entropy

Article

On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius

Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan; Frank.Nielsen@acm.org

Abstract: We generalize the Jensen-Shannon divergence and the Jensen-Shannon diversity index by considering a variational definition with respect to a generic mean, thereby extending the notion of Sibson's information radius. The variational definition applies to any arbitrary distance and yields a new way to define a Jensen-Shannon symmetrization of distances. When the variational optimization is further constrained to belong to prescribed families of probability measures, we get relative Jensen-Shannon divergences and their equivalent Jensen-Shannon symmetrizations of distances that generalize the concept of information projections. Finally, we touch upon applications of these variational Jensen-Shannon divergences and diversity indices to clustering and quantization tasks of probability measures, including statistical mixtures.

Keywords: Jensen-Shannon divergence; diversity index; Rényi entropy; information radius; information projection; exponential family; Bregman divergence; Fenchel–Young divergence; Bregman information; q-exponential family; q-divergence; Bhattacharyya distance; centroid; clustering

Citation: Nielsen, F. On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius. *Entropy* **2021**, *23*, 464. https://doi.org/10.3390/e23040464

Academic Editor: Yong Deng

Received: 12 March 2021
Accepted: 9 April 2021
Published: 14 April 2021

1. Introduction: Background and Motivations

The goal of the author is to methodologically contribute to an extension of the Sibson's information radius [1] and also concentrate on analysis of the specified families of distributions called exponential families [2].

Let $(\mathcal{X}, \mathcal{F})$ denote a measurable space [3] with sample space $\mathcal{X}$ and σ-algebra $\mathcal{F}$ on the set $\mathcal{X}$. The Jensen-Shannon divergence [4] (JSD) between two probability measures P and Q (or probability distributions) on $(\mathcal{X}, \mathcal{F})$ is defined by:

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

where D_{KL} denotes the Kullback–Leibler divergence [5,6] (KLD):

$$D_{\mathrm{KL}}[P : Q] := \begin{cases} \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}P(x)}{\mathrm{d}Q(x)}\right)\mathrm{d}P, & P \ll Q \\ +\infty, & P \not\ll Q \end{cases} \tag{2}$$

where $P \ll Q$ means that P is absolutely continuous with respect to Q [3], and $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon–Nikodym derivative of P with respect to Q. Equation (2) can be rewritten using the chain rule as:

$$D_{\mathrm{KL}}[P : Q] := \begin{cases} \int_{\mathcal{X}} \frac{\mathrm{d}P(x)}{\mathrm{d}Q(x)} \log\left(\frac{\mathrm{d}P(x)}{\mathrm{d}Q(x)}\right)\mathrm{d}Q, & P \ll Q \\ +\infty, & P \not\ll Q \end{cases} \tag{3}$$

Consider a measure μ for which both the Radon–Nikodym derivatives $p := \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q := \frac{\mathrm{d}P}{\mathrm{d}\mu}$ exist (e.g., $\mu = \frac{P+Q}{2}$). Subsequently the Kullback–Leibler divergence can be rewritten as (see Equation (2.5) page 5 of [5] and page 251 of the Cover & Thomas' textbook [6]):

$$D_{\mathrm{KL}}[p:q] := \int_{\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}\mu(x). \tag{4}$$

Denote by $\mathcal{D} = \mathcal{D}(\mathcal{X})$ the set of all densities with full support $\mathcal{X}$ (Radon–Nikodym derivatives of probability measures with respect to μ):

$$\mathcal{D}(\mathcal{X}) := \left\{p : \mathcal{X} \to \mathbb{R} : p(x) > 0\ \mu\text{-almost everywhere}, \int_{\mathcal{X}} p(x)\mathrm{d}\mu(x) = 1\right\}.$$

Subsequently, the Jensen-Shannon divergence [4] between two densities p and q of $\mathcal{D}$ is defined by:

$$D_{\mathrm{JS}}[p,q] := \frac{1}{2}\left(D_{\mathrm{KL}}\left[p : \frac{p+q}{2}\right] + D_{\mathrm{KL}}\left[q : \frac{p+q}{2}\right]\right). \tag{5}$$

Often, one considers the Lebesgue measure [3] $\mu = \mu_{\mathcal{L}}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel σ-algebra, or the counting measure [3] $\mu = \mu_{\#}$ on $(\mathcal{X}, 2^{\mathcal{X}})$ where $\mathcal{X}$ is a countable set, for defining the measure space $(\mathcal{X}, \mathcal{F}, \mu)$.

The JSD belongs to the class of f-divergences [7–9] which are known as the invariant decomposable divergences of information geometry (see [10], pp. 52–57). Although the KLD is asymmetric (i.e., $D_{\mathrm{KL}}[p:q] \neq D_{\mathrm{KL}}[q:p]$), the JSD is symmetric (i.e., $D_{\mathrm{JS}}[p,q] = D_{\mathrm{JS}}[q,p]$). The notation ':' is used as a parameter separator to indicate that the parameters are not permutation invariant, and that the order of parameters is important.

In this work, a distance $D(O_1 : O_2)$ is a measure of dissimilarity between two objects O_1 and O_2, which do not need to be symmetric or satisfy the triangle inequality of metric distances. A distance only satisfies the identity of indiscernibles: $D(O_1 : O_2) = 0$ if and only if $O_1 = O_2$. When the objects O_1 and O_2 are probability densities with respect to μ, we call this distance a statistical distance, use the brackets to enclose the arguments of the statistical distance (i.e., $D[O_1 : O_2]$), and we have $D[O_1 : O_2] = 0$ if and only if $O_1(x) = O_2(x)$ μ-almost everywhere.

The 2-point JSD of Equation (4) can be extended to a weighted set of n densities $\mathcal{P} := \{(w_1, p_1), \ldots, (w_n, p_n)\}$ (with positive w_i's normalized to sum up to unity, i.e., $\sum_{i=1}^n w_i = 1$) thus providing a diversity index, i.e., a n-point JSD for $\mathcal{P}$:

$$D_{\mathrm{JS}}(\mathcal{P}) := \sum_{i=1}^n w_i D_{\mathrm{KL}}[p_i : \bar{p}], \tag{6}$$

where $\bar{p} := \sum_{i=1}^n w_i p_i$ denotes the statistical mixture [11] of the densities of $\mathcal{P}$. We have $D_{\mathrm{JS}}[p:q] = D_{\mathrm{JS}}(\{(\frac{1}{2}, p), (\frac{1}{2}, q)\})$. We call $D_{\mathrm{JS}}(\mathcal{P})$ the Jensen-Shannon diversity index.

The KLD is also called the relative entropy since it can be expressed as the difference between the cross entropy $h[p:q]$ and the entropy $h[p]$:

$$\begin{aligned} D_{\mathrm{KL}}[p:q] &:= \int_{\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}\mu(x) && (7)\\ &= \int_{\mathcal{X}} p(x)\log p(x)\mathrm{d}\mu(x) - \int_{\mathcal{X}} p(x)\log q(x)\mathrm{d}\mu(x), && (8)\\ &= h[p:q] - h[p], && (9) \end{aligned}$$

with the cross-entropy and entropy defined, respectively, by

$$\begin{aligned} h[p:q] &:= -\int_{\mathcal{X}} p(x)\log q(x)\mathrm{d}\mu(x), && (10)\\ h[p] &:= -\int_{\mathcal{X}} p(x)\log p(x)\mathrm{d}\mu(x). && (11) \end{aligned}$$

Because $h[p] = h[p:p]$, we may say that the entropy is the self-cross-entropy.

When μ is the Lebesgue measure, the Shannon entropy is also called the differential entropy [6]. Although the discrete entropy $H[p] = -\sum_i p_i \log p_i$ (i.e., entropy with respect to the counting measure) is always positive and bounded by $\log |\mathcal{X}|$, the differential entropy may be negative (e.g., entropy of a Gaussian distribution with small variance).

The Jensen-Shannon divergence of Equation (6) can be rewritten as:

$$D_{\mathrm{JS}}[p,q] = h[\bar{p}] - \sum_{i=1}^{n} w_i h[p_i] := J_{-h}[p,q]. \tag{12}$$

The JSD representation of Equation (12) is a Jensen divergence [12] for the strictly convex negentropy $F(p) = -h[p]$, since the entropy function $h[.]$ is strictly concave. Therefore, it is appropriate to call this divergence the Jensen-Shannon divergence.

Because $\frac{p_i(x)}{\bar{p}(x)} \leq \frac{p_i(x)}{w_i p_i(x)} = \frac{1}{w_i}$, it can be shown that the Jensen-Shannon diversity index is upper bounded by $H(w) := -\sum_{i=1}^{n} w_i \log w_i$, the discrete Shannon entropy. Thus, the Jensen-Shannon diversity index is bounded by $\log n$, and the 2-point JSD is bounded by $\log 2$, although the KLD is unbounded and it may even be equal to $+\infty$ when the definite integral diverges (e.g., KLD between the standard Cauchy distribution and the standard Gaussian distribution). Another nice property of the JSD is that its square root yields a metric distance [13,14]. This property further holds for the quantum JSD [15]. The JSD has gained interest in machine learning. See, for example, the Generative Adversarial Networks [16] (GANs) in deep learning [17], where it was proven that minimizing the GAN objective function by adversarial training is equivalent to minimizing a JSD.

To delineate the different roles that are played by the factor $\frac{1}{2}$ in the ordinary Jensen-Shannon divergence (i.e., in weighting the two KLDs and in weighting the two densities), let us introduce two scalars $\alpha, \beta \in (0,1)$, and define a generic (α,β)-skewed Jensen-Shannon divergence, as follows:

$$\begin{aligned} D_{\mathrm{JS},\alpha,\beta}[p:q] &:= (1-\beta) D_{\mathrm{KL}}[p:m_\alpha] + \beta D_{\mathrm{KL}}[q:m_\alpha], && (13)\\ &= (1-\beta) h[p:m_\alpha] + \beta h[q:m_\alpha] - (1-\beta) h[p] - \beta h[q], && (14)\\ &= h[m_\beta : m_\alpha] - ((1-\beta) h[p] + \beta h[q]), && (15) \end{aligned}$$

where $m_\alpha := (1-\alpha)p + \alpha q$ and $m_\beta := (1-\beta)p + \beta q$. This identity holds, because $D_{\mathrm{JS},\alpha,\beta}$ is bounded by $(1-\beta)\log\frac{1}{1-\alpha} + \beta \log\frac{1}{\alpha}$, see [18]. Thus, when $\beta = \alpha$, we have $D_{\mathrm{JS},\alpha}[p,q] = D_{\mathrm{JS},\alpha,\alpha}[p,q] = h[m_\alpha] - ((1-\alpha)h[p] + \alpha h[q])$, since the self-cross entropy corresponds to the entropy: $h[m_\alpha : m_\alpha] = h[m_\alpha]$.

