entropy

MDPI

Article

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

Frank Nielsen

Sony Computer Science Laboratories, Takanawa Muse Bldg., 3-14-13, Higashigotanda, Shinagawa-ku,
Tokyo 141-0022, Japan; Frank.Nielsen@acm.org or nielsen@lix.polytechnique.fr

Received: 10 April 2019; Accepted: 9 May 2019; Published: 11 May 2019

Abstract: The Jensen–Shannon divergence is a renowned bounded symmetrization of the unbounded Kullback–Leibler divergence which measures the total Kullback–Leibler divergence to the average mixture distribution. However, the Jensen–Shannon divergence between Gaussian distributions is not available in closed form. To bypass this problem, we present a generalization of the Jensen–Shannon (JS) divergence using abstract means which yields closed-form expressions when the mean is chosen according to the parametric family of distributions. More generally, we define the JS-symmetrizations of any distance using parameter mixtures derived from abstract means. In particular, we first show that the geometric mean is well-suited for exponential families, and report two closed-form formula for (i) the geometric Jensen–Shannon divergence between probability densities of the same exponential family; and (ii) the geometric JS-symmetrization of the reverse Kullback–Leibler divergence between probability densities of the same exponential family. As a second illustrating example, we show that the harmonic mean is well-suited for the scale Cauchy distributions, and report a closed-form formula for the harmonic Jensen–Shannon divergence between scale Cauchy distributions. Applications to clustering with respect to these novel Jensen–Shannon divergences are touched upon.

Keywords: Jensen–Shannon divergence; Jeffreys divergence; resistor average distance; Bhattacharyya distance; f-divergence; Jensen/Burbea–Rao divergence; Bregman divergence; abstract weighted mean; quasi-arithmetic mean; mixture family; statistical M-mixture; exponential family; Gaussian family; Cauchy scale family; clustering

1. Introduction and Motivations

1.1. Kullback–Leibler Divergence and Its Symmetrizations

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space [1] where $\mathcal{X}$ denotes the sample space and $\mathcal{A}$ the σ-algebra of measurable events. Consider a positive measure μ (usually the Lebesgue measure μ_L with Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ or the counting measure μ_c with power set σ-algebra $2^{\mathcal{X}}$). Denote by $\mathcal{P}$ the set of probability distributions.

The Kullback–Leibler Divergence [2] (KLD) $\mathrm{KL} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is the most fundamental distance [2] between probability distributions, defined by:

$$\mathrm{KL}(P : Q) := \int p \log \frac{p}{q} \mathrm{d}\mu, \tag{1}$$

where p and q denote the Radon–Nikodym derivatives of probability measures P and Q with respect to μ (with $P, Q \ll \mu$). The KLD expression between P and Q in Equation (1) is independent of the dominating measure μ. Table A1 summarizes the various distances and their notations used in this paper.

The KLD is also called the relative entropy [2] because it can be written as the difference of the cross-entropy minus the entropy:

$$\mathrm{KL}(p:q) = h_\times(p:q) - h(p), \tag{2}$$

where $h_\times$ denotes the cross-entropy [2]:

$$h_\times(p:q) := \int p \log \frac{1}{q} \mathrm{d}\mu, \tag{3}$$

and

$$h(p) := \int p \log \frac{1}{p} \mathrm{d}\mu = h_\times(p:p), \tag{4}$$

denotes the *Shannon entropy* [2]. Although the formula of the Shannon entropy in Equation (4) unifies both the discrete case and the continuous case of probability distributions, the behavior of entropy in the discrete case and the continuous case is very different: When $\mu = \mu_c$, Equation (4) yields the discrete Shannon entropy which is always positive and upper bounded by $\log|\mathcal{X}|$. When $\mu = \mu_L$, Equation (4) defines the Shannon *differential entropy* which may be negative and unbounded [2] (e.g., the differential entropy of the Gaussian distribution $N(m,\sigma)$ is $\frac{1}{2}\log(2\pi e\sigma^2)$). See also [3] for further important differences between the discrete case and the continuous case.

In general, the KLD is an asymmetric distance (i.e., $\mathrm{KL}(p:q) \neq \mathrm{KL}(q:p)$, hence the argument separator notation using the delimiter ':') In information theory [2], it is customary to use the double bar notation '‖' instead of the comma ',' notation to avoid confusion with joint random variables. The *reverse KL divergence* or *dual KL divergence* is:

$$\mathrm{KL}^*(P:Q) := \mathrm{KL}(Q:P) = \int q \log \frac{q}{p} \mathrm{d}\mu. \tag{5}$$

In general, the *reverse distance* or *dual distance* for a distance D is written as:

$$D^*(p:q) := D(q:p). \tag{6}$$

One way to symmetrize the KLD is to consider the *Jeffreys Divergence* [4] (JD, Sir Harold Jeffreys (1891–1989) was a British statistician.):

$$J(p;q) := \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = \int (p-q) \log \frac{p}{q} \mathrm{d}\mu = J(q;p). \tag{7}$$

However, this symmetric distance is not upper bounded, and its sensitivity can raise numerical issues in applications. Here, we used the optional argument separator notation ';' to emphasize that the distance is symmetric but not necessarily a metric distance. This notation matches the notational convention of the mutual information if two joint random variables in information theory [2].

The symmetrization of the KLD may also be obtained using the harmonic mean instead of the arithmetic mean, yielding the *resistor average distance* [5] $R(p;q)$:

$$\frac{1}{R(p;q)} = \frac{1}{2}\left(\frac{1}{\mathrm{KL}(p:q)} + \frac{1}{\mathrm{KL}(q:p)}\right), \tag{8}$$

$$R(p;q) = \frac{2\left(\mathrm{KL}(p:q) + \mathrm{KL}(q:p)\right)}{\mathrm{KL}(p:q)\mathrm{KL}(q:p)} = \frac{2J(p;q)}{\mathrm{KL}(p:q)\mathrm{KL}(q:p)}. \tag{9}$$

Another famous symmetrization of the KLD is the *Jensen–Shannon Divergence* [6] (JSD) defined by:

$$\begin{aligned}
\mathrm{JS}(p;q) &:= \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right), && (10)\\
&= \frac{1}{2}\int\left(p\log\frac{2p}{p+q} + q\log\frac{2q}{p+q}\right)\mathrm{d}\mu. && (11)
\end{aligned}$$

This distance can be interpreted as the *total divergence to the average distribution* (see Equation (10)). The JSD can be rewritten as a *Jensen divergence* (or Burbea–Rao divergence [7]) for the negentropy generator $-h$ (called Shannon information):

$$\mathrm{JS}(p;q) = h\left(\frac{p+q}{2}\right) - \frac{h(p)+h(q)}{2}. \tag{12}$$

An important property of the Jensen–Shannon divergence compared to the Jeffreys divergence is that this distance is *always* bounded:

$$0 \leq \mathrm{JS}(p : q) \leq \log 2. \tag{13}$$

This follows from the fact that

$$\mathrm{KL}\left(p : \frac{p+q}{2}\right) = \int p\log\frac{2p}{p+q}\mathrm{d}\mu \leq \int p\log\frac{2p}{p}\mathrm{d}\mu = \log 2. \tag{14}$$

Finally, the square root of the JSD (i.e., $\sqrt{\mathrm{JS}}$) yields a *metric distance* satisfying the triangular inequality [8,9]. The JSD has found applications in many fields such as bioinformatics [10] and social sciences [11], just to name a few. Recently, the JSD has gained attention in the deep learning community with the *Generative Adversarial Networks* (GANs) [12]. In computer vision and pattern recognition, one often relies on information-theoretic techniques to perform registration and recognition tasks. For example, in [13], the authors use a mixture of Principal Axes Registrations (mPAR) whose parameters are estimated by minimizing the KLD between the considered two-point distributions. In [14], the authors parameterize both shapes and deformations using Gaussian Mixture Models (GMMs) to perform non-rigid shape registration. The lack of closed-form formula for the KLD between GMMs [15] spurred the use of other statistical distances which admit a closed-form expression for GMMs. For example, in [16], shape registration is performed by using the Jensen-Rényi divergence between GMMs. See also [17] for other information-theoretic divergences that admit closed-form formula for some statistical mixtures extending GMMs.

In information geometry [18], the KLD, JD and JSD are *invariant divergences* which satisfy the property of information monotonicity [18]. The class of (separable) distances satisfying the information monotonicity are exhaustively characterized as Csiszár's f-divergences [19]. A *f-divergence* is defined for a convex generator function f strictly convex at 1 (with $f(1) = f'(1) = 0$) by:

$$I_f(p : q) = \int pf\left(\frac{q}{p}\right)\mathrm{d}\mu. \tag{15}$$

The Jeffreys and Jensen–Shannon f-generators are:

$$\begin{aligned}
f_J(u) &:= (u-1)\log u, && (16)\\
f_{\mathrm{JS}}(u) &:= -(u+1)\log\frac{1+u}{2} + u\log u. && (17)
\end{aligned}$$

1.2. Statistical Distances and Parameter Divergences

In information and probability theory, the term "divergence" informally means a *statistical distance* [2]. However in information geometry [18], a divergence has a stricter meaning of being a smooth *parametric* distance (called a contrast function in [20]) from which a dual geometric structure can be derived [21,22].

Consider parametric distributions p_θ belonging to a parametric family of distributions $\{p_\theta \,:\, \theta \in \Theta\}$ (e.g., Gaussian family or Cauchy family), where Θ denotes the parameter space. Then a statistical distance D between distributions p_θ and $p_{\theta'}$ amount to an equivalent *parameter distance*:

$$P(\theta : \theta') := D(p_\theta : p_{\theta'}). \tag{18}$$

For example, the KLD between two distributions belonging to the same exponential family (e.g., Gaussian family) amount to a reverse Bregman divergence for the cumulant generator F of the exponential family [23]:

$$\mathrm{KL}(p_\theta : p_{\theta'}) = B_F^*(\theta : \theta') = B_F(\theta' : \theta). \tag{19}$$

A *Bregman divergence* B_F is defined for a strictly convex and differentiable generator F as:

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - \langle \theta - \theta', \nabla F(\theta') \rangle, \tag{20}$$

where $\langle \cdot, \cdot \rangle$ is an inner product (usually the Euclidean dot product for vector parameters).

