum_{i=1}^{n} w_i (F(p_i) + F(c) - F(\alpha p_i + (1-\alpha)c) - F(\alpha c + (1-\alpha)p_i)). \tag{51}$$

Removing the constant terms (i.e., independent of c), this amounts to minimize the following energy functional ($\sum_i w_i = 1$):

$$\min E(c) \equiv \min_c E'(c) = \min_c F(c) - \sum_i w_i (F(\alpha p_i + (1-\alpha)c) + F(\alpha c + (1-\alpha)p_i)). \tag{52}$$

Since F is convex, E is the minimization of a sum of a convex function plus a concave function. Therefore, we can apply the ConCave-Convex Procedure [14] (CCCP) that guarantees to converge to a minimum. We thus bypass using a gradient steepest descent numerical optimization that requires to tune a learning parameter.

Initializing

$$c_0 = \sum_{i=1}^{n} w_i p_i \tag{53}$$

to the Euclidean barycenter, we iteratively update as follows:

$$\nabla F(c_{t+1}) = \sum_{i=1}^{n} w_i ((1-\alpha)\nabla F(\alpha p_i + (1-\alpha)c_t) + \alpha \nabla F(\alpha c_t + (1-\alpha)p_i)), \tag{54}$$

or

$$c_{t+1} = (\nabla F)^{-1} \left(\sum_{i=1}^{n} w_i ((1-\alpha)\nabla F(\alpha p_i + (1-\alpha)c_t) + \alpha \nabla F(\alpha c_t + (1-\alpha)p_i)) \right) \tag{55}$$

(Observe that since F is strictly convex, its Hessian $\nabla^2 F$ is positive-definite, and ∇F is strictly increasing, so that ∇F^{-1} is well-defined.)

In the limit case, we get the following *fixed point* equation:

$$c^* = (\nabla F)^{-1} \left(\sum_{i=1}^{n} w_i ((1-\alpha)\nabla F(\alpha p_i + (1-\alpha)c^*) + \alpha \nabla F(\alpha c^* + (1-\alpha)p_i)) \right). \tag{56}$$

This rule is a quasi-arithmetic mean, and can alternatively be initialized using $c'_0 = \nabla F^{-1}(\sum_{i=1}^n w_i \nabla F(p_i))$. Let us instantiate this updating rule for $\alpha = \frac{1}{2}$ and $w_i = \frac{1}{n}$ on Shannon and Burg information functions:

Shannon information $F(x) = x \log x - x$ $\nabla F(x) = \log x, (\nabla F)^{-1}(x) = \exp x$	Burg information $F(x) = -\log x$ $\nabla F(x) = -1/x, (\nabla F)^{-1}(x) = -1/x$
$c_{t+1} = \left(\prod_{i=1}^n \frac{c_t + p_i}{2}\right)^{\frac{1}{n}}$ $\rightarrow$ Geometric update	$c_{t+1} = \frac{n}{\sum_{i=1}^n \frac{2}{c_t + p_i}}$ $\rightarrow$ Harmonic update

A Java(TM) source code implementing this CCCP centroid method with respect to symmetrized α-Jensen divergences is available online at:
http://www.informationgeometry.org/sJS/

Note that for Jeffreys ($\alpha = 0$) and Jensen-Shannon ($\alpha = \frac{1}{2}$) divergences, the energy function is convex, and therefore the minimum is necessarily unique. (In fact, both Jeffreys and Jensen-Shannon are two instances of the class of convex Ali-Silvey-Csiszár divergences [5, 1].)

Since α-JS divergences are ϕ-divergences (convex in both arguments), the barycenter with respect to α-JS is unique, and can be computed using any convex optimization technique. Ben-Tal et al. [3] called those center points entropic means; They consider scalar values that can be extended to dimension-wise separable divergences, but *not* to normalized nor continuous distributions.

Theorem 3 *The centroid of members of the same exponential family with respect to the symmetrized α-Bhattacharyya divergence can be computed equivalently as the centroid of their natural parameters with respect to the symmetrized α-Jensen divergence using the concave-convex procedure.*

Note that for members of the same exponential family, both c_0 or c'_0 initializations are interpreted as left-sided or right-sided Kullback-Leibler centroids [12].

5 Concluding remarks

We have introduced a novel parametric family of symmetric divergences based on Jensen's inequality called symmetrized α-skew Jensen divergences. Instantiating this family for the Shannon information generator, we have exhibited a one-parameter family of symmetrized Kullback-Leibler divergences. Furthermore, we showed that for distributions belonging to the *same* exponential family, the symmetrized α-Bhattacharyya divergence amounts to compute a symmetrized α-Jensen divergence defined on the parameter space, thus yielding a closed-form formula.

For applications requiring symmetric statistical distances, the choice is therefore not whether to decide between Jeffreys or Jensen-Shannon divergences, but rather to choose or tune the best α parameter according to the application and input data. It would be interesting to study the impact of α in the performance of information retrieval applications.

References

[1] Syed Mumtaz Ali and Samuel David Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

[2] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.

[3] Aharon Ben-Tal, Abraham Charnes, and Marc Teboulle. Entropic means. *Journal of Mathematical Analysis and Applications*, 139(2):537 – 551, 1989.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-Interscience, New York, NY, USA, 1991.

[5] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229318, 1967.

[6] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. volume 2, pages 524–531. IEEE Computer Society, June 2005.

[7] Harold Jeffreys. *Scientific Inference*. Cambridge University Press, 1973.

[8] Johan L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, December 1906. Available online from Springer.

[9] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[10] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151, 1991.

[11] Frank Nielsen and Sylvain Boltz. The Burbea-Rao and Bhattacharyya centroids. *Computing Research Repository (CoRR)*, http://arxiv.org/, April 2010.

[12] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2048–2059, June 2009.

[13] Alfred Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pages 547–561, 1961.

[14] Bharath Sriperumbudur and Gert Lanckriet. On the convergence of the concave-convex procedure. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1759–1767. 2009.