A f-divergence [9,19,20] is defined for a convex generator f, which is strictly convex at 1 (to satisfy the identity of the indiscernibles) and that satisfies $f(1) = 0$, by

$$I_f[p:q] := \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x) \geq f(1) = 0, \tag{16}$$

where the right-hand-side follows from Jensen's inequality [20]. For example, the total variation distance $D_{\mathrm{TV}}[p:q] = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)| \mathrm{d}\mu(x)$ is a f-divergence for the generator $f_{\mathrm{TV}}(u) = |u-1|$: $D_{\mathrm{TV}}[p:q] = I_{f_{\mathrm{TV}}}[p:q]$. The generator $f_{\mathrm{TV}}(u)$ is convex on $\mathbb{R}$, strictly convex at 1, and it satisfies $f(u) = 1$.

The $D_{\mathrm{JS},\alpha,\beta}$ divergence is a f-divergence

$$D_{\mathrm{JS},\alpha,\beta}[p:q] = I_{f_{\mathrm{JS},\alpha,\beta}}[p:q], \tag{17}$$

for the generator:

$$f_{\mathrm{JS},\alpha,\beta}(u) = -\left((1-\beta)\log(\alpha u + (1-\alpha)) + \beta u \log\left(\frac{1-\alpha}{u} + \alpha\right)\right). \tag{18}$$

We check that the generator $f_{\mathrm{JS},\alpha,\beta}$ is strictly convex, since, for any $a \in (0,1)$ and $b \in (0,1)$, we have

$$f''_{\mathrm{JS},\alpha,\beta}(u) = \frac{a^2(1-b)u + (a-1)^2 b}{a^2u^3 + 2a(1-a)u^2 + (a-1)^2 u} > 0, \tag{19}$$

when $u > 0$.

The Jensen-Shannon principle of taking the average of the (Kullback–Leibler) divergences between the source parameters to the mid-parameter can be applied to other distances. For example, the Jensen–Bregman divergence is a Jensen-Shannon symmetrization of the Bregman divergence B_F [12]:

$$B_F^{\mathrm{JS}}(\theta_1 : \theta_2) := \frac{1}{2}\left(B_F\left(\theta_1 : \frac{\theta_1+\theta_2}{2}\right) + B_F\left(\theta_2 : \frac{\theta_1+\theta_2}{2}\right)\right), \tag{20}$$

where the Bregman divergence [21] B_F is defined by

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta'). \tag{21}$$

The Jensen–Bregman divergence B_F^{JS} can also be written as an equivalent Jensen divergence J_F:

$$B_F^{\mathrm{JS}}(\theta_1 : \theta_2) = J_F(\theta_1 : \theta_2) := \frac{F(\theta_1)+F(\theta_2)}{2} - F\left(\frac{\theta_1+\theta_2}{2}\right), \tag{22}$$

where F is a strictly convex function ensuring $J_F(\theta_1 : \theta_2) \geq 0$ with equality if $\theta_1 = \theta_2$.

Because of its use in various fields of information sciences [22], various generalizations of the JSD have been proposed: These generalizations are either based on Equation (5) [23] or Equation (12) [18,24,25]. For example, the (arithmetic) mixture $\bar{p} = \sum_i w_i p_i$ in Equation (6) was replaced by an abstract statistical mixture with respect to a generic mean M in [23] (e.g., the geometric mixture induced by the geometric mean), and the two KLDS defining the JSD in Equation (5) was further averaged using another abstract mean N, thus yielding the following generic (M,N)-*Jensen-Shannon divergence* [23] (abbreviated as (M,N)-JSD):

$$D_{\mathrm{JS}}^{M,N}[p:q] := N\left(D_{\mathrm{KL}}\left[p : (pq)^M_{\frac{1}{2}}\right], D_{\mathrm{KL}}\left[q : (pq)^M_{\frac{1}{2}}\right]\right), \tag{23}$$

where $(pq)^M_\alpha$ denotes the statistical weighted M-mixture:

$$(pq)^M_\alpha := \frac{M_\alpha(p(x),q(x))}{\int_{\mathcal{X}} M_\alpha(p(x),q(x))\mathrm{d}\mu(x)}. \tag{24}$$

Notice that, when $M = N = A$ (the arithmetic mean), Equation (23) of the (A,A)-JSD reduces to the ordinary JSD of Equation (5). When the means M and N are symmetric, the (M,N)-JSD is symmetric.

In general, a weighted mean $M_\alpha(a,b)$ for any $\alpha \in [0,1]$ shall satisfy the in-betweeness property [26] (i.e., a mean should be contained inside its extrema):

$$\min\{a,b\} \leq M_\alpha(a,b) \leq \max\{a,b\}. \tag{25}$$

The three Pythagorean means defined for positive scalars $a > 0$ and $b > 0$ are classic examples of means:

- The arithmetic mean $A(a,b) = \frac{a+b}{2}$,
- the geometric mean $G(a,b) = \sqrt{ab}$, and
- the harmonic mean $H(a,b) = \frac{2ab}{a+b}$.

These Pythagorean means may be interpreted as special instances of another parametric family of means: The power means

$$P_\alpha(a,b) := \left(\frac{a^\alpha + b^\alpha}{2}\right)^{\frac{1}{\alpha}}, \tag{26}$$

defined for $\alpha \in \mathbb{R}\backslash\{0\}$ (also called Hölder means). The power means can be extended to the full range $\alpha \in \mathbb{R}$ by using the property that $\lim_{\alpha\to 0} P_\alpha(a,b) = G(a,b)$. The power means are homogeneous means: $P_\alpha(\lambda a, \lambda b) = \lambda P_\alpha(a,b)$ for any $\lambda > 0$. We refer to the handbook of means [27] to obtain definitions and principles of other means beyond these power means.

A weighted mean (also called barycenter) can be built from a non-weighted mean $M(a,b)$ (i.e., $\alpha = \frac{1}{2}$) by using the dyadic expansion of the real weight $\alpha \in [0,1]$, see [28]. That is, we can define the weighted mean $M(p,q;w,1-w)$ for $w = \frac{i}{2^k}$ with $i \in \{0,\ldots,2^k\}$ and k an integer. For example, consider a symmetric mean $M(p,q) = M(q,p)$. Subsequently, we get the following weighted means when $k = 3$:

$$\begin{aligned}
M\left(p,q;\frac{0}{8}=0,\frac{8}{8}=1\right) &= q\\
M\left(p,q;\frac{1}{8},\frac{7}{8}\right) &= M(M(M(p,q),q),q)\\
M\left(p,q;\frac{2}{8}=\frac{1}{4},\frac{6}{8}=\frac{3}{4}\right) &= M(M(p,q),q)\\
M\left(p,q;\frac{3}{8},\frac{5}{8}\right) &= M(M(M(p,q),p),q)\\
M\left(p,q;\frac{4}{8}=\frac{1}{2},\frac{4}{8}=\frac{1}{2}\right) &= M(p,q)\\
M\left(p,q;\frac{5}{8},\frac{3}{8}\right) &= M(M(M(p,q),q),p)\\
M\left(p,q;\frac{6}{8}=\frac{3}{4},\frac{2}{8}=\frac{1}{4}\right) &= M(M(p,q),p)\\
M\left(p,q;\frac{7}{8},\frac{1}{8}\right) &= M(M(M(p,q),p),p)\\
M\left(p,q;\frac{8}{8}=1,\frac{0}{8}=0\right) &= p
\end{aligned}$$

Let $w = \sum_{i=1}^{\infty} \frac{d_i}{2^i}$ be the unique dyadic expansion of the real number $w \in (0,1)$, where the d_i's are binary digits (i.e., $d_i \in \{0,1\}$). We define the weighted mean $M(x,y;w,1-w)$ of two positive reals p and q for a real weight $w \in (0,1)$ as

$$M(x,y;w,1-w) := \lim_{n\to\infty} M\left(x,y;\sum_{i=1}^{n}\frac{d_i}{2^i},1-\sum_{i=1}^{n}\frac{d_i}{2^i}\right). \tag{27}$$

Choosing the abstract mean M in accordance with the family $\mathcal{R} = \{p_\theta : \theta \in \Theta\}$ of the densities allows one to obtain closed-form formula for the (M,N)-JSDs that rely on definite integral calculations [23]. For example, the JSD between two Gaussian densities does not admit a closed-form formula because of the log-sum integral, but the (G,N)-JSD admits a closed-form formula when using geometric statistical mixtures (i.e., when $M = G$). The calculus trick is to find a weighted mean M_α, such that, for two densities p_{θ_1} and p_{θ_2}, the weighted mean distribution $M_\alpha(p_{\theta_1}(x),p_{\theta_2}(x)) = \frac{p_{\theta_{1,2,\alpha}}(x)}{Z_{M_\alpha}(\theta_1,\theta_2)}$, where $Z_{M_\alpha}(\theta_1,\theta_2)$ is the normalizing coefficient and $p_{\theta_{1,2,\alpha}} \in \mathcal{R}$. Thus, the integral calculation can be simply calculated as $\int M_\alpha(p_{\theta_1}(x),p_{\theta_2}(x))\mathrm{d}\mu(x) = \frac{1}{Z_{M_\alpha}(\theta_1,\theta_2)}$ since $p_{\theta_{1,2,\alpha}}(x)$, and, therefore, $\int p_{\theta_{1,2,\alpha}}(x)\mathrm{d}\mu(x) = 1$. This trick has also been used in Bayesian hypothesis testing for upper bounding the probability of error between two densities of a parametric family of distributions by replacing the usual geometric mean (Section 11.7 of [6], page 375) by a more general quasi-arithmetic mean [29]. For example, the harmonic mean is well-suited to Cauchy distributions, and the power means to Student t-distributions [29].

As an application of these generalized JSDs, Deasy et al. [30] used the skewed geometric JSD (namely, the $(G_\alpha, A_{1-\alpha})$-JSD for $\alpha \in (0,1)$), which admits a closed-form formula between normal densities [23], and showed how regularizing an optimization task with this G-JSD divergence improved reconstruction and generation of Variational AutoEncoders (VAEs).

More generally, instead of using the KLD, one can also use any arbitrary distance D to define its JS-symmetrization, as follows:

$$D^{\mathrm{JS}}_{M,N}[p:q] := N\left(D\left[p:(pq)^M_{\frac{1}{2}}\right], D\left[q:(pq)^M_{\frac{1}{2}}\right]\right). \tag{28}$$

These symmetrizations may further be skewed by using M_α and/or N_β for $\alpha \in (0,1)$ and $\beta \in (0,1)$, yielding the definition [23]:

$$D^{\mathrm{JS}}_{M_\alpha,N_\beta}[p:q] := N_\beta\left(D\left[p:(pq)^M_\alpha\right], D\left[q:(pq)^M_\alpha\right]\right). \tag{29}$$

With these notations, the ordinary JSD is $D_{\mathrm{JS}} = D_{\mathrm{KL}}{}^{\mathrm{JS}}_{A,A}$, the (A,A) JS-symmetrization of the KLD with respect to the arithmetic means $M = A$ and $N = A$.