Similar to the interpretation of the Jensen–Shannon divergence (statistical divergence) as a Jensen divergence for the negentropy generator, the *Jensen–Bregman divergence* [7] JB_F (parametric divergence JBD) amounts to a Jensen divergence J_F for a strictly convex generator $F : \Theta \to \mathbb{R}$:

$$\begin{aligned}
\mathrm{JB}_F(\theta : \theta') &:= \frac{1}{2}\left(B_F\left(\theta : \frac{\theta + \theta'}{2}\right) + B_F\left(\theta' : \frac{\theta + \theta'}{2}\right)\right), &(21)\\
&= \frac{F(\theta) + F(\theta')}{2} - F\left(\frac{\theta + \theta'}{2}\right) =: J_F(\theta : \theta'), &(22)
\end{aligned}$$

Let us introduce the notation $(\theta_p \theta_q)_\alpha := (1-\alpha)\theta_p + \alpha\theta_q$ to denote the *linear interpolation* (LERP) of the parameters. Then we have more generally that the skew Jensen–Bregman divergence $\mathrm{JB}_F^\alpha(\theta : \theta')$ amounts to a skew Jensen divergence $J_F^\alpha(\theta : \theta')$:

$$\begin{aligned}
\mathrm{JB}_F^\alpha(\theta : \theta') &:= (1-\alpha) B_F\left(\theta : (\theta\theta')_\alpha\right) + \alpha B_F\left(\theta' : (\theta\theta')_\alpha\right), &(23)\\
&= (F(\theta)F(\theta'))_\alpha - F\left((\theta\theta')_\alpha\right) =: J_F^\alpha(\theta : \theta'), &(24)
\end{aligned}$$

1.3. J-Symmetrization and JS-Symmetrization of Distances

For any arbitrary distance $D(p : q)$, we can define its *skew J-symmetrization* for $\alpha \in [0,1]$ by:

$$J_D^\alpha(p : q) := (1-\alpha) D\left(p : q\right) + \alpha D\left(q : p\right), \tag{25}$$

and its *JS-symmetrization* by:

$$\begin{aligned}
\mathrm{JS}_D^\alpha(p : q) &:= (1-\alpha) D\left(p : (1-\alpha)p + \alpha q\right) + \alpha D\left(q : (1-\alpha)p + \alpha q\right), &(26)\\
&= (1-\alpha) D\left(p : (pq)_\alpha\right) + \alpha D\left(q : (pq)_\alpha\right). &(27)
\end{aligned}$$

Usually, $\alpha = \frac{1}{2}$, and for notational brevity, we drop the superscript: $\mathrm{JS}_D(p : q) := \mathrm{JS}_D^{\frac{1}{2}}(p : q)$. The Jeffreys divergence is twice the J-symmetrization of the KLD, and the Jensen–Shannon divergence is the JS-symmetrization of the KLD.

The J-symmetrization of a f-divergence I_f is obtained by taking the generator

$$f_\alpha^J(u) = (1-\alpha)f(u) + \alpha f^\diamond(u), \tag{28}$$

where $f^\diamond(u) = uf(\frac{1}{u})$ is the *conjugate* generator:

$$I_{f^\diamond}(p:q) = I_f^*(p:q) = I_f(q:p). \tag{29}$$

The JS-symmetrization of a f-divergence

$$I_f^\alpha(p:q) {:=} (1-\alpha)I_f(p:(pq)_\alpha) + \alpha I_f(q:(pq)_\alpha), \tag{30}$$

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

$$f_\alpha^{\mathrm{JS}}(u) {:=} (1-\alpha)f(\alpha u + 1 - \alpha) + \alpha f\left(\alpha + \frac{1-\alpha}{u}\right). \tag{31}$$

We check that we have:

$$I_f^\alpha(p:q) = (1-\alpha)I_f(p:(pq)_\alpha) + \alpha I_f(q:(pq)_\alpha) = I_f^{1-\alpha}(q:p) = I_{f_\alpha^{\mathrm{JS}}}(p:q). \tag{32}$$

A family of symmetric distances unifying the Jeffreys divergence with the Jensen–Shannon divergence was proposed in [24]. Finally, let us mention that once we have symmetrized a distance D, we may also metrize this symmetric distance by choosing (when it exists) the largest exponent $\delta > 0$ such that D^δ becomes a metric distance [8,25–28].

1.4. Contributions and Paper Outline

The paper is organized as follows:

Section 2 reports the special case of mixture families in information geometry [18] for which the Jensen–Shannon divergence can be expressed as a Bregman divergence (Theorem 1), and highlight the lack of closed-form formula when considering exponential families. This fact precisely motivated this work.

Section 3 introduces the generalized Jensen–Shannon divergences using statistical mixtures derived from abstract weighted means (Definitions 2 and 5), presents the JS-symmetrization of statistical distances, and report a sufficient condition to get bounded JS-symmetrizations (Property 1).

In Section 4.1, we consider the calculation of the geometric JSD between members of the same exponential family (Theorem 2) and instantiate the formula for the multivariate Gaussian distributions (Corollary 1). We discuss about applications for k-means clustering in Section 4.1.2. In Section 4.2, we illustrate the method with another example that calculates in closed form the harmonic JSD between scale Cauchy distributions (Theorem 4).

Finally, we wrap up and conclude this work in Section 5.

2. Jensen–Shannon Divergence in Mixture and Exponential Families

We are interested to calculate the JSD between densities belonging to parametric families of distributions.

A trivial example is when $p = (p_0, \ldots, p_D)$ and $q = (q_0, \ldots, q_D)$ are categorical distributions: The average distribution $\frac{p+q}{2}$ is a again categorical distribution, and the JSD is expressed plainly as:

$$\mathrm{JS}(p,q) = \frac{1}{2}\sum_{i=0}^{D}\left(p_i \log\frac{2p_i}{p_i+q_i} + q_i \log\frac{2q_i}{p_i+q_i}\right). \tag{33}$$

Another example is when $p = m_{\theta_p}$ and $q = m_{\theta_q}$ both belong to the same *mixture family* [18] $\mathcal{M}$:

$$\mathcal{M} := \left\{ m_\theta(x) = \left(1 - \sum_{i=1}^{D} \theta_i p_i(x)\right) p_0(x) + \sum_{i=1}^{D} \theta_i p_i(x) : \theta_i > 0, \sum_i \theta_i < 1 \right\}, \tag{34}$$

for linearly independent component distributions $p_0, p_1, \ldots, p_D$. We have [29]:

$$\mathrm{KL}(m_{\theta_p} : m_{\theta_q}) = B_F(\theta_p : \theta_q), \tag{35}$$

where B_F is a Bregman divergence defined in Equation (20) obtained for the convex negentropy generator [29] $F(\theta) = -h(m_\theta)$. The proof that $F(\theta)$ is a strictly convex function is not trivial [30].

The mixture families include the family of categorical distributions over a finite alphabet $\mathcal{X} = \{E_0, \ldots, E_D\}$ (the D-dimensional probability simplex) since those categorical distributions form a mixture family with $p_i(x) := \Pr(X = E_i) = \delta_{E_i}(x)$. Beware that mixture families impose to prescribe the component distributions. Therefore, a density of a mixture family is a special case of statistical mixtures (e.g., GMMs) with prescribed component distributions.

The mathematical identity of Equation (35) that does not yield a practical formula since $F(\theta)$ is usually not itself available in closed form. Worse, the Bregman generator can be non-analytic [31]. Nevertheless, this identity is useful for computing the right-sided Bregman centroid (left KL centroid of mixtures) since this centroid is equivalent to the center of mass, and independent of the Bregman generator [29].

Since the mixture of mixtures is also a mixture, specifically

$$\frac{m_{\theta_p} + m_{\theta_q}}{2} = m_{\frac{\theta_p + \theta_q}{2}} \in \mathcal{M}, \tag{36}$$

it follows that we get a closed-form expression for the JSD between mixtures belonging to $\mathcal{M}$.

Theorem 1 (JSD between mixtures)**.** *The Jensen–Shannon divergence between two distributions $p = m_{\theta_p}$ and $q = m_{\theta_q}$ belonging to the same mixture family $\mathcal{M}$ is expressed as a Jensen–Bregman divergence for the negentropy generator F:*

$$\mathrm{JS}(m_{\theta_p}, m_{\theta_q}) = \frac{1}{2}\left(B_F\left(\theta_p : \frac{\theta_p + \theta_q}{2}\right) + B_F\left(\theta_q : \frac{\theta_p + \theta_q}{2}\right)\right). \tag{37}$$

This amounts to calculate the Jensen divergence:

$$\mathrm{JS}(m_{\theta_p}, m_{\theta_q}) = J_F(\theta_1; \theta_2) = (F(\theta_1)F(\theta_2))_{\frac{1}{2}} - F((\theta_1\theta_2)_{\frac{1}{2}}), \tag{38}$$

where $(v_1 v_2)_\alpha := (1-\alpha)v_1 + \alpha v_2$.

Now, consider distributions $p = e_{\theta_p}$ and $q = e_{\theta_q}$ belonging to the same *exponential family* [18] $\mathcal{E}$:

$$\mathcal{E} := \left\{ e_\theta(x) = \exp\left(\theta^\top x - F(\theta)\right) : \theta \in \Theta \right\}, \tag{39}$$

where

$$\Theta := \left\{ \theta \in \mathbb{R}^D : \int \exp(\theta^\top x) \mathrm{d}\mu < \infty \right\}, \tag{40}$$

denotes the natural parameter space. We have [18]:

$$\mathrm{KL}(e_{\theta_p} : e_{\theta_q}) = B_F(\theta_q : \theta_p), \tag{41}$$

where F denotes the log-normalizer or cumulant function of the exponential family [18].

However, $\frac{e_{\theta_p}+e_{\theta_q}}{2}$ *does not* belong to $\mathcal{E}$ in general, except for the case of the categorical/multinomial family which is both an exponential family and a mixture family [18].

For example, the mixture of two Gaussian distributions with distinct components is *not* a Gaussian distribution. Thus, it is not obvious to get a closed-form expression for the JSD in that case. This limitation precisely motivated the introduction of generalized JSDs defined in the next section.

Notice that in [32,33], it is shown how to express or approximate the f-divergences using expansions of power χ pseudo-distances. These power chi distances can all be expressed in closed form when dealing with isotropic Gaussians. This result holds for the JSD since the JSD is a f-divergence [33].

3. Generalized Jensen–Shannon Divergences

We first define abstract means M, and then generic statistical M-mixtures from which generalized Jensen–Shannon divergences are built thereof.

Definitions

Consider an *abstract mean* [34] M. That is, a continuous bivariate function $M(\cdot,\cdot) : I \times I \to I$ on an interval $I \subset \mathbb{R}$ that satisfies the following *in-betweenness* property:

$$\inf\{x,y\} \leq M(x,y) \leq \sup\{x,y\}, \quad \forall x,y \in I. \tag{42}$$

Using the unique *dyadic expansion* of real numbers, we can always build a corresponding *weighted mean* $M_\alpha(p,q)$ (with $\alpha \in [0,1]$) following the construction reported in [34] (page 3) such that $M_0(p,q) = p$ and $M_1(p,q) = q$. In the remainder, we consider $I = (0,\infty)$.