The JS-symmetrization can be interpreted as the N_β-Jeffreys' symmetrization of a generalization of Lin's α-skewed K-divergence [4] $D^K_{M_\alpha}[p:q]$:

$$\begin{aligned} D^{\mathrm{JS}}_{M_\alpha,N_\beta}[p:q] &= N_\beta(D^K_{M_\alpha}[p:q], D^K_{M_\alpha}[p:q]), & (30)\\ D^K_{M_\alpha}[p:q] &:= D\left[p:(pq)^{M_\alpha}_\alpha\right]. & (31) \end{aligned}$$

In this work, we consider symmetrizing an arbitrary distance D (including the KLD), generalizing the Jensen-Shannon divergence by using a variational formula for the JSD. Namely, we observe that the Jensen-Shannon divergence can also be defined as the following minimization problem:

$$D_{\mathrm{JS}}[p,q] := \min_{c \in \mathcal{D}} \frac{1}{2}(D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c]), \tag{32}$$

since the optimal density c is proven unique using the calculus of variation [1,31,32] and it corresponds to the mid density $\frac{p+q}{2}$, a statistical (arithmetic) mixture.

Proof. Let $S(c) = D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c] \geq 0$. We use the method of the Lagrange multipliers for the constrained optimization problem $\min_c S(c)$ such that $\int c(x)\mathrm{d}\mu(x) = 1$. Let us minimize $S(c) + \lambda(\int c(x)\mathrm{d}\mu(x) - 1)$. The density c realizing the minimum $S(c)$ satisfies the Euler–Lagrange equation $\frac{\partial L}{\partial c} = 0$, where $L(c) := p\log\frac{p}{c} + q\log\frac{q}{c} + \lambda c$ is the Lagrangian. That is, $-\frac{p}{c} - \frac{q}{c} + \lambda = 0$ or, equivalently, $c = \frac{1}{\lambda}(p+q)$. Parameter λ is then evaluated from the constraint $\int_{\mathcal{X}} c(x)\mathrm{d}\mu(x) = 1$: we get $\lambda = 2$ since $\int_{\mathcal{X}}(p(x)+q(x))\mathrm{d}\mu(x) = 2$. Therefore, we find that $c(x) = \frac{p(x)+q(x)}{2}$, the mid density of $p(x)$ and $q(x)$. □

Considering Equation (32) instead of Equation (5) for defining the Jensen-Shannon divergence is interesting, because it allows one to consider a novel approach for generalizing the Jensen-Shannon divergence. This variational approach was first considered by Sibson [1] to define the α-information radius of a set of weighted distributions while using Rényi α-entropies that are based on Rényi principled α-means [33]. The α-information radius includes the Jensen-Shannon diversity index when $\alpha = 1$. Sibson's work is our point of departure for generalizing the Jensen-Shannon divergence and proposing the Jensen-Shannon symmetrizations of arbitrary distances.

The paper is organized, as follows: in Section 2, we recall the rationale and definitions of the Rényi α-entropy and the Rényi α-divergence [33], and explain the information radius of Sibson [1], which includes, as a special case, the ordinary Jensen-Shannon divergence and that can be interpreted as generalized skew Bhattacharyya distances. We report, in Theorem 2, a closed-form formula for calculating the information radius of order α between two densities of an exponential family when $\frac{1}{\alpha}$ is an integer. It is noteworthy to point out that Sibson's work (1969) includes, as a particular case of the information radius, a definition of the JSD, prior to the well-known reference paper of Lin [4] (1991). In Section 3, we present the JS-symmetrization variational definition that is based on a generalization of the information radius with a generic mean (Equation (88) and Definition 3). In Section 4, we constrain the mixture density to belong to a prescribed class of (parametric) probability densities, like an exponential family [2], and obtain a relative information radius generalizing information radius and related to the concept of information projections. Our Definition 5 generalizes the (relative) normal information radius of Sibson [1], who considered the multivariate normal family (Proposition 4). We illustrate this notion of relative information radius by calculating the density of an exponential family minimizing the reverse Kullback–Leibler divergence between a mixture of densities of that exponential family (Proposition 6). Moreover, we get a semi-closed-form formula for the Kullback–Leibler divergence between the densities of two different exponential families (Proposition 5), generalizing the Fenchel–Young divergence [34]. As an application of these relative variational JSDs, we touch upon the problems of clustering and quantization of probability densities in Section 4.2. Finally, we conclude by summarizing our contributions and discussing related works in Section 5.

2. Rényi Entropy and Divergence, and Sibson Information Radius

Rényi [33] investigated a generalization of the four axioms of Fadeev [35], yielding the unique Shannon entropy [20]. In doing so, Rényi replaced the ordinary weighted arithmetic mean by a more general class of averaging schemes. Namely, Rényi considered the weighted quasi-arithmetic means [36]. A weighted quasi-arithmetic mean can be induced by a strictly monotonous and continuous function g, as follows:

$$M_g(x_1,\ldots,x_n;w_1,\ldots,w_n) := g^{-1}\left(\sum_{i=1}^{n} w_i g(x_i)\right), \tag{33}$$

where the x_i's and the w_i's are positive (the weights are normalized, so that $\sum_{i=1}^n w_i = 1$). Because $M_g = M_{-g}$, we may assume without loss of generality that g is a strictly increasing and continuous function. The quasi-arithmetic means were investigated independently by Kolmogorov [36], Nagumo [37], and de Finetti [38].

For example, the power means $P_\alpha(a,b) = \left(\frac{a^\alpha+b^\alpha}{2}\right)^{\frac{1}{\alpha}}$ introduced earlier are quasi-arithmetic means for the generator $g_\alpha^P(u) := u^\alpha$:

$$P_\alpha(a,b) = M_{g_\alpha^P}\left(a,b;\frac{1}{2},\frac{1}{2}\right). \tag{34}$$

Rényi proved that, among the class of weighted quasi-arithmetic means, only the means induced by the family of functions

$$g_\alpha(u) := 2^{(\alpha-1)u}, \tag{35}$$

$$g_\alpha^{-1}(v) := \frac{1}{\alpha-1}\log_2 v, \tag{36}$$

for $\alpha > 0$ and $\alpha \neq 1$ yield a proper generalization of Shannon entropy, nowadays called the Rényi α-entropy. The Rényi α-mean is

$$\begin{aligned} M_\alpha^R(x_1,\ldots,x_n;w_1,\ldots,w_n) &= M_{g_\alpha}(x_1,\ldots,x_n;w_1,\ldots,w_n), && (37)\\ &= \frac{1}{\alpha-1}\log_2\left(\sum_{i=1}^n w_i 2^{(\alpha-1)x_i}\right). && (38)\end{aligned}$$

The Rényi α-means M_α^R are not power means: They are not homogeneous means [31]. Let $M_\alpha^R(p,q) = M_\alpha^R\left(p,q;\frac{1}{2},\frac{1}{2}\right) = \frac{1}{\alpha-1}\log_2\frac{2^{(\alpha-1)p}+2^{(\alpha-1)q}}{2}$. Subsequently, we have $\lim_{\alpha\to\infty} M_\alpha^R(p,q) = \max\{p,q\}$ and $\lim_{\alpha\to 1} M_\alpha^R(p,q) = A(p,q) = \frac{p+q}{2}$. Indeed, we have

$$\begin{aligned} M_\alpha^R(p,q) &= \frac{1}{\alpha-1}\log_2\frac{2^{(\alpha-1)p}+2^{(\alpha-1)q}}{2},\\ &= \frac{1}{\alpha-1}\log_2\frac{e^{(\alpha-1)p\log 2}+e^{(\alpha-1)q\log 2}}{2},\\ &\approx_{\alpha\to 1} \frac{1}{\alpha-1}\log_2\left(1+(\alpha-1)\frac{p+q}{2}\log 2\right),\\ &\approx_{\alpha\to 1} \frac{1}{\alpha-1}\frac{1}{\log 2}(\alpha-1)\frac{p+q}{2}\log 2,\\ &\approx_{\alpha\to 1} \frac{p+q}{2} = A(p,q),\end{aligned}$$

using the following first-order approximations: $e^x \approx_{x\to 0} = 1+x$ and $\log(1+x) \approx_{x\to 0} = x$.

To obtain an intuition of the Rényi entropy, we may consider generalized entropies derived from quasi-arithmetic means, as follows:

$$h_g[p] := -M_g(\log_2 p_1,\ldots,\log_2 p_n;p_1,\ldots,p_n). \tag{39}$$

When $g(u) = u$, we recover Shannon entropy. When $g_2(u) = 2^u$, we get $h_{g_2}[p] = -\log_2\sum_i p_i^2$, called the collision entropy, since $-\log\Pr[X_1 = X_2] = h_{g_2}[p]$, when X_1 and X_2 are independent and identically distributed random variables with $X_1 \sim p$ and $X_2 \sim p$. When $g(u) = g_\alpha(u) = 2^{(\alpha-1)u}$, we get

$$\begin{aligned} h_{g_\alpha}[p] &= -\frac{1}{\alpha-1}\log_2\left(\sum_i p_i 2^{(\alpha-1)\log_2 p_i}\right), && (40)\\ &= \frac{1}{1-\alpha}\log_2\sum_i p_i p_i^{\alpha-1} = \frac{1}{1-\alpha}\log_2\sum_i p_i^\alpha. && (41)\end{aligned}$$

The formula of Equation (41) is the discrete Rényi α-entropy [33], which can be defined more generally on a measure space $(\mathcal{X},\mathcal{F},\mu)$, as follows:

$$h_\alpha^R[p] := \frac{1}{1-\alpha}\log\left(\int_{\mathcal{X}} p^\alpha(x)\mathrm{d}\mu(x)\right), \quad \alpha \in (0,1)\cup(1,\infty). \tag{42}$$

In the limit case $\alpha \to 1$, the Rényi α-entropy converges to Shannon entropy: $\lim_{\alpha\to 1} h_\alpha^R[p] = h[p]$. Rényi α-entropies are non-increasing with respect to increasing α: $h_\alpha^R[p] \geq h_{\alpha'}^R[p]$ for $\alpha < \alpha'$. In the discrete case (i.e., counting measure μ on a finite alphabet $\mathcal{X}$), we can further define $h_0[p] = \log|\mathcal{X}|$ for $\alpha = 0$ (also called max-entropy or Hartley entropy). The Rényi $+\infty$-entropy

$$h_{+\infty}[p] = -\log\max_{x\in\mathcal{X}} p(x)$$

is also called the min-entropy, since the sequence h_α is non-increasing with respect to increasing α.

Similarly, Rényi obtained the α-divergences for $\alpha > 0$ and $\alpha \neq 1$ (originally called information gain of order α):

$$D_\alpha^R[p:q] := \frac{1}{\alpha-1}\log_2\left(\int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha}\mathrm{d}\mu(x)\right), \tag{43}$$

generalizing the Kullback–Leibler divergence, since $\lim_{\alpha\to 1} D_\alpha^R[p:q] = D_{\mathrm{KL}}[p:q]$. Rényi α-divergences are non-decreasing with respect to increasing α [39]: $D_\alpha^R[p:q] \leq D_{\alpha'}^R[p:q]$ for $\alpha' \geq \alpha$.