Examples of common weighted means are:

- the *arithmetic mean* $A_\alpha(x,y) = (1-\alpha)x + \alpha_y$,
- the *geometric mean* $G_\alpha(x,y) = x^{1-\alpha}y^\alpha$, and
- the *harmonic mean* $H_\alpha(x,y) = \frac{xy}{(1-\alpha)y+\alpha x}$.

These means can be unified using the concept of *quasi-arithmetic means* [34] (also called Kolmogorov–Nagumo means):

$$M_\alpha^h(x,y) {:=} h^{-1}\left((1-\alpha)h(x) + \alpha h(y)\right), \tag{43}$$

where h is a strictly monotonous function. For example, the geometric mean $G_\alpha(x,y)$ is obtained as $M_\alpha^h(x,y)$ for the generator $h(u) = \log(u)$. Rényi used the concept of quasi-arithmetic means instead of the arithmetic mean to define axiomatically the Rényi entropy [35] of order α in information theory [2].

For any abstract weighted mean, we can build a statistical mixture called a M-mixture as follows:

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:*

$$(pq)_\alpha^M(x) {:=} \frac{M_\alpha(p(x),q(x))}{Z_\alpha^M(p:q)}, \tag{44}$$

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* $\mathcal{X}$*.)*

The A-mixture $(pq)_\alpha^A(x) = (1-\alpha)p(x) + \alpha q(x)$ ('A' standing for the arithmetic mean) represents the usual statistical mixture [36] (with $Z_\alpha^A(p:q) = 1$). The G-mixture $(pq)_\alpha^G(x) = \frac{p(x)^{1-\alpha}q(x)^\alpha}{Z_\alpha^G(p:q)}$ of two distributions $p(x)$ and $q(x)$ ('G' standing for the geometric mean G) is an exponential family of order [37] 1:

$$(pq)^G_\alpha(x) = \exp\left((1-\alpha)p(x) + \alpha q(x) - \log Z^G_\alpha(p:q)\right). \tag{46}$$

The two-component M-mixture can be generalized to a k-component M-mixture with $\alpha \in \Delta_{k-1}$, the $(k-1)$-dimensional standard simplex:

$$(p_1 \ldots p_k)^M_\alpha := \frac{p_1(x)^{\alpha_1} \times \ldots \times p_k(x)^{\alpha_k}}{Z_\alpha(p_1, \ldots, p_k)}, \tag{47}$$

where $Z_\alpha(p_1, \ldots, p_k) := \int_{\mathcal{X}} p_1(x)^{\alpha_1} \times \ldots \times p_k(x)^{\alpha_k} \mathrm{d}\mu(x)$.

For a given pair of distributions p and q, the set $\{M_\alpha(p(x), q(x)) : \alpha \in [0,1]\}$ describes a path in the space of probability density functions. This density interpolation scheme was investigated for quasi-arithmetic weighted means in [38–40]. In [41], the authors study the Fisher information matrix for the α-mixture models (using α-power means).

We call $(pq)^M_\alpha$ the α-weighted M-mixture, thus extending the notion of α-mixtures [42] obtained for power means P_α. Notice that abstract means have also been used to generalize Bregman divergences using the concept of (M,N)-*convexity* [43].

Let us state a first generalization of the Jensen–Shannon divergence:

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

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

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:

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*

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

By notation, we have $\mathrm{JS}^{M_\alpha}(p:q) = \mathrm{JS}^{M_\alpha}_{\mathrm{KL}}(p:q)$. That is, the arithmetic JS-symmetrization of the KLD is the JSD.

Let us define the α-skew K-divergence [6,44] $K_\alpha(p:q)$ as

$$K_\alpha\,(p:q) := \mathrm{KL}(p : (1-\alpha)p + \alpha q) = \mathrm{KL}(p : (pq)_\alpha), \tag{50}$$

where $(pq)_\alpha(x) := (1-\alpha)p(x) + \alpha q(x)$. Then the Jensen–Shannon divergence and the Jeffreys divergence can be rewritten [24] as

$$\mathrm{JS}\,(p;q) = \frac{1}{2}\left(K_{\frac{1}{2}}\,(p:q) + K_{\frac{1}{2}}\,(q:p)\right), \tag{51}$$

$$J\,(p;q) = K_1(p:q) + K_1(q:p), \tag{52}$$

since $\mathrm{KL}(p:q) = K_1(p:q)$. Then $\mathrm{JS}_\alpha(p:q) = (1-\alpha)K_\alpha(p:q) + \alpha K_{1-\alpha}(q:p)$. Similarly, we can define the generalized skew K-divergence:

$$K^{M_\alpha}_D(p:q) := D\left(p : (pq)^M_\alpha\right). \tag{53}$$

The success of the JSD compared to the JD in applications is partially due to the fact that the JSD is upper bounded by $\log 2$. So, one question to ask is whether those generalized JSDs are upper bounded or not?

To report a sufficient condition, let us first introduce the dominance relationship between means: We say that a mean M dominates a mean N when $M(x,y) \geq N(x,y)$ for all $x, y \geq 0$, see [34]. In that case we write concisely $M \geq N$. For example, the Arithmetic-Geometric-Harmonic (AGH) inequality states that $A \geq G \geq H$.

Consider the term

$$\mathrm{KL}(p : (pq)_\alpha^M) = \int p(x) \log \frac{p(x) Z_\alpha^M(p,q)}{M_\alpha(p(x), q(x))} \mathrm{d}\mu(x), \quad (54)$$

$$= \log Z_\alpha^M(p,q) + \int p(x) \log \frac{p(x)}{M_\alpha(p(x), q(x))} \mathrm{d}\mu(x). \quad (55)$$

When mean M_α dominates the arithmetic mean A_α, we have

$$\int p(x) \log \frac{p(x)}{M_\alpha(p(x), q(x))} \mathrm{d}\mu(x) \leq \int p(x) \log \frac{p(x)}{A_\alpha(p(x), q(x))} \mathrm{d}\mu(x),$$

and

$$\int p(x) \log \frac{p(x)}{A_\alpha(p(x), q(x))} \mathrm{d}\mu(x) \leq \int p(x) \log \frac{p(x)}{(1-\alpha)p(x)} \mathrm{d}\mu(x) = -\log(1-\alpha).$$

Notice that $Z_\alpha^A(p : q) = 1$ (when $M = A$ is the arithmetic mean), and we recover the fact that the α-skew Jensen–Shannon divergence is upper bounded by $-\log(1-\alpha)$ (e.g., $\log 2$ when $\alpha = \frac{1}{2}$).

We summarize the result in the following property:

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

Let us observe that dominance of means can be used to define distances: For example, the celebrated α-divergences

$$I_\alpha(p : q) = \int \left(\alpha p(x) + (1-\alpha) q(x) - p(x)^\alpha q(x)^{1-\alpha} \right) \mathrm{d}\mu(x), \quad \alpha \notin \{0,1\} \quad (56)$$

can be interpreted as a difference of two means, the arithmetic mean and the geometry mean:

$$I_\alpha(p : q) = \int (A_\alpha(q(x) : p(x)) - G_\alpha(q(x) : p(x))) \, \mathrm{d}\mu(x). \quad (57)$$

We can also define the generalized Jeffreys divergence as follows:

Definition 4 (N-Jeffreys divergence)**.** *For a mean N, the skew N-Jeffreys divergence (for $\beta \in [0,1]$) is defined by*

$$J^{N_\beta}(p : q) := N_\beta(\mathrm{KL}\,(p : q)\,, \mathrm{KL}\,(q : p)). \quad (58)$$

This definition includes the (scaled) *resistor average distance* [5] $R(p;q)$, obtained for the *harmonic mean* $N = H$ for the KLD with skew parameter $\beta = \frac{1}{2}$:

$$\frac{1}{R(p;q)} = \frac{1}{2} \left(\frac{1}{\mathrm{KL}(p : q)} + \frac{1}{\mathrm{KL}(q : p)} \right), \quad (59)$$

$$R(p;q) = \frac{2J(p;q)}{\mathrm{KL}(p : q)\mathrm{KL}(q : p)}. \quad (60)$$

In [5], the factor $\frac{1}{2}$ is omitted to keep the spirit of the original Jeffreys divergence.

We can further extend this definition for any arbitrary divergence D as follows:

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

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

We now show how to choose the abstract mean according to the parametric family of distributions to obtain some closed-form formula for some statistical distances.

4. Some Closed-Form Formula for the M-Jensen–Shannon Divergences

Our motivation to introduce these novel families of M-Jensen–Shannon divergences is to obtain closed-form formula when probability densities belong to some given parametric families $\mathcal{P}_\Theta$. We shall illustrate the principle of the method to choose the right abstract mean for the considered parametric family, and report corresponding formula for the following two case studies:

1. The *geometric G-Jensen–Shannon divergence* for the *exponential families* (Section 4.1), and
2. the *harmonic H-Jensen–Shannon divergence* for the family of *Cauchy scale distributions* (Section 4.2).

Recall that the arithmetic A-Jensen–Shannon divergence is well-suited for mixture families (Theorem 1).

4.1. The Geometric G-Jensen–Shannon Divergence

Consider an exponential family [37] $\mathcal{E}_F$ with log-normalizer F:

$$\mathcal{E}_F = \left\{p_\theta(x)\mathrm{d}\mu = \exp(\theta^\top x - F(\theta))\mathrm{d}\mu : \theta \in \Theta\right\}, \tag{62}$$

and natural parameter space

$$\Theta = \left\{\theta : \int_{\mathcal{X}} \exp(\theta^\top x)\mathrm{d}\mu < \infty\right\}. \tag{63}$$

The log-normalizer (a log-Laplace function also called log-partition or cumulant function) is a real analytic convex function.

We seek for a mean M such that the weighted M-mixture density $(p_{\theta_1} p_{\theta_2})_\alpha^M$ of two densities p_{θ_1} and p_{θ_2} of the same exponential family yields another density of that exponential family (e.g., $p_{(\theta_1\theta_2)_\alpha}$). When considering exponential families, choose the *weighted geometric mean G_α* for the abstract mean $M_\alpha(x,y)$: $M_\alpha(x,y) = G_\alpha(x,y) = x^{1-\alpha}y^\alpha$, for $x, y > 0$. Indeed, it is well-known that the normalized weighted product of distributions belonging to the same exponential family also belongs to this exponential family [45]:

$$\begin{aligned} \forall x \in \mathcal{X}, \quad (p_{\theta_1} p_{\theta_2})_\alpha^G(x) &:= \frac{G_\alpha(p_{\theta_1}(x), p_{\theta_2}(x))}{\int G_\alpha(p_{\theta_1}(t), p_{\theta_2}(t))\mathrm{d}\mu(t)} = \frac{p_{\theta_1}^{1-\alpha}(x) p_{\theta_2}^\alpha(x)}{Z_\alpha^G(p:q)}, & (64) \\ &= p_{(\theta_1\theta_2)_\alpha}(x), & (65) \end{aligned}$$

where the normalization factor is

$$Z_\alpha^G(p:q) = \exp(-J_F^\alpha(\theta_1 : \theta_2)), \tag{66}$$

for the skew Jensen divergence J_F^α defined by:

$$J_F^\alpha(\theta_1 : \theta_2) := (F(\theta_1)F(\theta_2))_\alpha - F((\theta_1\theta_2)_\alpha). \tag{67}$$

Notice that since the natural parameter space Θ is convex, the distribution $p_{(\theta_1\theta_2)_\alpha} \in \mathcal{E}_F$ (since $(\theta_1\theta_2)_\alpha \in \Theta$).

Thus, it follows that we have:

$$\begin{aligned} \mathrm{KL}\left(p_\theta : (p_{\theta_1}p_{\theta_2})^G_\alpha\right) &= \mathrm{KL}\left(p_\theta : p_{(\theta_1\theta_2)_\alpha}\right), && (68)\\ &= B_F((\theta_1\theta_2)_\alpha : \theta). && (69)\end{aligned}$$

This allows us to conclude that the *G-Jensen–Shannon divergence* admits the following closed-form expression between densities belonging to the same exponential family:

$$\begin{aligned} \mathrm{JS}^G_\alpha(p_{\theta_1} : p_{\theta_2}) &:= (1-\alpha)\mathrm{KL}(p_{\theta_1} : (p_{\theta_1}p_{\theta_2})^G_\alpha) + \alpha\mathrm{KL}(p_{\theta_2} : (p_{\theta_1}p_{\theta_2})^G_\alpha), && (70)\\ &= (1-\alpha)B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2). && (71)\end{aligned}$$

Please note that since $(\theta_1\theta_2)_\alpha - \theta_1 = \alpha(\theta_2-\theta_1)$ and $(\theta_1\theta_2)_\alpha - \theta_2 = (1-\alpha)(\theta_1-\theta_2)$, it follows that $(1-\alpha)B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha) = J^\alpha_F(\theta_1 : \theta_2)$.

The *dual divergence* [46] D^* (with respect to the reference argument) or *reverse divergence* of a divergence D is defined by swapping the calling arguments: $D^*(\theta : \theta') := D(\theta' : \theta)$.

Thus, if we defined the Jensen–Shannon divergence for the dual KL divergence $\mathrm{KL}^*(p : q) := \mathrm{KL}(q : p)$

$$\begin{aligned} \mathrm{JS}_{\mathrm{KL}^*}(p:q) &:= \frac{1}{2}\left(\mathrm{KL}^*\left(p : \frac{p+q}{2}\right) + \mathrm{KL}^*\left(q : \frac{p+q}{2}\right)\right), && (72)\\ &= \frac{1}{2}\left(\mathrm{KL}\left(\frac{p+q}{2} : p\right) + \mathrm{KL}\left(\frac{p+q}{2} : q\right)\right), && (73)\end{aligned}$$

then we obtain:

$$\begin{aligned} \mathrm{JS}^{G_\alpha}_{\mathrm{KL}^*}(p_{\theta_1} : p_{\theta_2}) &:= (1-\alpha)\mathrm{KL}((p_{\theta_1}p_{\theta_2})^G_\alpha : p_{\theta_1}) + \alpha\mathrm{KL}((p_{\theta_1}p_{\theta_2})^G_\alpha : p_{\theta_2}), && (74)\\ &= (1-\alpha)B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha) = \mathrm{JB}^\alpha_F(\theta_1 : \theta_2), && (75)\\ &= (1-\alpha)F(\theta_1) + \alpha F(\theta_2) - F((\theta_1\theta_2)_\alpha), && (76)\\ &= J^\alpha_F(\theta_1 : \theta_2). && (77)\end{aligned}$$

Please note that $\mathrm{JS}_{D^*} \neq \mathrm{JS}_D{}^*$.

In general, the JS-symmetrization for the reverse KL divergence is

$$\begin{aligned} \mathrm{JS}_{\mathrm{KL}^*}(p;q) &= \frac{1}{2}\left(\mathrm{KL}\left(\frac{p+q}{2} : p\right) + \mathrm{KL}\left(\frac{p+q}{2} : q\right)\right), && (78)\\ &= \int m \log\frac{m}{\sqrt{pq}}\mathrm{d}\mu = \int A(p,q)\log\frac{A(p,q)}{G(p,q)}\mathrm{d}\mu, && (79)\end{aligned}$$

where $m = \frac{p+q}{2} = A(p,q)$ and $G(p,q) = \sqrt{pq}$. Since $A \geq G$ (arithmetic-geometric inequality), it follows that $\mathrm{JS}_{\mathrm{KL}^*}(p;q) \geq 0$.

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:*

$$\begin{aligned} \mathrm{JS}^{G_\alpha}(p_{\theta_1} : p_{\theta_2}) &= (1-\alpha)B_F\left((\theta_1\theta_2)_\alpha : \theta_1\right) + \alpha B_F\left((\theta_1\theta_2)_\alpha : \theta_2\right), && (80)\\ \mathrm{JS}^{G_\alpha}_{\mathrm{KL}^*}(p_{\theta_1} : p_{\theta_2}) &= \mathrm{JB}^\alpha_F(\theta_1 : \theta_2) = J^\alpha_F(\theta_1 : \theta_2). && (81)\end{aligned}$$

4.1.1. Case Study: The Multivariate Gaussian Family

Consider the *exponential family* [18,37] of multivariate Gaussian distributions [47–49]

$$\{N(\mu,\Sigma) : \mu \in \mathbb{R}^d, \Sigma \succ 0\}. \tag{82}$$

The multivariate Gaussian family is also called the *multivariate normal* family in the literature, or MVN family for short.

Let $\lambda{:=}(\lambda_v, \lambda_M) = (\mu, \Sigma)$ denote the *composite* (vector,matrix) parameter of an MVN. The d-dimensional MVN density is given by

$$p_\lambda(x;\lambda) \quad := \quad \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\lambda_M|}} \exp\left(-\frac{1}{2}(x-\lambda_v)^\top \lambda_M^{-1}(x-\lambda_v)\right), \tag{83}$$

where $|\cdot|$ denotes the matrix determinant. The natural parameters θ are also expressed using both a vector parameter θ_v and a matrix parameter θ_M in a compound object $\theta = (\theta_v, \theta_M)$. By defining the following *compound inner product* on a composite (vector,matrix) object

$$\langle \theta, \theta' \rangle {:=} \theta_v^\top \theta_v' + \mathrm{tr}\left({\theta_M'}^\top \theta_M\right), \tag{84}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace, we rewrite the MVN density of Equation (83) in the canonical form of an exponential family [37]:

$$p_\theta(x;\theta) \quad := \quad \exp\left(\langle t(x), \theta \rangle - F_\theta(\theta)\right) = p_\lambda(x; \lambda(\theta)), \tag{85}$$

where

$$\theta = (\theta_v, \theta_M) = \left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right) = \theta(\lambda) = \left(\lambda_M^{-1}\lambda_v, -\frac{1}{2}\lambda_M^{-1}\right), \tag{86}$$

is the *compound natural parameter* and

$$t(x) = (x, -xx^\top) \tag{87}$$

is the *compound sufficient statistic*. The function F_θ is the strictly convex and continuously differentiable log-normalizer defined by:

$$F_\theta(\theta) = \frac{1}{2}\left(d \log \pi - \log|\theta_M| + \frac{1}{2}\theta_v^\top \theta_M^{-1}\theta_v\right), \tag{88}$$

The log-normalizer can be expressed using the ordinary parameters, $\lambda = (\mu, \Sigma)$, as:

$$\begin{aligned} F_\lambda(\lambda) &= \frac{1}{2}\left(\lambda_v^\top \lambda_M^{-1}\lambda_v + \log|\lambda_M| + d\log 2\pi\right), && (89) \\ &= \frac{1}{2}\left(\mu^\top \Sigma^{-1}\mu + \log|\Sigma| + d\log 2\pi\right). && (90) \end{aligned}$$

The *moment/expectation parameters* [18,49] are

$$\eta = (\eta_v, \eta_M) = E[t(x)] = \nabla F(\theta). \tag{91}$$

We report the conversion formula between the three types of coordinate systems (namely the ordinary parameter λ, the natural parameter θ and the moment parameter η) as follows:

$$\left\{\begin{array}{l}\theta_v(\lambda)=\lambda_M^{-1}\lambda_v=\Sigma^{-1}\mu\\ \theta_M(\lambda)=\frac{1}{2}\lambda_M^{-1}=\frac{1}{2}\Sigma^{-1}\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\lambda_v(\theta)=\frac{1}{2}\theta_M^{-1}\theta_v=\mu\\ \lambda_M(\theta)=\frac{1}{2}\theta_M^{-1}=\Sigma\end{array}\right. \quad (92)$$

$$\left\{\begin{array}{l}\eta_v(\theta)=\frac{1}{2}\theta_M^{-1}\theta_v\\ \eta_M(\theta)=-\frac{1}{2}\theta_M^{-1}-\frac{1}{4}(\theta_M^{-1}\theta_v)(\theta_M^{-1}\theta_v)^\top\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\theta_v(\eta)=-(\eta_M+\eta_v\eta_v^\top)^{-1}\eta_v\\ \theta_M(\eta)=-\frac{1}{2}(\eta_M+\eta_v\eta_v^\top)^{-1}\end{array}\right. \quad (93)$$

$$\left\{\begin{array}{l}\lambda_v(\eta)=\eta_v=\mu\\ \lambda_M(\eta)=-\eta_M-\eta_v\eta_v^\top=\Sigma\end{array}\right. \Leftrightarrow \left\{\begin{array}{l}\eta_v(\lambda)=\lambda_v=\mu\\ \eta_M(\lambda)=-\lambda_M-\lambda_v\lambda_v^\top=-\Sigma-\mu\mu^\top\end{array}\right. \quad (94)$$

The dual Legendre convex conjugate [18,49] is

$$F_\eta^*(\eta)=-\frac{1}{2}\left(\log(1+\eta_v^\top\eta_M^{-1}\eta_v)+\log|-\eta_M|+d(1+\log 2\pi)\right), \quad (95)$$

and $\theta=\nabla_\eta F_\eta^*(\eta)$.