Sibson (Robin Sibson (1944–2017) is also renown for inventing the natural neighbour interpolation [40]) [1] considered both the Rényi α-divergence [33] D_α^R and the Rényi α-weighted mean $M_\alpha^R := M_{g_\alpha}$ to define the information radius R_α of order α of a weighted set $\mathcal{P} = \{(w_i, p_i)\}_{i=1}^n$ of densities p_i's as the following minimization problem:

$$R_\alpha(\mathcal{P}) := \min_{c\in\mathcal{D}} R_\alpha(\mathcal{P}, c), \tag{44}$$

where

$$R_\alpha(\mathcal{P}, c) := M_\alpha^R\left(D_\alpha^R[p_1:c], \ldots, D_\alpha^R[p_n:c]; w_1, \ldots, w_n\right). \tag{45}$$

The Rényi α-weighted mean M_α^R can be rewritten as

$$M_\alpha^R(x_1,\ldots,x_n;w_1,\ldots,w_n) = \frac{1}{\alpha-1}\mathrm{LSE}((\alpha-1)x_1\log 2+\log w_1,\ldots,(\alpha-1)x_i\log 2+\log w_i), \tag{46}$$

where function $\mathrm{LSE}(a_1,\ldots,a_n) := \log(\sum_{i=1}^n e^{a_i})$ denotes the log-sum-exp (convex) function [41,42].

Notice that $2^{(\alpha-1)D_\alpha^R[p:q]} = \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha}\mathrm{d}\mu(x)$, the Bhattacharyya α-coefficient [12] (also called Chernoff α-coefficient [43,44]):

$$C_{\mathrm{Bhat},\alpha}[p:q] := \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha}\mathrm{d}\mu(x). \tag{47}$$

Thus, we have

$$R_\alpha(\mathcal{P}, c) = \frac{1}{\alpha-1}\log_2\left(\sum w_i C_{\mathrm{Bhat},\alpha}[p_i:c]\right). \tag{48}$$

The ordinary Bhattacharyya coefficient is obtained for $\alpha = \frac{1}{2}$: $C_{\mathrm{Bhat}}[p:q] := \int_{\mathcal{X}}\sqrt{p(x)}\sqrt{q(x)}\mathrm{d}\mu(x)$.

Sibson [1] also considered the limit case $\alpha \to \infty$ when defining the information radius:

$$D_\infty^R[p:q] := \log_2 \sup_{x\in\mathcal{X}} \frac{p(x)}{q(x)}. \tag{49}$$

Sibson reported the following theorem in his information radius study [1]:

Theorem 1 (Theorem 2.2 and Corollary 2.3 of [1])**.** *The optimal density* $c_\alpha^* = \arg\min_{c\in\mathcal{D}} R_\alpha(\mathcal{P}, c)$ *is unique, and we have:*

$$\begin{aligned}
c_1^*(x) &= \textstyle\sum_i w_i p_i(x), & R_1(\mathcal{P}) &= R_1(\mathcal{P}, c_1^*) = \textstyle\int_{\mathcal{X}}\sum_i w_i p_i \log_2 \frac{p_i}{\sum_j w_j p_j(x)}\mathrm{d}\mu(x),\\
c_\alpha^*(x) &= \frac{(\sum_i w_i p_i(x)^\alpha)^{\frac{1}{\alpha}}}{\int_{\mathcal{X}}(\sum_i w_i p_i(x)^\alpha)^{\frac{1}{\alpha}}\mathrm{d}\mu(x)}, & R_\alpha(\mathcal{P}) &= R_\alpha(\mathcal{P}, c_\alpha^*) = \tfrac{1}{\alpha-1}\log_2\left(\textstyle\int_{\mathcal{X}}(\sum_i w_i p_i(x)^\alpha)^{\frac{1}{\alpha}}\mathrm{d}\mu(x)\right)^\alpha,\\
& & & \alpha \in (0,1)\cup(1,\infty)\\
c_\infty^*(x) &= \frac{\max_i p_i(x)}{\int_{\mathcal{X}}(\max_i p_i(x))\mathrm{d}\mu(x)}, & R_\infty(\mathcal{P}) &= R_\infty(\mathcal{P}, c_\infty^*) = \log_2 \textstyle\int_{\mathcal{X}}(\max_i p_i(x))\mathrm{d}\mu(x),
\end{aligned}$$

Observe that $R_\infty(\mathcal{P})$ does not depend on the (positive) weights.

The proof follows from the following decomposition of the information radius:

Proposition 1. *We have:*

$$R_\alpha(\mathcal{P}, c) - R_\alpha(\mathcal{P}, c^*_\alpha) = D^R_\alpha(c^*_\alpha, c) \geq 0. \tag{50}$$

Because the proof is omitted in [1], we report it here:

Proof. Let $\Delta(c, c^*_\alpha) := R_\alpha(\mathcal{P}, c) - R_\alpha(\mathcal{P}, c^*_\alpha)$. We handle the three cases, depending on the α values:

- Case $\alpha \in (0,1) \cup (1, \infty)$: Let $P_\alpha(\mathcal{P})(x) := (\sum_i w_i p_i(x)^\alpha)^{\frac{1}{\alpha}}$. We have $(c^*_\alpha(x))^\alpha = \frac{\sum_i w_i p_i(x)^\alpha}{(\int P_\alpha(\mathcal{P})(x) \mathrm{d}\mu(x))^\alpha}$. We obtain

$$\begin{aligned} \Delta(c, c^*_\alpha) &= \frac{1}{\alpha - 1} \log_2 \left(\sum_i w_i \int p_i(x)^\alpha c(x)^{1-\alpha} \mathrm{d}\mu(x) \right) - \frac{1}{\alpha - 1} \log_2 \left(\int P_\alpha(\mathcal{P})(x) \mathrm{d}\mu(x) \right)^\alpha, && (51) \\ &= \frac{1}{\alpha - 1} \log_2 \frac{\sum_i w_i \int p_i(x)^\alpha c(x)^{1-\alpha} \mathrm{d}\mu}{(\int P_\alpha(\mathcal{P})(x) \mathrm{d}\mu(x))^\alpha}, && (52) \\ &= \frac{1}{\alpha - 1} \log_2 \int \frac{(\sum_i w_i p_i(x)^\alpha) c(x)^{1-\alpha}}{(\int P_\alpha(\mathcal{P})(x) \mathrm{d}\mu(x))^\alpha} \mathrm{d}\mu(x), && (53) \\ &= \frac{1}{\alpha - 1} \log_2 \int (c^*_\alpha(x))^\alpha c(x)^{1-\alpha} \mathrm{d}\mu(x), && (54) \\ &:= D^R_\alpha(c^*_\alpha, c). && (55) \end{aligned}$$

- Case $\alpha = 1$: we have $\Delta(c, c^*_1) := R_1(\mathcal{P}, c) - R_1(\mathcal{P}, c^*_1)$ with $c^*_1 = \sum_i w_i p_i$. Because $R_1(\mathcal{P}, c) = \sum_i w_i D_{\mathrm{KL}}[p_i : c]$, we have

$$\begin{aligned} R_1(\mathcal{P}, c) &= \sum_i w_i h[p_i : c] - w_i h[p_i], && (56) \\ &= h[\sum_i w_i p_i : c] - \sum_i w_i h[p_i], && (57) \\ &= h[c^*_1 : c] - \sum_i w_i h[p_i]. && (58) \end{aligned}$$

It follows that

$$\begin{aligned} \Delta(c, c^*_1) &= h[c^*_1 : c] - \sum_i w_i h[p_i] - \left(h[c^*_1 : c^*_1] - \sum_i w_i h[p_i] \right), && (59) \\ &= h^{[} c^*_1 : c] - h[c^*_1], && (60) \\ &= D_{\mathrm{KL}}[c^*_1 : c] = D^R_1[c^*_1 : c]. && (61) \end{aligned}$$

- Case $\alpha = \infty$: we have $c^*_\infty = \frac{\max_i p_i(x)}{\int (\max_i p_i(x)) \mathrm{d}\mu(x)}$, $R_\infty(\mathcal{P}, c^*_\infty) = \log_2 \int (\max_i p_i(x)) \mathrm{d}\mu(x)$, and $D^R_\infty[p : q] = \log_2 \sup_x \frac{p(x)}{q(x)}$. We have $R_\infty(\mathcal{P}, c) = \log_2 \sup_x \frac{p_i(x)}{c(x)}$ Thus, $\Delta(c, c^*_\alpha) := R_\infty(\mathcal{P}, c) - R_\infty(\mathcal{P}, c^*_\infty) = \log_2 \sup_x \frac{c^*_\infty(x)}{c(x)} = D^R_\infty[c^*_\infty : c]$.

□

It follows that

$$\min_c R_\alpha(\mathcal{P}, c) = \min_c R_\alpha(\mathcal{P}, c^*_\alpha) + D^R_\alpha(c^*_\alpha, c) \equiv \min_c D^R_\alpha(c^*_\alpha, c) \geq 0.$$

Thus we have $c = c^*_\alpha$ since $D^R_\alpha(c^*_\alpha, c)$ is minimized for $c = c^*_\alpha$.

Notice that $c^*_\infty(x) = \frac{\max\{p_1(x), \ldots, p_n(x)\}}{\int_{\mathcal{X}} (\max_i p_i(x)) \mathrm{d}\mu(x)}$ is the upper envelope of the densities $p_i(x)$'s normalized to be a density. Provided that the densities p_i's intersect pairwise in at most s locations (i.e., $|\{p_i(x) \cap p_j(x)\}| \leq s$ for $i \neq j$), we can efficiently compute this upper envelope using an output-sensitive algorithm [45] of computational geometry.

When the point set is $\mathcal{P} = \left\{ \left(\frac{1}{2}, p\right), \left(\frac{1}{2}, q\right) \right\}$ with $w_1 = w_2 = \frac{1}{2}$, the information radius defines a (2-point) symmetric distance, as follows:

$$\begin{array}{ll} R_1(p,q) = \frac{1}{2}\int_{\mathcal{X}} p(x)\log_2 \frac{2p}{p(x)+q(x)}\mathrm{d}\mu(x) + \frac{1}{2}\int_{\mathcal{X}} q(x)\log_2 \frac{2q(x)}{p(x)+q(x)}\mathrm{d}\mu(x), & \alpha = 1 \\ R_\alpha(p,q) = \frac{\alpha}{\alpha-1}\log_2 \int_{\mathcal{X}} \left(\frac{p(x)^\alpha + q(x)^\alpha}{2}\right)^{\frac{1}{\alpha}}\mathrm{d}\mu(x) = \frac{\alpha}{\alpha-1}\log_2 \int_{\mathcal{X}} P_\alpha(p(x),q(x))\mathrm{d}\mu(x), & \alpha \in (0,1)\cup(1,\infty) \\ R_\infty(p,q) = \log_2 \int_{\mathcal{X}} \max\{p(x),q(x)\}\mathrm{d}\mu(x), & \alpha = \infty. \end{array}$$

This family of symmetric divergences may be called the Sibson's α-divergences, and the Jensen-Shannon divergence is interpreted as a limit case when $\alpha \to 1$. Notice that, since we have $\lim_{\alpha\to\infty} P_\alpha(p,q) = \max\{p,q\}$ and $\lim_{\alpha\to\infty}\frac{\alpha}{\alpha-1} = 1$, we have $\lim_{\alpha\to\infty} R_\alpha(p,q) = R_\infty(p,q)$. Notice that, for $\alpha = 1$, the integral and logarithm operations are swapped as compared to R_α for $\alpha \in (0,1)\cup(1,\infty)$.