We check the Fenchel-Young equality when $\eta=\nabla F(\theta)$ and $\theta=\nabla F^*(\eta)$:

$$F_\theta(\theta)+F_\eta^*(\eta)-\langle\theta,\eta\rangle=0. \quad (96)$$

The Kullback–Leibler divergence between two d-dimensional Gaussians distributions $p_{(\mu_1,\Sigma_1)}$ and $p_{(\mu_2,\Sigma_2)}$ (with $\Delta_\mu=\mu_2-\mu_1$) is

$$\mathrm{KL}(p_{(\mu_1,\Sigma_1)}:p_{(\mu_2,\Sigma_2)}) \;=\; \frac{1}{2}\left\{\mathrm{tr}(\Sigma_2^{-1}\Sigma_1)+\Delta_\mu^\top\Sigma_2^{-1}\Delta_\mu+\log\frac{|\Sigma_2|}{|\Sigma_1|}-d\right\}=\mathrm{KL}(p_{\lambda_1}:p_{\lambda_2}). \quad (97)$$

We check that $\mathrm{KL}(p_{(\mu,\Sigma)}:p_{(\mu,\Sigma)})=0$ since $\Delta_\mu=0$ and $\mathrm{tr}(\Sigma^{-1}\Sigma)=\mathrm{tr}(I)=d$. Notice that when $\Sigma_1=\Sigma_2=\Sigma$, we have

$$\mathrm{KL}(p_{(\mu_1,\Sigma)}:p_{(\mu_2,\Sigma)})=\frac{1}{2}\Delta_\mu^\top\Sigma^{-1}\Delta_\mu=\frac{1}{2}D_{\Sigma^{-1}}^2(\mu_1,\mu_2), \quad (98)$$

that is half the squared Mahalanobis distance for the precision matrix Σ^{-1} (a positive-definite matrix: $\Sigma^{-1}\succ 0$), where the Mahalanobis distance is defined for any positive matrix $Q\succ 0$ as follows:

$$D_Q(p_1:p_2)=\sqrt{(p_1-p_2)^\top Q(p_1-p_2)}. \quad (99)$$

The Kullback–Leibler divergence between two probability densities of the same exponential families amount to a Bregman divergence [18]:

$$\mathrm{KL}(p_{(\mu_1,\Sigma_1)}:p_{(\mu_2,\Sigma_2)})=\mathrm{KL}(p_{\lambda_1}:p_{\lambda_2})=B_F(\theta_2:\theta_1)=B_{F^*}(\eta_1:\eta_2), \quad (100)$$

where the Bregman divergence is defined by

$$B_F(\theta:\theta'):=F(\theta)-F(\theta')-\langle\theta-\theta',\nabla F(\theta')\rangle, \quad (101)$$

with $\eta'=\nabla F(\theta')$. Define the canonical divergence [18]

$$A_F(\theta_1:\eta_2)=F(\theta_1)+F^*(\eta_2)-\langle\theta_1,\eta_2\rangle=A_{F^*}(\eta_2:\theta_1), \quad (102)$$

since $F^{**}=F$. We have $B_F(\theta_1:\theta_2)=A_F(\theta_1:\eta_2)$.

Now, observe that $p_\theta(0,\theta)=\exp(-F(\theta))$ when $\langle t(0),\theta\rangle=0$. In particular, this holds for the multivariate normal family. Thus, we have the following proposition.

Proposition 1. *For the MVN family, we have*

$$p_\theta(x;(\theta_1\theta_2)_\alpha) = \frac{p_\theta(x,\theta_1)^{1-\alpha} p_\theta(x,\theta_2)^\alpha}{Z_\alpha^G(p_{\theta_1} : p_{\theta_2})}, \tag{103}$$

with the scaling normalization factor:

$$Z_\alpha^G(p_{\theta_1} : p_{\theta_2}) = \exp(-J_F^\alpha(\theta_1 : \theta_2)) = \frac{p_\theta(0;\theta_1)^{1-\alpha} p_\theta(0;\theta_2)^\alpha}{p_\theta(0;(\theta_1\theta_2)_\alpha)}. \tag{104}$$

More generally, we have for a k-dimensional weight vector α belonging to the $(k-1)$-dimensional standard simplex:

$$Z_\alpha^G(p_{\theta_1}, \ldots p_{\theta_k}) = \frac{\prod_{i=1}^k p_\theta(0,\theta_i)^{\alpha_i}}{p_\theta(0;\bar{\theta})}, \tag{105}$$

where $\bar{\theta} = \sum_{i=1}^k \alpha_i \theta_i$.

Finally, we state the formulas for the G-JS divergence between MVNs for the KL and reverse KL, respectively:

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

$$\begin{aligned}
\mathrm{JS}^{G_\alpha}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_\alpha,\Sigma_\alpha)}) + \alpha\mathrm{KL}(p_{(\mu_2,\Sigma_2)} : p_{(\mu_\alpha,\Sigma_\alpha)}), && (106)\\
&= (1-\alpha)B_F((\theta_1\theta_2)_\alpha : \theta_1) + \alpha B_F((\theta_1\theta_2)_\alpha : \theta_2), && (107)\\
&= \frac{1}{2}\Bigg(\mathrm{tr}\left(\Sigma_\alpha^{-1}((1-\alpha)\Sigma_1 + \alpha\Sigma_2)\right) + \log\frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha} + \\
&\quad (1-\alpha)(\mu_\alpha-\mu_1)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_1) + \alpha(\mu_\alpha-\mu_2)^\top\Sigma_\alpha^{-1}(\mu_\alpha-\mu_2) - d\Bigg) && (108)\\
\mathrm{JS}_*^{G_\alpha}(p_{(\mu_1,\Sigma_1)} : p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_1,\Sigma_1)}) + \alpha\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)} : p_{(\mu_2,\Sigma_2)}), && (109)\\
&= (1-\alpha)B_F(\theta_1 : (\theta_1\theta_2)_\alpha) + \alpha B_F(\theta_2 : (\theta_1\theta_2)_\alpha), && (110)\\
&= J_F(\theta_1 : \theta_2), && (111)\\
&= \frac{1}{2}\left((1-\alpha)\mu_1^\top\Sigma_1^{-1}\mu_1 + \alpha\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}{|\Sigma_\alpha|}\right), && (112)
\end{aligned}$$

where

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

(matrix harmonic barycenter) and

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

Notice that the *α-skew Bhattacharyya distance* [7]:

$$B_\alpha(p:q) = -\log\int_{\mathcal{X}} p^{1-\alpha}q^\alpha \mathrm{d}\mu \tag{115}$$

between two members of the same exponential family amounts to a α-skew Jensen divergence between the corresponding natural parameters:

$$B_\alpha(p_{\theta_1} : p_{\theta_2}) = J_F^\alpha(\theta_1 : \theta_2). \tag{116}$$

A simple proof follows from the fact that

$$\int p_{(\theta_1\theta_2)_\alpha}(x)\mathrm{d}\mu(x) = 1 = \int \frac{p_{\theta_1}^{1-\alpha}(x)p_{\theta_2}^{\alpha}(x)}{Z_\alpha^G(p_{\theta_1}:p_{\theta_2})}\mathrm{d}\mu(x). \tag{117}$$

Therefore, we have

$$\log 1 = 0 = \log \int p_{\theta_1}^{1-\alpha}(x)p_{\theta_2}^{\alpha}(x)\mathrm{d}\mu(x) - \log Z_\alpha^G(p_{\theta_1}:p_{\theta_2}), \tag{118}$$

with $Z_\alpha^G(p_{\theta_1}:p_{\theta_2}) = \exp(-J_F(p_{\theta_1}:p_{\theta_2}))$. Thus, it follows that

$$\begin{aligned} B_\alpha(p_{\theta_1}:p_{\theta_2}) &= -\log \int p_{\theta_1}^{1-\alpha}(x)p_{\theta_2}^{\alpha}(x)\mathrm{d}\mu(x), && (119)\\ &= -\log Z_\alpha^G(p_{\theta_1}:p_{\theta_2}), && (120)\\ &= J_F(p_{\theta_1}:p_{\theta_2}). && (121)\end{aligned}$$

Corollary 2. *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.*

4.1.2. Applications to k-Means Clustering

Let $P = \{p_1, \ldots, p_n\}$ denote a point set, and $C = \{c_1, \ldots, c_k\}$ denote a set of k (cluster) centers. The generalized k-means objective [23] with respect to a distance D is defined by:

$$E_D(P,C) = \frac{1}{n}\sum_{i=1}^{n} \min_{j\in\{1,\ldots,k\}} D(p_i:c_j). \tag{122}$$

By defining the distance $D(p,C) = \min_{j\in\{1,\ldots,k\}} D(p:c_j)$ of a point to a set of points, we can rewrite compactly the objective function as $E_D(P,C) = \frac{1}{n}\sum_{i=1}^{n} D(p_i,C)$. Denote by $E_D^*(P,k)$ the minimum objective loss for a set of $k = |C|$ clusters: $E_D^*(P,k) = \min_{|C|=k} E_D(P,C)$. It is NP-hard [50] to compute $E_D^*(P,k)$ when $k > 1$ and the dimension $d > 1$. The most common heuristic is Lloyd's batched k-means [23] that yields a local minimum.

The performance of the *probabilistic k-means++ initialization* [51] has been extended to arbitrary distances in [52] as follows:

Theorem 3 (Generalized k-means++ performance, [53]). *Let κ_1 and κ_2 be two constants such that κ_1 defines the quasi-triangular inequality property:*

$$D(x:z) \leq \kappa_1\left(D(x:y)+D(y:z)\right), \quad \forall x,y,z \in \Delta^d, \tag{123}$$

and κ_2 handles the symmetry inequality:

$$D(x:y) \leq \kappa_2 D(y:x), \quad \forall x,y \in \Delta^d. \tag{124}$$

Then the generalized k-means++ seeding guarantees with high probability a configuration C of cluster centers such that:

$$E_D(P,C) \leq 2\kappa_1^2(1+\kappa_2)(2+\log k)E_D^*(P,k). \tag{125}$$

To bound the constants κ_1 and κ_2, we rewrite the generalized Jensen–Shannon divergences using quadratic form expressions: That is, using a squared Mahalanobis distance:

$$D_Q(p:q) = \sqrt{(p-q)^\top Q(p-q)}, \tag{126}$$

for a positive-definite matrix $Q \succ 0$. Since the Bregman divergence can be interpreted as the tail of a first-order Taylor expansion, we have:

$$B_F(\theta_1 : \theta_2) = \frac{1}{2}(\theta_1 - \theta_2)^\top \nabla^2 F(\xi)(\theta_1 - \theta_2), \tag{127}$$

for $\xi \in \Theta$ (open convex). Similarly, the Jensen divergence can be interpreted as a Jensen–Bregman divergence, and thus we have

$$J_F(\theta_1 : \theta_2) \frac{1}{2}(\theta_1 - \theta_2)^\top \nabla^2 F(\xi')(\theta_1 - \theta_2), \tag{128}$$

for $\xi' \in \Theta$. More precisely, for a prescribed point set $\{\theta_1, \ldots, \theta_n\}$, we have $\xi, \xi' \in \mathrm{CH}(\{\theta_1, \ldots, \theta_n\})$, where CH denotes the closed convex hull. We can therefore upper bound κ_1 and κ_2 using the ratio $\frac{\max_{\theta \in \mathrm{CH}(\{\theta_1,\ldots,\theta_n\})} \|\nabla^2 F(\theta)\|}{\max_{\theta \in \mathrm{CH}(\{\theta_1,\ldots,\theta_n\})} \|\nabla^2 F(\theta)\|}$. See [54] for further details.

A centroid for a set of parameters $\theta_1, \ldots, \theta_n$ is defined as the minimizer of the functional

$$E_D(\theta) = \frac{1}{n} \sum_i D(\theta_i : \theta). \tag{129}$$

In particular, the *symmetrized Bregman centroids* have been studied in [55] (for JS^{G_α}), and the *Jensen centroids* (for $\mathrm{JS}_*^{G_\alpha}$) have been investigated in [7] using the convex-concave iterative procedure.

4.2. The Harmonic Jensen–Shannon Divergence (H-JS)

The *principle* to get closed-form formula for generalized Jensen–Shannon divergences between distributions belonging to a parametric family $\mathcal{P}_\Theta = \{p_\theta : \theta \in \Theta\}$ consists of finding an abstract mean M such that the M-mixture $(p_{\theta_1} p_{\theta_2})^M_\alpha$ belongs to the family $\mathcal{P}_\Theta$. In particular, when Θ is a convex domain, we seek a mean M such that $(p_{\theta_1} p_{\theta_2})^M_\alpha = p_{(\theta_1 \theta_2)_\alpha}$ with $(\theta_1 \theta_2)_\alpha \in \Theta$.

Let us consider the *weighted harmonic mean* [34] (induced by the harmonic mean) H:

$$H_\alpha(x, y) := \frac{1}{(1-\alpha)\frac{1}{x} + \alpha\frac{1}{y}} = \frac{xy}{(1-\alpha)y + \alpha x} = \frac{xy}{(xy)_{1-\alpha}}, \quad \alpha \in [0, 1]. \tag{130}$$

The harmonic mean is a quasi-arithmetic mean $H_\alpha(x, y) = M^h_\alpha(x, y)$ obtained for the monotone (decreasing) function $h(u) = \frac{1}{u}$ (or equivalently for the increasing monotone function $h(u) = -\frac{1}{u}$).

This harmonic mean is well-suited for the *scale family* $\mathcal{C}$ of Cauchy probability distributions (also called Lorentzian distributions):

$$\mathcal{C}_\Gamma := \left\{ p_\gamma(x) = \frac{1}{\gamma} p_{\mathrm{std}}\left(\frac{x}{\gamma}\right) = \frac{\gamma}{\pi(\gamma^2 + x^2)} : \gamma \in \Gamma = (0, \infty) \right\}, \tag{131}$$

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

Using the computer algebra system Maxima (http://maxima.sourceforge.net/) we find that (see Appendix B)

$$(p_{\gamma_1} p_{\gamma_2})^H_{\frac{1}{2}}(x) = \frac{H_\alpha(p_{\gamma_1}(x) : p_{\gamma_2}(x))}{Z^H_\alpha(\gamma_1, \gamma_2)} = p_{(\gamma_1 \gamma_2)_\alpha} \tag{132}$$

where the normalizing coefficient is

$$Z^H_\alpha(\gamma_1, \gamma_2) := \sqrt{\frac{\gamma_1 \gamma_2}{(\gamma_1 \gamma_2)_\alpha (\gamma_1 \gamma_2)_{1-\alpha}}} = \sqrt{\frac{\gamma_1 \gamma_2}{(\gamma_1 \gamma_2)_\alpha (\gamma_2 \gamma_1)_\alpha}}, \tag{133}$$

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

The H-Jensen–Shannon symmetrization of a distance D between distributions writes as:

$$\mathrm{JS}_D^{H_\alpha}(p:q) = (1-\alpha)D(p:(pq)_\alpha^H) + \alpha D(q:(pq)_\alpha^H), \tag{134}$$

where H_α denote the weighted harmonic mean. When D is available in closed form for distributions belonging to the scale Cauchy distributions, so is $\mathrm{JS}_D^{H_\alpha}(p:q)$.

For example, consider the KL divergence formula between two scale Cauchy distributions:

$$\mathrm{KL}(p_{\gamma_1}:p_{\gamma_2}) = 2\log\frac{A(\gamma_1,\gamma_2)}{G(\gamma_1,\gamma_2)} = 2\log\frac{\gamma_1+\gamma_2}{2\sqrt{\gamma_1\gamma_2}}, \tag{135}$$

where A and G denote the arithmetic and geometric means, respectively. The formula initially reported in [56] has been corrected by the authors. Since $A \geq G$ (and $\frac{A}{G} \geq 1$), it follows that $\mathrm{KL}(p_{\gamma_1}:p_{\gamma_2}) \geq 0$. Notice that the KL divergence is symmetric for Cauchy scale distributions. We note in passing that for exponential families, the KL divergence is symmetric only for the location Gaussian family (since the only symmetric Bregman divergences are the squared Mahalanobis distances [57]). The cross-entropy between scale Cauchy distributions is $h^\times(p_{\gamma_1}:p_{\gamma_2}) = \log\pi\frac{(\gamma_1+\gamma_2)^2}{\gamma_2}$, and the differential entropy is $h(p_\gamma) = h^\times(p_\gamma:p_\gamma) = \log 4\pi\gamma$.

Then the H-JS divergence between $p = p_{\gamma_1}$ and $q = p_{\gamma_2}$ is:

$$\begin{aligned}
\mathrm{JS}^H(p:q) &= \frac{1}{2}\left(\mathrm{KL}\left(p:(pq)_{\frac{1}{2}}^H\right) + \mathrm{KL}\left(q:(pq)_{\frac{1}{2}}^H\right)\right), &(136)\\
\mathrm{JS}^H(p_{\gamma_1}:p_{\gamma_2}) &= \frac{1}{2}\left(\mathrm{KL}\left(p_{\gamma_1}:p_{\frac{\gamma_1+\gamma_2}{2}}\right) + \mathrm{KL}\left(p_{\gamma_2}:p_{\frac{\gamma_1+\gamma_2}{2}}\right)\right), &(137)\\
&= \log\frac{(3\gamma_1+\gamma_2)(3\gamma_2+\gamma_1)}{8\sqrt{\gamma_1\gamma_2}(\gamma_1+\gamma_2)}. &(138)
\end{aligned}$$

We check that when $\gamma_1 = \gamma_2 = \gamma$, we have $\mathrm{JS}^{H_\alpha}(p_\gamma:p_\gamma) = 0$.

Theorem 4 (Harmonic JSD between scale Cauchy distributions)**.** *The harmonic Jensen–Shannon divergence between two scale Cauchy distributions p_{γ_1} and p_{γ_2} is $\mathrm{JS}^H(p_{\gamma_1}:p_{\gamma_2}) = \log\frac{(3\gamma_1+\gamma_2)(3\gamma_2+\gamma_1)}{8\sqrt{\gamma_1\gamma_2}(\gamma_1+\gamma_2)}$.*

Let us report some numerical examples: Consider $p_{\gamma_1} = 0.1$ and $p_{\gamma_1} = 0.5$, we find that $\mathrm{JS}^H(p_{\gamma_1}:p_{\gamma_2}) \simeq 0.176$. When $p_{\gamma_1} = 0.2$ and $p_{\gamma_1} = 0.8$, we find that $\mathrm{JS}^H(p_{\gamma_1}:p_{\gamma_2}) \simeq 0.129$.

Notice that KL formula is scale-invariant, and this property holds for any scale family:

Lemma 1. *The Kullback–Leibler divergence between two distributions p_{s_1} and p_{s_2} belonging to the same scale family $\{p_s(x) = \frac{1}{s}p(\frac{x}{s})\}_{s\in(0,\infty)}$ with standard density p is scale-invariant: $\mathrm{KL}(p_{\lambda s_1}:p_{\lambda s_2}) = \mathrm{KL}(p_{s_1}:p_{s_2}) = \mathrm{KL}(p:p_{\frac{s_2}{s_1}}) = \mathrm{KL}(p_{\frac{s_1}{s_2}}:p)$ for any $\lambda > 0$.*

A direct proof follows from a change of variable in the KL integral with $y = \frac{x}{\lambda}$ and $\mathrm{d}x = \lambda\mathrm{d}y$. Please note that although the KLD between scale Cauchy distributions is symmetric, it is not the case for all scale families: For example, the Rayleigh distributions form a scale family with the KLD amounting to compute a Bregman asymmetric Itakura–Saito divergence between parameters [37].

Instead of the KLD, we can choose the total variation distance for which a formula has been reported in [38] between two Cauchy distributions. Notice that the Cauchy distributions are alpha-stable distributions for $\alpha = 1$ and q Gaussian distributions for $q = 2$ ([58], p. 104). A closed-form formula for the divergence between two q-Gaussians is given in [58] when $q < 2$. The definite integral $h_q(p) = \int_{-\infty}^{+\infty} p(x)^q \mathrm{d}\mu$ is available in closed form for Cauchy distributions. When $q = 2$, we have $h_2(p_\gamma) = \frac{1}{2\pi\gamma}$.

We refer to [38] for yet other illustrative examples considering the family of Pearson type VII distributions and central multivariate t-distributions which use the power means (quasi-arithmetic means M^h induced by $h(u) = u^\alpha$ for $\alpha > 0$) for defining mixtures.

Table 1 summarizes the various examples introduced in the paper.