Theorem 2. *When $\alpha = \frac{1}{k}$ for an integer $k \geq 2$, the Sibson α-divergences between two densities p_{θ_1} and p_{θ_2} of an exponential family $\{p_\theta \,:\, \theta \in \Theta\}$ with cumulant function $F(\theta)$ is available in closed form:*

$$R_\alpha(p_{\theta_1}, p_{\theta_2}) = -\frac{1}{k-1}\log_2\left(\frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}\exp\left(F\left(\frac{i}{k}\theta_1 + \left(1-\frac{i}{k}\right)\theta_2\right) - \left(\frac{i}{k}F(\theta_1) + \left(1-\frac{i}{k}\right)F(\theta_2)\right)\right)\right).$$

Proof. Let $p = p_{\theta_1}$ and $q = p_{\theta_2}$ be two densities of an exponential family [2] with cumulant function $F(\theta)$ and natural parameter space Θ. Without a loss of generality, we may consider a natural exponential family [2] with densities written canonically as $p_\theta(x) = \exp(x^\top\theta - F(\theta))$ for $\theta \in \Theta$. It can be shown that the cumulant function $F(\theta) = \log\int_{\mathcal{X}}\exp(x^\top\theta)\mathrm{d}\mu(x)$ is strictly convex and analytic on the open convex natural parameter space Θ [2].

When $\alpha = \frac{1}{2}$ (i.e., $k = 2$), we have:

$$R_{\frac{1}{2}}(p,q) = -\log_2 \int_{\mathcal{X}}\left(\frac{\sqrt{p(x)}+\sqrt{q(x)}}{2}\right)^2\mathrm{d}\mu(x), \tag{62}$$

$$= -\log_2\left(\frac{1}{2} + \frac{1}{2}\int_{\mathcal{X}}\sqrt{p(x)}\sqrt{q(x)}\mathrm{d}\mu(x)\right), \tag{63}$$

$$= -\log_2\left(\frac{1}{2} + \frac{1}{2}C_{\mathrm{Bhat}}[p:q]\right) \geq 0, \tag{64}$$

where $C_{\mathrm{Bhat}}[p:q] := \int_{\mathcal{X}}\sqrt{p(x)}\sqrt{q(x)}\mathrm{d}\mu(x)$ is the Bhattacharyya coefficient (with $0 \leq C_{\mathrm{Bhat}}[p:q] \leq 1$). Using Theorem 3 of [12], we have

$$C_{\mathrm{Bhat}}[p_{\theta_1}, p_{\theta_2}] = \exp\left(F\left(\frac{\theta_p + \theta_q}{2}\right) - \frac{F(\theta_p) + F(\theta_q)}{2}\right),$$

so that we obtain the following closed-form formula:

$$R_{\frac{1}{2}}(p_{\theta_1}, p_{\theta_2}) = -\log_2\left(\frac{1}{2} + \frac{1}{2}\exp\left(F\left(\frac{\theta_p+\theta_q}{2}\right) - \frac{F(\theta_p)+F(\theta_q)}{2}\right)\right) \geq 0,$$

Now, assume that $k = \frac{1}{\alpha} \geq 2$ is an arbitrary integer, and let us apply the binomial expansion for $P_\alpha(p_{\theta_1}, p_{\theta_2})$ in the spirit of [46,47]:

$$\int_{\mathcal{X}} P_\alpha(p_{\theta_1}(x), p_{\theta_2}(x))\mathrm{d}\mu(x) = \int_{\mathcal{X}}\left(\frac{p_{\theta_1}(x)^{\frac{1}{k}} + p_{\theta_2}(x)^{\frac{1}{k}}}{2}\right)^k\mathrm{d}\mu(x), \tag{65}$$

$$= \frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}\int_{\mathcal{X}}\left(p_{\theta_1}(x)^{\frac{1}{k}}\right)^i\left(p_{\theta_2}(x)^{\frac{1}{k}}\right)^{k-i}\mathrm{d}\mu(x). \tag{66}$$

Let $I_{k,i}(\theta_1,\theta_2) := \int_{\mathcal{X}} \left(p_{\theta_1}(x)^{\frac{1}{k}}\right)^i \left(p_{\theta_2}(x)^{\frac{1}{k}}\right)^{k-i} \mathrm{d}\mu(x)$. Because $\frac{i}{k}\theta_1 + \frac{k-i}{k}\theta_2 = \theta_2 + \frac{i}{k}(\theta_1-\theta_2) \in \Theta$ for $i \in \{0,\ldots,k\}$, we get by following the calculation steps in [12]:

$$I_{k,i}(\theta_1,\theta_2) := \exp\left(F\left(\frac{i}{k}\theta_1 + \left(1-\frac{i}{k}\right)\theta_2\right) - \left(\frac{i}{k}F(\theta_1) + \left(1-\frac{i}{k}\right)F(\theta_2)\right)\right) < \infty.$$

Notice that $I_{2,1} = C_{\mathrm{Bhat}}[p_{\theta_1}, p_{\theta_2}]$, and $I_{k,0} = I_{k,k} = 1$.

Thus, we get the following closed-form formula:

$$R_\alpha(p_{\theta_1}, p_{\theta_2}) = -\frac{1}{k-1}\log_2\left(\frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}\exp\left(F\left(\frac{i}{k}\theta_1 + \left(1-\frac{i}{k}\right)\theta_2\right) - \left(\frac{i}{k}F(\theta_1) + \left(1-\frac{i}{k}\right)F(\theta_2)\right)\right)\right). \quad (67)$$

□

This closed-form formula applies, in particular, to the family $\{\mathcal{N}(\mu,\Sigma)\}$ of (multivariate) normal distributions: In this case, the natural parameters θ are expressed using both a vector parameter component v and a matrix parameter component M:

$$\theta = (v, M) = \left(\Sigma^{-1}m, -\frac{1}{2}\Sigma^{-1}\right), \quad (68)$$

and the cumulant function is:

$$F_{\mathcal{N}}(\theta) = \frac{d}{2}\log\pi - \frac{1}{2}\log|-2M| - \frac{1}{4}v^\top M^{-1}v, \quad (69)$$

where $|\cdot|$ denotes the matrix determinant.

In general, the optimal density $c_\alpha^* = \arg\min_{c\in\mathcal{D}} R_\alpha(\mathcal{P}, c)$ yielding the information radius $R_\alpha(\mathcal{P})$ can be interpreted as a generalized centroid (extending the notion of Fréchet means [48]) with respect to (M_α^R, D_α^R), where a (M,D)-centroid is defined by:

Definition 1 ((M,D)-centroid)**.** *Let $\mathcal{P} = \{(w_1,p_1),\ldots,(w_n,p_n)\}$ be a normalized weighted parameter set, M a mean, and D a distance. Subsequently, the (M,D)-centroid is defined as*

$$c_{M,D}(\mathcal{P}) = \arg\min_c M(D(p_1 : c),\ldots,D(p_n : c); w_1,\ldots,w_n).$$

Here, we give a general definition of the (M,D)-centroid for an arbitrary distance (not necessarily a symmetric nor metric distance). The parameter set can either be probability measures having densities with respect to a given measure μ or a set of vectors. In the first case, the distance D is called a statistical distance. When the densities belong to a parametric family of densities $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$, the statistical distance $D[p_{\theta_1} : p_{\theta_2}]$ amounts to a parameter distance: $D_{\mathcal{P}}(\theta_1 : \theta_2) := D[p_{\theta_1} : p_{\theta_2}]$. For example, when all of the densities p_i's belong to a same natural exponential family [2]

$$\mathcal{P} = \{p_\theta(x) = \exp(\theta^\top t(x) - F(\theta)) \ : \ \theta \in \Theta\}$$

with cumulant function $F(\theta) = \log\int\exp(\theta^\top t(x))\mathrm{d}\mu(x)$ (i.e., $p_i = p_{\theta_i}$) and sufficient statistic vector $t(x)$, we have $D_{\mathrm{KL}}[p_\theta : p_{\theta_i}] = B_F^*(\theta : \theta_i) := B_F(\theta_i : \theta)$, where B_F^* denotes the reverse Bregman divergence (by parameter order swapping) the Bregman divergence [21] B_F defined by

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta'). \quad (70)$$

Thus, we have $D_{\mathcal{P}}(\theta_1 : \theta_2) := B_F^*(\theta_1 : \theta_2) = D_{\mathrm{KL}}[p_{\theta_1} : p_{\theta_2}]$.

Let $\mathcal{V} = \{(w_1,\theta_1),\ldots,(w_n,\theta_n)\}$ be the parameter set corresponding to $\mathcal{P}$. Define

$$R_F(\mathcal{V},\theta) := \sum_{i=1}^{n} w_i B_F(\theta_i : \theta). \quad (71)$$

Subsequently, we have the equivalent decomposition of Proposition 1:

$$R_F(\mathcal{V},\theta) - R_F(\mathcal{V},\theta^*) = B_F(\theta^* : \theta), \tag{72}$$

with $\theta^* = \bar{\theta} := \sum_{i=1}^n w_i\theta_i$. (this decomposition is used to prove Proposition 1 of [21]). The quantity $R_F(\mathcal{V}) = R_F(\mathcal{V},\theta^*)$ was termed the Bregman information [21,49]. The Bregman information generalizes the variance that was obtained when the Bregman divergence is the squared Euclidean distance. $R_F(\mathcal{V})$ could also be called Bregman information radius according to Sibson. Because $R_F(\mathcal{V}) = \sum_{i=1}^n w_i D_{\mathrm{KL}}[p_{\bar{\theta}} : p_{\theta_i}]$, we can interpret the Bregman information as a Sibson's information radius for densities of an exponential family with respect to the arithmetic mean $M_1^R = A$ and the reverse Kullback–Leibler divergence: $D_{\mathrm{KL}}^*[p : q] := D_{\mathrm{KL}}[q : p]$. This observation yields us the JS-symmetrization of distances based on generalized information radii in Section 3.

More generally, we may consider the densities belonging to a deformed q-exponential family (see [10], page 85–89 and the monograph [50]). Deformed q-exponential families generalize the exponential families, and include the q-Gaussians [10]. A common way to measure the statistical distance between two densities of a q-exponential family is the q-divergence [10], which is related to Tsallis' entropy [51]. We may also define another statistical divergence between two densities of a q-exponential family which amounts to Bregman divergence. For example, we refer to [52] for details concerning the family of Cauchy distributions, which are q-Gaussians for $q = 2$.