Table 1. Summary of the weighted means M chosen according to the parametric family in order to ensure that the family is closed under M-mixturing: $(p_{\theta_1} p_{\theta_2})_\alpha^M = p_{(\theta_1\theta_2)_\alpha}$.

$\mathbf{JS}^{M_\alpha}$	**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	$Z_\alpha^H(\theta_1 : \theta_2) = \sqrt{\frac{\theta_1\theta_2}{(\theta_1\theta_2)_\alpha(\theta_1\theta_2)_{1-\alpha}}}$

4.3. The M-Jensen–Shannon Matrix Distances

In this section, we consider distances between matrices which play an important role in quantum computing [59,60]. We refer to [61] for the matrix Jensen–Bregman logdet divergence. The *Hellinger distance* can be interpreted as the difference of an arithmetic mean A and a geometric mean G:

$$D_H(p,q) = \sqrt{1 - \int_{\mathcal{X}} \sqrt{p(x)}\sqrt{q(x)} \mathrm{d}\mu(x)} = \sqrt{\int_{\mathcal{X}} (A(p(x),q(x)) - G(p(x),q(x)))\mathrm{d}\mu(x)}. \tag{139}$$

Notice that since $A \geq G$, we have $D_H(p,q) \geq 0$. The scaled and squared Hellinger distance is an α-divergence I_α for $\alpha = 0$. Recall that the α-divergence can be interpreted as the difference of a weighted arithmetic minus a weighted geometry mean.

In general, if a mean M_1 dominates a mean M_2, we may define the distance as

$$D_{M_1,M_2}(p,q) = \int_{\mathcal{X}} (M_1(p,q) - M_2(p,q))\, \mathrm{d}\mu(x). \tag{140}$$

When considering matrices [62], there is *not* a unique definition of a geometric matrix mean, and thus we have different notions of matrix Hellinger distances [62], some of them are divergences (smooth distances defining a dualistic structure in information geometry).

We define the *matrix M-Jensen–Shannon divergence* for a matrix divergence [63,64] D as follows:

$$\mathrm{JS}_D^M(X_1, X_2) = \frac{1}{2}\left(D(X_1 : M(X_1,X_2)) + D(X_2 : M(X_1,X_2))\right) = \mathrm{JS}_D^M(X_2, X_1). \tag{141}$$

For example, we can choose the *von Neumann matrix divergence* [63]:

$$D_{\mathrm{vN}}(X_1 : X_2) := \mathrm{tr}\left(X_1 \log X_1 - X_1 \log X_2 - X_1 + X_2\right), \tag{142}$$

or the *LogDet matrix divergence* [63]:

$$D_{\mathrm{ld}}(X_1 : X_2) := \mathrm{tr}(X_1 X_2^{-1}) - \log|X_1 X_2^{-1}| - d, \tag{143}$$

where square matrices X_1 and X_2 have dimension d.

5. Conclusions and Perspectives

We introduced a generalization of the celebrated Jensen–Shannon divergence [6], termed the (M,N)*-Jensen–Shannon divergences*, based on *M-mixtures* derived from abstract means M. This new family of divergences includes the ordinary Jensen–Shannon divergence when both M and N are set to the arithmetic mean. We reported closed-form expressions of the M Jensen–Shannon divergences for mixture families and exponential families in information geometry by choosing the arithmetic and geometric weighted mean, respectively. The α-skewed geometric Jensen–Shannon divergence (G-Jensen–Shannon divergence) between densities p_{θ_1} and p_{θ_2} of the same exponential family with cumulant function F is

$$\mathrm{JS}_{\mathrm{KL}}^{G_\alpha}[p_{\theta_1} : p_{\theta_2}] = \mathrm{JS}_{B_{F^*}}^{A_\alpha}(\theta_1 : \theta_2).$$

Here, we used the bracket notation to emphasize that the statistical distance $\mathrm{JS}_{\mathrm{KL}}^{G_\alpha}$ is between densities, and the parenthesis notation to emphasize that the distance $\mathrm{JS}_{B_{F^*}}^{A_\alpha}$ is between parameters. We also have $\mathrm{JS}_{\mathrm{KL}^*}^{G_\alpha}[p_{\theta_1} : p_{\theta_2}] = J_F^\alpha(\theta_1 : \theta_2)$. We also show how to get a closed-form formula for the harmonic Jensen–Shannon divergence of Cauchy scale distributions by taking harmonic mixtures.

For an arbitrary distance D, we define the *skew N-Jeffreys symmetrization*:

$$J_D^{N_\beta}(p_1 : p_2) = N_\beta(D(p_1 : p_2), D(p_2 : p_1)), \tag{144}$$

and the *skew* (M,N)*-JS-symmetrization*:

$$\mathrm{JS}_D^{M_\alpha,N_\beta}(p_1 : p_2) = N_\beta(D(p_1 : (p_1p_2)_\alpha^M), D(p_2 : (p_1p_2)_\alpha^M)). \tag{145}$$

A Java$^{\mathrm{TM}}$ source code for computing the geometric Jensen–Shannon divergence between multivariate Gaussian distributions is available online at https://franknielsen.github.io/M-JS/.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Summary of Distances and Their Notations

Table A1 lists the main distances with their notations.

Table A1. Summary of Distances and Their Notations.

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 $Z_\alpha^M(p,q) = \int_{t\in\mathcal{X}} M_\alpha(p(t),q(t))\mathrm{d}\mu(t)$
Statistical distance	$D(p:q)$
Dual/reverse distance D^*	$D^*(p:q){:=}D(q:p)$
Kullback-Leibler divergence	$\mathrm{KL}(p:q) = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x)$
reverse Kullback-Leibler divergence	$\mathrm{KL}^*(p:q) = \mathrm{KL}(q:p) = \int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}\mu(x)$
Jeffreys divergence	$J(p;q) = \mathrm{KL}(p:q) + \mathrm{KL}(q:p) = \int(p(x)-q(x))\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x)$
Resistor divergence	$\frac{1}{R(p;q)} = \frac{1}{2}\left(\frac{1}{\mathrm{KL}(p:q)} + \frac{1}{\mathrm{KL}(q:p)}\right)$. $R(p;q) = \frac{2J(p;q)}{\mathrm{KL}(p:q)\mathrm{KL}(q:p)}$
skew K-divergence	$K_\alpha(p:q) = \int p(x)\log\frac{p(x)}{(1-\alpha)p(x)+\alpha q(x)}\mathrm{d}\mu(x)$
Jensen-Shannon divergence	$\mathrm{JS}(p,q) = \frac{1}{2}\left(\mathrm{KL}\left(p : \frac{p+q}{2}\right) + \mathrm{KL}\left(q : \frac{p+q}{2}\right)\right)$
skew Bhattacharrya divergence	$B_\alpha(p:q) = -\log\int_{\mathcal{X}} p(x)^{1-\alpha}q(x)^\alpha\mathrm{d}\mu(x)$
Hellinger distance	$D_H(p,q) = \sqrt{1 - \int_{\mathcal{X}}\sqrt{p(x)}\sqrt{q(x)}\mathrm{d}\mu(x)}$
α-divergences	$I_\alpha(p:q) = \int\left(\alpha p(x) + (1-\alpha)q(x) - p(x)^\alpha q(x)^{1-\alpha}\right)\mathrm{d}\mu(x), \alpha \notin \{0,1\}$ $I_\alpha(p:q) = A_\alpha(q:p) - G_\alpha(q:p)$
Mahalanobis distance	$D_Q(p:q) = \sqrt{(p-q)^\top Q(p-q)}$ for a positive-definite matrix $Q \succ 0$
f-divergence	$I_f(p:q) = \int p(x)f\left(\frac{q(x)}{p(x)}\right)\mathrm{d}\mu(x)$, with $f(1) = f'(1) = 0$ f strictly convex at 1
reverse f-divergence	$I_f^*(p:q) = \int q(x)f\left(\frac{p(x)}{q(x)}\right)\mathrm{d}\mu(x) = I_{f^\diamond}(p:q)$ for $f^\diamond(u) = uf(\frac{1}{u})$
J-symmetrized f-divergence	$J_f(p;q) = \frac{1}{2}(I_f(p:q) + I_f(q:p))$
JS-symmetrized f-divergence	$I_f^\alpha(p;q){:=}(1-\alpha)I_f(p:(pq)_\alpha) + \alpha I_f(q:(pq)_\alpha) = I_{f_\alpha^{\mathrm{JS}}}(p:q)$ for $f_\alpha^{\mathrm{JS}}(u){:=}(1-\alpha)f(\alpha u + 1 - \alpha) + \alpha f\left(\alpha + \frac{1-\alpha}{u}\right)$

Table A1. *Cont.*

Parameter distance	
Bregman divergence	$B_F(\theta:\theta'):=F(\theta)-F(\theta')-\langle\theta-\theta',\nabla F(\theta')\rangle$
skew Jeffreys-Bregman divergence	$S_F^\alpha=(1-\alpha)B_F(\theta:\theta')+\alpha B_F(\theta':\theta)$
skew Jensen divergence	$J_F^\alpha(\theta:\theta'):=(F(\theta)F(\theta'))_\alpha-F((\theta\theta')_\alpha)$
Jensen-Bregman divergence	$\mathrm{JB}_F(\theta;\theta')=\frac{1}{2}\left(B_F\left(\theta:\frac{\theta+\theta'}{2}\right)+B_F\left(\theta':\frac{\theta+\theta'}{2}\right)\right)=J_F(\theta;\theta').$
Generalized Jensen-Shannon divergences	
skew J-symmetrization	$J_D^\alpha(p:q):=(1-\alpha)D\,(p:q)+\alpha D\,(q:p)$
skew JS-symmetrization	$\mathrm{JS}_D^\alpha(p:q):=(1-\alpha)D\,(p:(1-\alpha)p+\alpha q)+\alpha D\,(q:(1-\alpha)p+\alpha q)$
skew M-Jensen-Shannon divergence	$\mathrm{JS}^{M_\alpha}(p:q):=(1-\alpha)\mathrm{KL}\left(p:(pq)_\alpha^M\right)+\alpha\mathrm{KL}\left(q:(pq)_\alpha^M\right)$
skew M-JS-symmetrization	$\mathrm{JS}_D^{M_\alpha}(p:q):=(1-\alpha)D\left(p:(pq)_\alpha^M\right)+\alpha D\left(q:(pq)_\alpha^M\right)$
N-Jeffreys divergence	$J^{N_\beta}(p:q):=N_\beta(\mathrm{KL}\,(p:q)\,,\mathrm{KL}\,(q:p))$
N-J D divergence	$J_D^{N_\beta}(p:q)=N_\beta(D(p:q),D(q:p))$
skew (M,N)-D JS divergence	$\mathrm{JS}_D^{M_\alpha,N_\beta}(p:q):=N_\beta\left(D\left(p:(pq)_\alpha^M\right),D\left(q:(pq)_\alpha^M\right)\right)$

Appendix B. Symbolic Calculations in Maxima

The program below calculates the normalizer Z for the harmonic H-mixtures of Cauchy distributions (Equation (133)).

```
assume(gamma>0);
Cauchy(x,gamma) := gamma/(%pi*(x**2+gamma**2));
assume(alpha>0);
assume(alpha<1);
h(x,y,alpha) :=  (x*y)/((1-alpha)*y+alpha*x);
assume(gamma1>0);
assume(gamma2>0);
m(x,alpha) := ratsimp(h(Cauchy(x,gamma1),Cauchy(x,gamma2),alpha));
/* calculate Z */
integrate(m(x,alpha),x,-inf,inf);
```

References

1. Billingsley, P. *Probability and Measure*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
3. Ho, S.W.; Yeung, R.W. On the discontinuity of the Shannon information measures. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Adelaide, Australia, 4–9 September 2005; pp. 159–163.
4. Nielsen, F. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Process. Lett.* **2013**, *20*, 657–660. [CrossRef]
5. Johnson, D.; Sinanovic, S. Symmetrizing the Kullback-Leibler Distance. Technical report of Rice University (US). 2001. Available online: https://scholarship.rice.edu/handle/1911/19969 (accessed on 11 May 2019).
6. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
7. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [CrossRef]
8. Vajda, I. On metric divergences of probability measures. *Kybernetika* **2009**, *45*, 885–900.
9. Fuglede, B.; Topsoe, F. Jensen-Shannon divergence and Hilbert space embedding. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Waikiki, HI, USA, 29 June–4 July 2014; p. 31.
10. Sims, G.E.; Jun, S.R.; Wu, G.A.; Kim, S.H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 2677–2682. [CrossRef]
11. DeDeo, S.; Hawkins, R.X.; Klingenstein, S.; Hitchcock, T. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy* **2013**, *15*, 2246–2276. [CrossRef]
12. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.

13. Wang, Y.; Woods, K.; McClain, M. Information-theoretic matching of two point sets. *IEEE Trans. Image Process.* **2002**, *11*, 868–872. [CrossRef]
14. Peter, A.M.; Rangarajan, A. Information geometry for landmark shape analysis: Unifying shape representation and deformation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 337–350. [CrossRef] [PubMed]
15. Nielsen, F.; Sun, K. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy* **2016**, *18*, 442. [CrossRef]
16. Wang, F.; Syeda-Mahmood, T.; Vemuri, B.C.; Beymer, D.; Rangarajan, A. Closed-form Jensen-Rényi divergence for mixture of Gaussians and applications to group-wise shape registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer: Berlin, Germany, 2009; pp. 648–655.
17. Nielsen, F. Closed-form information-theoretic divergences for statistical mixtures. In Proceedings of the IEEE 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 1723–1726.
18. Amari, S.I. *Information Geometry and Its Applications*; Springer: Berlin, Germany, 2016.
19. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
20. Eguchi, S. Geometry of minimum contrast. *Hiroshima Math. J.* **1992**, *22*, 631–647. [CrossRef]
21. Amari, S.I.; Cichocki, A. Information geometry of divergence functions. *Bull. Pol. Acad. Sci. Tech. Sci.* **2010**, *58*, 183–195. [CrossRef]
22. Ciaglia, F.M.; Di Cosmo, F.; Felice, D.; Mancini, S.; Marmo, G.; Pérez-Pardo, J.M. Hamilton-Jacobi approach to potential functions in information geometry. *J. Math. Phys.* **2017**, *58*, 063506. [CrossRef]
23. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
24. Nielsen, F. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv* **2010**, arXiv:1009.4004.
25. Chen, P.; Chen, Y.; Rao, M. Metrics defined by Bregman divergences. *Commun. Math. Sci.* **2008**, *6*, 915–926. [CrossRef]
26. Chen, P.; Chen, Y.; Rao, M. Metrics defined by Bregman divergences: Part 2. *Commun. Math. Sci.* **2008**, *6*, 927–948. [CrossRef]
27. Kafka, P.; Österreicher, F.; Vincze, I. On powers of f-divergences defining a distance. *Stud. Sci. Math. Hung.* **1991**, *26*, 415–422.
28. Österreicher, F.; Vajda, I. A new class of metric divergences on probability spaces and its applicability in statistics. *Ann. Inst. Stat. Math.* **2003**, *55*, 639–653. [CrossRef]
29. Nielsen, F.; Nock, R. On the geometry of mixtures of prescribed distributions. In Proceeding of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 Aprli 2018; pp. 2861–2865.
30. Nielsen, F.; Hadjeres, G. Monte Carlo Information Geometry: The dually flat case. *arXiv* **2018**, arXiv:1803.07225.
31. Watanabe, S.; Yamazaki, K.; Aoyagi, M. Kullback information of normal mixture is not an analytic function. *IEICE Tech. Rep. Neurocomput.* **2004**, *104*, 41–46.
32. Nielsen, F.; Nock, R. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Process. Lett.* **2014**, *21*, 10–13. [CrossRef]
33. Nielsen, F.; Hadjeres, G. On power chi expansions of f-divergences. *arXiv* **2019**, arXiv:1903.05818.
34. Niculescu, C.; Persson, L.E. *Convex Functions and Their Applications*, 2nd ed.; Springer: Berlin, Germany, 2018.
35. Rényi, A. 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: Oakland, CA, USA, 1961.
36. McLachlan, G.J.; Lee, S.X.; Rathnayake, S.I. Finite mixture models. *Ann. Rev. Stat. Appl.* **2019**, *6*, 355–378. [CrossRef]
37. Nielsen, F.; Garcia, V. Statistical exponential families: A digest with flash cards. *arXiv* **2009**, arXiv:0911.4863.
38. Nielsen, F. Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. *Pattern Recognit. Lett.* **2014**, *42*, 25–34. [CrossRef]

39. Eguchi, S.; Komori, O. Path connectedness on a space of probability density functions. In *Geometric Science of Information (GSI)*; Springer: Cham, Switzerland, 2015; pp. 615–624. [CrossRef]
40. Eguchi, S.; Komori, O.; Ohara, A. Information geometry associated with generalized means. In *Information Geometry and its Applications IV*; Springer: Berlin, Germany, 2016; pp. 279–295.
41. Asadi, M.; Ebrahimi, N.; Kharazmi, O.; Soofi, E.S. Mixture models, Bayes Fisher information, and divergence measures. *IEEE Trans. Inf. Theory* **2019**, *65*, 2316–2321. [CrossRef]
42. Amari, S.I. Integration of stochastic models by minimizing α-divergence. *Neural Comput.* **2007**, *19*, 2780–2796. [CrossRef]
43. Nielsen, F.; Nock, R. Generalizing skew Jensen divergences and Bregman divergences with comparative convexity. *IEEE Signal Process. Lett.* **2017**, *24*, 1123–1127. [CrossRef]
44. Lee, L. Measures of distributional similarity. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 20–26 June 1999; pp. 25–32. [CrossRef]
45. Nielsen, F. The statistical Minkowski distances: Closed-form formula for Gaussian mixture models. *arXiv* **2019**, arXiv:1901.03732.
46. Zhang, J. Reference duality and representation duality in information geometry. *AIP Conf. Proc.* **2015**, *1641*, 130–146.
47. Yoshizawa, S.; Tanabe, K. Dual differential geometry associated with the Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations. *SUT J. Math.* **1999**, *35*, 113–137.
48. Nielsen, F.; Nock, R. A closed-form expression for the Sharma–Mittal entropy of exponential families. *J. Phys. A Math. Theor.* **2011**, *45*, 032003. [CrossRef]
49. Nielsen, F. An elementary introduction to information geometry. *arXiv* **2018**, arXiv:1808.08271.
50. Nielsen, F.; Nock, R. Optimal interval clustering: Application to Bregman clustering and statistical mixture learning. *IEEE Signal Process. Lett.* **2014**, *21*, 1289–1292. [CrossRef]
51. Arthur, D.; Vassilvitskii, S. *k*-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics*; ACM: New York, NY, USA, 2007; pp. 1027–1035.
52. Nielsen, F.; Nock, R.; Amari, S.I. On clustering histograms with *k*-means by using mixed α-divergences. *Entropy* **2014**, *16*, 3273–3301. [CrossRef]
53. Nielsen, F.; Nock, R. Total Jensen divergences: definition, properties and clustering. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, Australia, 19–24 August 2015; pp. 2016–2020.
54. Ackermann, M.R.; Blömer, J. Bregman clustering for separable instances. In *Scandinavian Workshop on Algorithm Theory*; Springer: Berlin, Germany, 2010; pp. 212–223.
55. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904. [CrossRef]
56. Tzagkarakis, G.; Tsakalides, P. A statistical approach to texture image retrieval via alpha-stable modeling of wavelet decompositions. In Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services, Instituto Superior Técnico, Lisboa, Portugal, 21–23 April 2004; pp. 21–23.
57. Boissonnat, J.D.; Nielsen, F.; Nock, R. Bregman Voronoi diagrams. *Discrete Comput. Geom.* **2010**, *44*, 281–307. [CrossRef]
58. Naudts, J. *Generalised Thermostatistics*; Springer Science & Business Media: Berlin, Germany, 2011.
59. Briët, J.; Harremoës, P. Properties of classical and quantum Jensen-Shannon divergence. *Phys. Rev. A* **2009**, *79*, 052311. [CrossRef]
60. Audenaert, K.M. Quantum skew divergence. *J. Math. Phys.* **2014**, *55*, 112202. [CrossRef]
61. Cherian, A.; Sra, S.; Banerjee, A.; Papanikolopoulos, N. Jensen-Bregman logdet divergence with application to efficient similarity search for covariance matrices. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2161–2174. [CrossRef]
62. Bhatia, R.; Jain, T.; Lim, Y. Strong convexity of sandwiched entropies and related optimization problems. *Rev. Math. Phys.* **2018**, *30*, 1850014. [CrossRef]

63. Kulis, B.; Sustik, M.A.; Dhillon, I.S. Low-rank kernel learning with Bregman matrix divergences. *J. Mach. Learn. Res.* **2009**, *10*, 341–376.
64. Nock, R.; Magdalou, B.; Briys, E.; Nielsen, F. Mining matrix data with Bregman matrix divergences for portfolio selection. In *Matrix Information Geometry*; Springer: Berlin, Germany, 2013; pp. 373–402.