Sibson proved that the information radii of any order are all upper bounded (Theorem 2.8 and Theorem 2.9 of [1]) as follows:

$$\begin{aligned} R_1(\mathcal{P}) &\leq \sum_i w_i \log_2 \frac{1}{w_j} \leq \log_2 n < \infty, && (73)\\ R_\alpha(\mathcal{P}) &\leq \frac{\alpha}{\alpha-1}\log_2\left(\sum_i w_i^{\frac{1}{\alpha}}\right) \leq \log_2 n < \infty, \quad \alpha \in (0,1)\cup(1,\infty) && (74)\\ R_\infty(\mathcal{P}) &\leq \log_2 n < \infty. && (75) \end{aligned}$$

We interpret Sibson's upper bounds of Equations (73)–(75), as follows:

Proposition 2 (Information radius upper bound)**.** *The information radius of order α of a weighted set of distributions is upper bounded by the discrete Rényi entropy of order $\frac{1}{\alpha}$ of the weight distribution: $R_\alpha(\mathcal{P}) \leq H_{\frac{1}{\alpha}}^R[w]$, where $H_\alpha^R[w] := \frac{1}{1-\alpha}\log(\sum_i w_i^\alpha)$.*

3. JS-Symmetrization of Distances Based on Generalized Information Radius

Let us give the following definitions generalizing the information radius (i.e., Jensen-Shannon symmetrization of the distance when $|\mathcal{P}| = 2$) and the ordinary Jensen-Shannon divergence:

Definition 2 ((M,D)-information radius)**.** *Let M be a weighted mean and D a distance. Subsequently, the generalized information radius for a weighted set of points (e.g., vectors or densities) $(w_1,p_1),\ldots,(w_n,p_n)$ is:*

$$R_{M,D}(\mathcal{P}) := \min_{c\in\mathcal{D}} M(D(p_1 : c),\ldots,D(p_n : c); w_1,\ldots,w_n).$$

Recall that we also defined the (M,D)-centroid in Definition 1 as follows:

$$c_{M,D}(\mathcal{P}) := \arg\min_{c\in\mathcal{D}} M(D(p_1 : c),\ldots,D(p_n : c); w_1,\ldots,w_n).$$

When $M = A$, we recover the notion of Fréchet mean [48]. Notice that, although the minimum $R_{M,D}(\mathcal{P})$ is unique, several generalized centroids $c_{M,D}(\mathcal{P})$ may potentially exist, depending on (M,D). In particular, Definition 2 and Definition 1 apply when D is a statistical distance, i.e., a distance between densities (Radon–Nikodym derivatives of corresponding probability measures with respect to a dominating measure μ).

The generalized information radius can be interpreted as a diversity index or an n-point distance. When $n = 2$, we get the following (2-point) distances, which are considered as a generalization of the Jensen-Shannon divergence or Jensen-Shannon symmetrization:

Definition 3 (M-vJS symmetrization of D)**.** *Let M be a mean and D a statistical distance. Subsequently, the variational Jensen-Shannon symmetrization of D is defined by the formula of a generalized information radius:*

$$D_M^{\mathrm{vJS}}[p:q] := \min_{c\in\mathcal{D}} M(D[p:c], D[q:c]).$$

We use the acronym vJS to distinguish it with the JS-symmetrization reported in [23]:

$$D_M^{\mathrm{JS}}[p:q] = D_{M,A}^{\mathrm{JS}}[p:q] := \frac{1}{2}\left(D\left[p:(pq)_{\frac{1}{2}}^M\right] + D\left[q:(pq)_{\frac{1}{2}}^M\right]\right).$$

We recover Sibson's information radius $R_\alpha[p:q]$ induced by two densities p and q from Definition 3 as the M_α^R*-vJS symmetrization of the Rényi divergence* D_α^R. We have ${B_F}_A^{\mathrm{vJS}}$, which is the Bregman information [21]. Notice that we may skew these generalized JSDs by taking weighted mean M_β instead of M for $\beta \in (0,1)$, yielding the general definition:

Definition 4 (Skew M_β-vJS symmetrization of D)**.** *Let M_β be a weighted mean and D a statistical distance. Subsequently, the variational skewed Jensen-Shannon symmetrization of D is defined by the formula of a generalized information radius:*

$$\boxed{D_{M_\beta}^{\mathrm{vJS}}[p:q] := \min_{c\in\mathcal{D}} M_\beta(D[p:c], D[q:c])}$$

Example 1. *For example, the skewed Jensen–Bregman divergence of Equation (20) can be interpreted as a Jensen-Shannon symmetrization of the Bregman divergence B_F [12] since we have:*

$$\begin{aligned} {B_F}_{A_\beta}^{\mathrm{vJS}}(\theta_1:\theta_2) &= \min_{\theta\in\Theta} A_\beta(B_F(\theta_1:\theta), B_F(\theta_2:\theta)), && (76)\\ &= \min_{\theta\in\Theta}(1-\beta)B_F(\theta_1:\theta) + \beta B_F(\theta_2:\theta), && (77)\\ &= (1-\beta)B_F(\theta_1:(1-\beta)\theta_1+\beta\theta_2) + \beta B_F(\theta_2:(1-\beta)\theta_1+\beta\theta_2), && (78)\\ &=: \mathrm{JB}_{F,\beta}(\theta_1:\theta_2). && (79)\end{aligned}$$

Indeed, the Bregman barycenter $\arg\min_{\theta\in\Theta}(1-\beta)B_F(\theta_1:\theta) + B_F(\theta_2:\theta)$ *is unique and it corresponds to $\theta = (1-\beta)\theta_1 + \beta\theta_2$, see [21]. The skewed Jensen–Bregman divergence* $\mathrm{JB}_{F,\beta}(\theta_1:\theta_2)$ *can also be rewritten as an equivalent skewed Jensen divergence (see Equation (22)):*

$$\begin{aligned} \mathrm{JB}_{F,\beta}(\theta_1:\theta_2) &= (1-\beta)B_F(\theta_1:(1-\beta)\theta_1+\beta\theta_2) + \beta B_F(\theta_2:(1-\beta)\theta_1+\beta\theta_2), && (80)\\ &= (1-\beta)F(\theta_1) + \beta F(\theta_2) - F((1-\beta)\theta_1+\beta\theta_2), && (81)\\ &=: J_{F,\beta}(\theta_1:\theta_2). && (82)\end{aligned}$$

Example 2. *Consider a conformal Bregman divergence [53] that is defined by*

$$B_{F,\rho}(\theta_1:\theta_2) = \rho(\theta_1)B_F(\theta_1:\theta_2), \quad (83)$$

where $\rho(\theta) > 0$ is a conformal factor. Subsequently, we have

$$
\begin{aligned}
B_{F,\rho}{}^{\mathrm{vJS}}_{A_\beta}(\theta_1 : \theta_2) &= \min_{\theta\in\Theta} A_\beta\big(B_{F,\rho}(\theta_1 : \theta), B_{F,\rho}(\theta_2 : \theta)\big), & (84)\\
&= \min_{\theta\in\Theta}(1-\beta)B_{F,\rho}(\theta_1 : \theta) + B_{F,\rho}(\theta_2 : \theta), & (85)\\
&= (1-\beta)B_F(\theta_1 : \gamma_1\theta_1 + \gamma_2\theta_2) + \beta B_F(\theta_2 : \gamma_1\theta_1 + \gamma_2\theta_2), & (86)
\end{aligned}
$$

where $\gamma_1 = \frac{(1-\beta)\rho(\theta_1)}{(1-\beta)\rho(\theta_1)+\beta\rho(\theta_2)}$ *and* $\gamma_2 = \frac{\beta\rho(\theta_2)}{(1-\beta)\rho(\theta_1)+\beta\rho(\theta_2)} = 1-\gamma_1$.

Notice that this definition is implicit and it can be made explicit when the centroid $c^*(p,q)$ is unique:

$$
D^{\mathrm{vJS}}_{M_\beta}[p:q] = M_\beta(D[p : c^*(p,q)], D[q : c^*(p,q)] \tag{87}
$$

In particular, when $D = D_{\mathrm{KL}}$, the KLD, we obtain generalized skewed Jensen-Shannon divergences for M_β a weighted mean with $\beta \in (0,1)$:

$$
D^{M_\beta}_{\mathrm{vJS}}[p:q] := \min_{c\in\mathcal{D}} M_\beta(D_{\mathrm{KL}}[p:c], D_{\mathrm{KL}}[q:c]). \tag{88}
$$

Example 3. *Amari [31] obtained the (A, D_α)-information radius and its corresponding unique centroid for D_α, the α-divergence of information geometry [10] (page 67).*

Example 4. *Brekelmans et al. [54] studied the geometric path $(p_1p_2)^G_\beta(x) \propto p_1^{1-\beta}(x)p_2^\beta(x)$ between two distributions p_1 and p_2 of $\mathcal{D}$, where $G_\beta(a,b) = a^{1-\beta}b^\beta$ (with $a, b > 0$) is the weighted geometric mean. They proved the variational formula:*

$$
(p_1p_2)^G_\beta = \min_{c\in\mathcal{D}}(1-\beta)D_{\mathrm{KL}}[c : p_1] + \beta D_{\mathrm{KL}}[c : p_2]. \tag{89}
$$

That is, $(p_1p_2)^G_\beta$ is a G_β-D^*_{KL} centroid, where D^*_{KL} is the reverse KLD. The corresponding $(G_\beta, D^*_{\mathrm{KL}})$-vJSD is studied is [23] and it is used in deep learning in [30].

It is interesting to study the link between (M_β, D)-variational Jensen-Shannon symmetrization of D and the (M'_α, N'_β)-JS symmetrization of [23]. In particular, the link between M_β for averaging in the minimization and M'_α the mean for generating abstract mixtures.

More generally, Brekelmans et al. [55] considered the α-divergences extended to positive measures (i.e., a separable divergence built as the different between a weighted arithmetic mean and a geometric mean [56]):

$$
D^e_\alpha[p:q] := \frac{4}{1-\alpha^2}\int_{\mathcal{X}}\left(\frac{1-\alpha}{2}p(x) + \frac{1+\alpha}{2}q(x) - p^{\frac{1-\alpha}{2}}(x)q^{\frac{1+\alpha}{2}}(x)\right)\mathrm{d}\mu(x) \tag{90}
$$

and proved that

$$
c^*_\beta = \arg\min_{c\in\mathcal{D}}\{(1-\beta)D^e_\alpha[p_1 : c] + \beta D^e_\alpha[p_2 : c]\} \tag{91}
$$

is a density of a likelihood ratio q-exponential family: $c^*_\beta = \frac{p_1(x)}{Z_{\beta,q}}\exp_q(\beta\log_q\frac{p_2(x)}{p_1(x)})$ for $q = \frac{1+\alpha}{2}$. That is, c^*_β is the (A_β, D^e_α)-generalized centroid, and the corresponding information radius is the variational JS symmetrization:

$$
D^{e\,\mathrm{vJS}}_\alpha[p_1 : p_2] = (1-\beta)D^e_\alpha[p_1 : c^*_\beta] + \beta D^e_\alpha[p_2 : c^*_\beta] \tag{92}
$$

Example 5. *The q-divergence [57] D_q between two densities of a q-exponential family amounts to a Bregman divergence [10,57]. Thus, D^{vJS}_q for $M = A$ is a generalized information radius that amounts to a Bregman information.*

For the case $\alpha = \infty$ in Sibson's information radius, we find that the information radius is related to the total variation:

Proposition 3 (Lemma 2.4 [1])**.** *:*

$$D_{\infty}^{\mathrm{vJS},R}[p:q] = \log_2(1 + D_{\mathrm{TV}}[p:q]), \tag{93}$$

where D_{TV} *denotes the total variation*

$$D_{\mathrm{TV}}[p:q] = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)| \mathrm{d}\mu(x). \tag{94}$$

Proof. Because $\max\{p(x), q(x)\} = \frac{p(x)+q(x)}{2} + \frac{1}{2}|q(x) - p(x)|$, it follows that we have:

$$\int_{\mathcal{X}} \max\{p(x), q(x)\} \mathrm{d}\mu(x) = 1 + D_{\mathrm{TV}}[p:q].$$

From Theorem 1, we have $R_{\infty}(\{(\frac{1}{2}, p), (\frac{1}{2}, q)) = \log_2 \int_{\mathcal{X}} \max\{p(x), q(x)\} \mathrm{d}\mu(x)$ and, therefore, $R_{\infty}(\{(\frac{1}{2}, p), (\frac{1}{2}, q)) = \log_2(1 + D_{\mathrm{TV}}[p:q])$. □

Notice that, when $M = M_g$ is a quasi-arithmetic mean, we may consider the divergence $D_g[p:q] = g^{-1}(D[p:q])$, so that the centroid of the (M_g, D_g)-JS symmetrization is:

$$\arg\min_c g^{-1}\left(\sum_{i=1}^n w_i D[p_i : c]\right) \equiv \arg\min_c \sum_{i=1}^n w_i D[p_i : c]. \tag{95}$$

The generalized α-skewed Bhattacharyya divergence [29] can also be considered with respect to a weighted mean M_α:

$$D_{\mathrm{Bhat},M_\alpha}[p:q] = -\log \int_{\mathcal{X}} M_\alpha(p(x), q(x)) \mathrm{d}\mu(x).$$

In particular, when M_α is a quasi-arithmetic weighted mean that is induced by a strictly continuous and monotone function g, we have

$$D_{\mathrm{Bhat},g,\alpha}[p:q] := -\log \int_{\mathcal{X}} M_g(p(x), q(x); \alpha) \mathrm{d}\mu(x) =: D_{\mathrm{Bhat},(M_g)_\alpha}[p:q].$$

Because $\min\{p(x), q(x)\} \leq M_g(p(x), q(x); \alpha) \leq \max\{p(x), q(x)\}$, $\min\{a, b\} = \frac{a+b}{2} - \frac{|b-a|}{2}$ and $\max\{a, b\} = \frac{a+b}{2} + \frac{|b-a|}{2}$, we deduce that we have:

$$0 \leq 1 - D_{\mathrm{TV}}[p, q] \leq \int_{\mathcal{X}} M_g(p(x), q(x); \alpha) \mathrm{d}\mu(x) \leq 1 + D_{\mathrm{TV}}[p, q] \leq 2. \tag{96}$$

The information radius of Sibson for $\alpha \in (0, 1) \cup (1, \infty)$ may be interpreted as generalized scaled α-skewed Bhattacharyya divergences with respect to the power means P_α, since we have $R_\alpha(p, q) = \frac{\alpha}{\alpha-1} \log_2 \int_{\mathcal{X}} P_\alpha(p(x), q(x); \alpha) \mathrm{d}\mu(x) = \frac{\alpha}{1-\alpha} D_{\mathrm{Bhat},P_\alpha}[p:q]$.

4. Relative Information Radius and Relative Jensen-Shannon Symmetrizations of Distances

4.1. Relative Information Radius

In this section, instead of considering the full space of densities $\mathcal{D}$ on $(\mathcal{X}, \mathcal{F}, \mu)$ for performing the variational optimization of the information radius, we rather consider a subfamily of (parametric) densities $\mathcal{R} \subset \mathcal{D}$. Subsequently, we define accordingly the $\mathcal{R}$-relative Jensen-Shannon divergence ($\mathcal{R}$-JSD for short) as

$$D_{\mathrm{vJS}}^{\mathcal{R}}[p:q] := \min_{c \in \mathcal{R}} \left\{ \frac{1}{2} D_{\mathrm{KL}}[p:c] + \frac{1}{2} D_{\mathrm{KL}}[q:c] \right\}. \tag{97}$$

In particular, Sibson [1] considered the normal information radius, i.e., the $\mathcal{R}$-relative Jensen-Shannon divergence with $\mathcal{R} = \{\mathcal{N}(\mu, \Sigma) \ : \ (\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{P}^d_{++}\}$, where $\mathbb{P}^d_{++}$ denotes the open cone of $d \times d$ positive-definite matrices (positive-definite covariance matrices of Gaussian distributions). More generally, we may consider any exponential family $\mathcal{E}$ [2].

Definition 5 (Relative $(\mathcal{R}, M)$-JS symmetrization of D)**.** *Let M be a mean and D a statistical distance. Subsequently, the relative $(\mathcal{R}, M)$-JS symmetrization of D is:*

$$D^{\mathrm{vJS}}_{M,\mathcal{R}}[p : q] := \min_{c \in \mathcal{R}} M(D[p : c], D[q : c]).$$

We obtain the relative Jensen-Shannon divergences when $D = D_{\mathrm{KL}}$.

Example 6. *Grosse et al. [58] considered geometric and moment average paths for annealing. They proved that, when $p_1 = p_{\theta_1}$ and $p_2 = p_{\theta_2}$ belong to an exponential family [2] $\mathcal{E}_F$ with cumulant function F, we have*

$$(p_1 p_2)^G_\beta = \frac{p_1(x)^{1-\beta} p_2(x)^\beta}{\int p_1(x)^{1-\beta} p_2(x)^\beta \mathrm{d}\mu(x)} = \arg\min_{c \in \mathcal{E}_F} \{(1-\beta) D_{\mathrm{KL}}[c : p_1] + \beta D_{\mathrm{KL}}[c : p_2]\}, \quad (98)$$

and

$$p_{\bar{\eta}} = \arg\min_{c \in \mathcal{E}_F} \{(1-\beta) D_{\mathrm{KL}}[p_1 : c] + \beta D_{\mathrm{KL}}[c : p_2]\}, \quad (99)$$

where $\bar{\eta} = (1-\beta)\eta_1 + \beta\eta_2$, $\eta_i = E_{p_{\theta_i}}[t(x)]$ (this is not an arithmetic mixture, but an exponential family density moment parameter that is a mixture of the parameters).

The corresponding minima can be interpreted as relative skewed Jensen-Shannon symmetrization for the reverse KLD D^*_{KL} (Equation (98)) and the relative skewed Jensen-Shannon divergence (Equation (99)):

$$\begin{aligned} D^{*\ \mathrm{vJS}}_{\mathrm{KL}\,A_\beta,\mathcal{E}_F}[p_1 : p_2] &= \min_{c \in \mathcal{E}_F} \{(1-\beta) D^*_{\mathrm{KL}}[p_1 : c] + \beta D^*_{\mathrm{KL}}[p_2 : c]\}, &(100)\\ D^{\mathrm{vJS}}_{A_\beta,\mathcal{E}_F}[p_1 : p_2] &= \min_{c \in \mathcal{E}_F} \{(1-\beta) D_{\mathrm{KL}}[c : p_1] + \beta D_{\mathrm{KL}}[c : p_2]\}, &(101)\end{aligned}$$

where $A_\beta(a, b) := (1-\beta)a + \beta b$ is the weighted arithmetic mean for $\beta \in (0, 1)$.

Notice that, when $p = q$, we have $D^{\mathrm{vJS}}_{M,\mathcal{R}}[p : p] = \min_{c \in \mathcal{R}} D[p : c]$, which is the information projection [59] with respect to D of density p to the submanifold $\mathcal{R}$. Thus, when $p \notin \mathcal{R}$, we have $D^{\mathrm{vJS}}_{M,\mathcal{R}}[p : p] > 0$, i.e., the relative JSDs are not proper divergences, since a proper divergence ensures that $D[p : q] \geq 0$ with equality if $p = q$. Figure 1 illustrates the main cases of the relative Jensen-Shannon divergenc between p and q: Either p and q are both inside or outside $\mathcal{R}$, or one point is inside $\mathcal{R}$, while the other point is outside $\mathcal{R}$. When $p = q$, we get an information projection when both of the points are outside $\mathcal{R}$, and $D^{\mathcal{R}}_{\mathrm{vJS}}[p : p] = 0$ when $p \in \mathcal{R}$. When $p, q \in \mathcal{R}$ with $p \neq q$, the value $D^{\mathcal{R}}_{\mathrm{vJS}}[p : q]$ corresponds to the information radius (and the arg min to the right-sided Kullback–Leibler centroid).

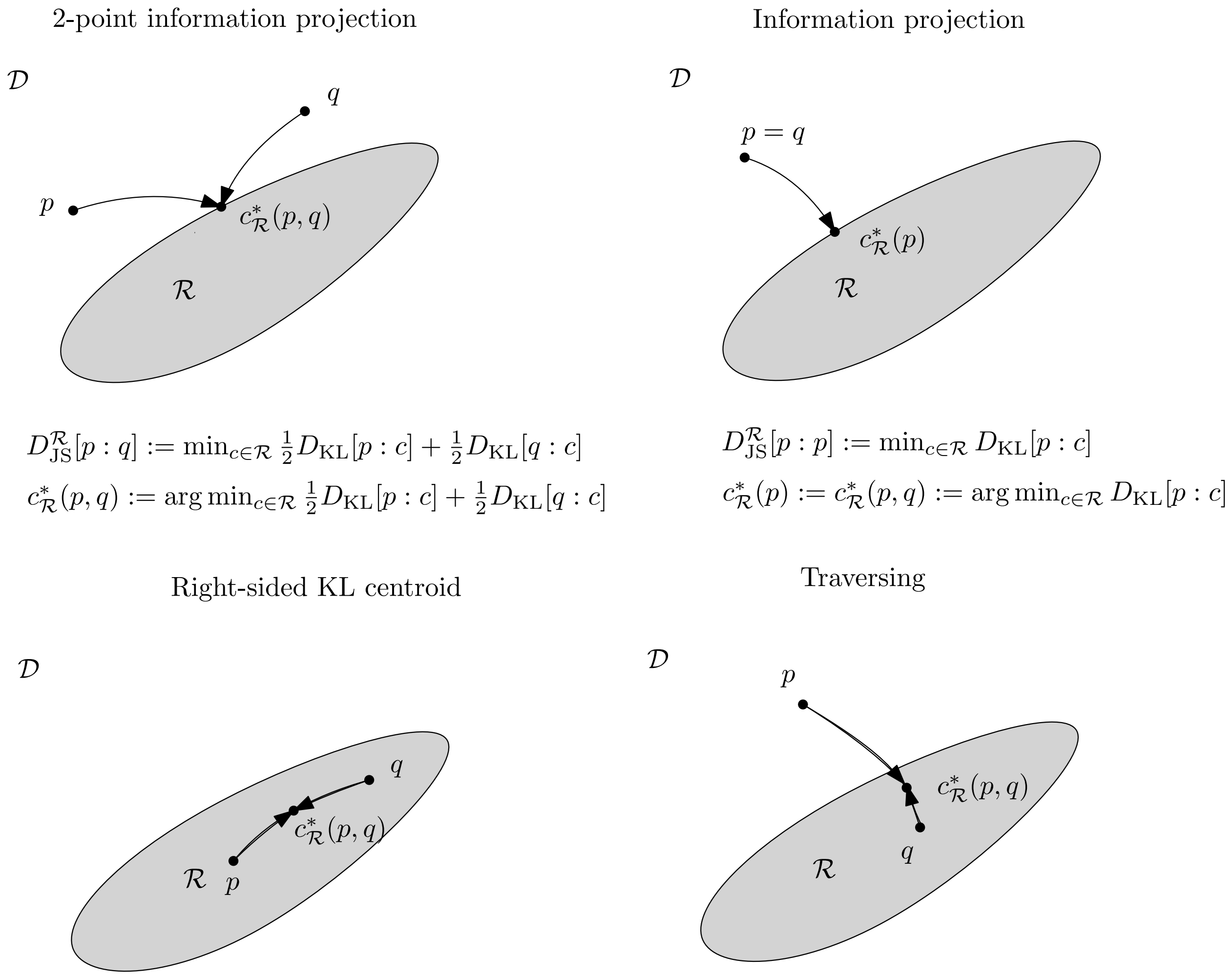
2-point information projection
$\mathcal{D}$
q
p
$c^*_{\mathcal{R}}(p,q)$
$\mathcal{R}$
$D^{\mathcal{R}}_{\mathrm{JS}}[p:q] := \min_{c\in\mathcal{R}} \frac{1}{2}D_{\mathrm{KL}}[p:c]+\frac{1}{2}D_{\mathrm{KL}}[q:c]$
$c^*_{\mathcal{R}}(p,q) := \arg\min_{c\in\mathcal{R}} \frac{1}{2}D_{\mathrm{KL}}[p:c]+\frac{1}{2}D_{\mathrm{KL}}[q:c]$
Information projection
$\mathcal{D}$
$p=q$
$c^*_{\mathcal{R}}(p)$
$\mathcal{R}$
$D^{\mathcal{R}}_{\mathrm{JS}}[p:p] := \min_{c\in\mathcal{R}} D_{\mathrm{KL}}[p:c]$
$c^*_{\mathcal{R}}(p) := c^*_{\mathcal{R}}(p,q) := \arg\min_{c\in\mathcal{R}} D_{\mathrm{KL}}[p:c]$
Right-sided KL centroid
$\mathcal{D}$
q
$c^*_{\mathcal{R}}(p,q)$
$\mathcal{R}$
p
Traversing
$\mathcal{D}$
p
$c^*_{\mathcal{R}}(p,q)$
q
$\mathcal{R}$

Figure 1. Illustrating several cases of the relative Jensen-Shannon divergence based on whether $p \in \mathcal{R}$ and $q \in \mathcal{R}$ or not.

4.2. *Relative Jensen-Shannon Divergences: Applications to Density Clustering and Quantization*

Let $D_{\mathrm{KL}}[p : q_\theta]$ be the Kullback–Leibler divergence between an *arbitrary* density p and a density q_θ of an exponential family $\mathcal{Q} = \{q_\theta \ : \ \theta \in \Theta\}$. Let us canonically express [2,60] the density $q_\theta(x)$, as

$$q_\theta(x) = \exp\left(\theta^\top t_{\mathcal{Q}}(x) - F_{\mathcal{Q}}(\theta) + k_{\mathcal{Q}}(x)\right),$$

where $t_{\mathcal{Q}}(x)$ denotes the sufficient statistics, $k_{\mathcal{Q}}(x)$ is an auxiliary carrier measure term (e.g., $k(x) = 0$ for the Gaussian family and $k(x) = \log(x)$ for the Rayleigh family [60]), and $F_{\mathcal{Q}}(\theta)$ the cumulant function. Assume that we know in closed-form the following quantities:

- $m_p := E_p[t_{\mathcal{Q}}(x)] = \int p(x) t_{\mathcal{Q}}(x) \mathrm{d}\mu(x)$ and
- the Shannon entropy $h[p] = -\int p(x) \log p(x) \mathrm{d}\mu(x)$ of p.

Subsequently, we can express the KLD using a semi-closed-form formula.

Proposition 4. *Let $q_\theta \in \mathcal{Q}$ be a density of an exponential family and p an arbitrary density with $m_p = E_p[t_{\mathcal{Q}}(x)]$. Subsequently, the Kullback–Leibler divergence between p and q_θ is expressed as:*

$$D_{\mathrm{KL}}[p : q_\theta] = F_{\mathcal{Q}}(\theta) - m_p^\top \theta - E_p[k_{\mathcal{Q}}(x)] - h[p], \tag{102}$$

where $h[p : q_\theta] = F_{\mathcal{Q}}(\theta) - m_p^\top \theta - E_p[k_{\mathcal{Q}}(x)]$ is the cross-entropy between p and q_θ.

Proof. The proof is straightforward since $\log q_\theta(x) = \theta^\top t_{\mathcal{Q}}(x) - F_{\mathcal{Q}}(\theta) + k_{\mathcal{Q}}(x)$. Therefore, we have:

$$\begin{aligned} D_{\mathrm{KL}}[p : q_\theta] &= h[p : q_\theta] - h[p], & (103)\\ &= -\int_{\mathcal{X}} p(x) \log q_\theta(x) \mathrm{d}\mu(x) - h[p], & (104)\\ &= F_{\mathcal{Q}}(\theta) - m_p^\top \theta - E_p[k_{\mathcal{Q}}(x)] - h[p]. & (105)\end{aligned}$$

□

Example 7. *For example, when $q_\theta = q_{\mu,\Sigma}$ is the density of a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ (with $k_{\mathcal{N}}(x) = 0$), we have*

$$D_{\mathrm{KL}}[p : q_{\mu,\Sigma}] = \frac{1}{2}\Big(\log|2\pi\Sigma| + (\mu - m)^\top \Sigma^{-1}(\mu - m) + \mathrm{tr}(\Sigma^{-1}S)\Big) - h[p], \qquad (106)$$

where $m = \mu(p) = E_p[X]$ and $S = \mathrm{Cov}(p) := E_p[XX^\top] - E_p[X]E_p[X]^\top$.

The formula of Proposition 4 is said in semi-closed-form, because it relies on knowing both the entropy h of p and the sufficient statistic moments $E_p[t_{\mathcal{Q}}(x)]$. Yet, this semi-closed formula may prove to be useful in practice: For example, we can answer the comparison predicate

"Is $D_{\mathrm{KL}}[p : q_{\theta_1}] \geq D_{\mathrm{KL}}[p : q_{\theta_2}]$ or not?"

by checking whether $F_{\mathcal{Q}}(\theta_1) - F_{\mathcal{Q}}(\theta_2) - m_p^\top(\theta_1 - \theta_2) \geq 0$ or not (i.e., the terms $-E_p[k_{\mathcal{Q}}(x)] - h[p]$ in Equation (102) cancel out). Thus, we get a closed-form predicate, although D_{KL} is only known in semi-closed-form. This KLD comparison predicate shall be used later on when clustering densities with respect to centroids in Section 4.2.

Remark 1. *Note that when $Y = f(X)$ for an invertible and differentiable transformation f then we have $h[Y] = h[X] + E_X[\log|J_f(X)|]$ where J_f denotes the Jacobian matrix. For example, when $Y = f(X) = AX$, we have $h[Y] = h[X] + \log|A|$.*

When p belongs to an exponential family $\mathcal{P}$ ($\mathcal{P}$ may be different from $\mathcal{Q}$) with cumulant function $F_{\mathcal{P}}$, sufficient statistics $t_{\mathcal{P}}(x)$, auxiliary carrier term $k_{\mathcal{P}}(x)$, and natural parameter θ, we have the entropy [61] expressed, as follows:

$$\begin{aligned} h[p] &= F_{\mathcal{P}}(\theta) - \theta^\top \nabla F_{\mathcal{P}}(\theta) - E_p[k_{\mathcal{P}}(x)], & (107)\\ &= -F_{\mathcal{P}}^*(\eta) - E_p[k_{\mathcal{P}}(x)], & (108)\end{aligned}$$

where $F_{\mathcal{P}}^*(\eta) = \theta^\top \nabla F(\theta) - F(\theta)$ is the Legendre transform of $F(\theta)$ and $\eta = \eta(\theta) = \nabla F(\theta)$ is called the moment parameter since we have $\eta(\theta) = E_p[t_{\mathcal{P}}(x)]$ [2,60].

It follows the following proposition refining Proposition 4 when $p = p_\theta \in \mathcal{P}$:

Proposition 5. *Let p_θ be a density of an exponential family $\mathcal{P}$ and $q_{\theta'}$ be a density of an exponential family $\mathcal{Q}$. Subsequently, the Kullback–Leibler divergence between p_θ and $q_{\theta'}$ is expressed as:*

$$D_{\mathrm{KL}}[p_\theta : q_{\theta'}] = F_{\mathcal{Q}}(\theta') + F_{\mathcal{P}}^*(\eta) - E_{p_\theta}[t_{\mathcal{Q}}(x)]^\top \theta' + E_{p_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)]. \qquad (109)$$

Proof. We have

$$\begin{aligned} D_{\mathrm{KL}}[p_\theta : q_{\theta'}] &= h[p_\theta : q_{\theta'}] - h[p_\theta], & (110)\\ &= F_{\mathcal{Q}}(\theta') - m_{p_\theta}^\top \theta' - E_{p_\theta}[k_{\mathcal{Q}}(x)] + F_{\mathcal{P}}^*(\eta) + E_{p_\theta}[k_{\mathcal{P}}(x)], & (111)\\ &= F_{\mathcal{Q}}(\theta') + F_{\mathcal{P}}^*(\eta) - E_{p_\theta}[t_{\mathcal{Q}}(x)]^\top \theta' + E_{p_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)]. & (112)\end{aligned}$$

□

In particular, when p and q belong both to the same exponential family (i.e., $\mathcal{P} = \mathcal{Q}$ with $k_{\mathcal{P}}(x) = k_{\mathcal{Q}}(x)$), we have $F(\theta) := F_{\mathcal{P}}(\theta) := F_{\mathcal{Q}}(\theta)$ and $E_{p_\theta}[t_{\mathcal{Q}}(x)] = \nabla F(\theta) =: \eta$, and

$$D_{\mathrm{KL}}[p_\theta : q_{\theta'}] = F(\theta') + F^*(\eta) - \theta'^\top \eta.$$

This last equation is the Fenchel–Young divergence in Bregman manifolds [34,62] (called dually flat spaces in information geometry [10]). Thus the divergence can be rewritten as equivalent dual Bregman divergences:

$$\begin{aligned} D_{\mathrm{KL}}[p_\theta : q_{\theta'}] &= F(\theta') + F^*(\eta) - \eta^\top \theta', && (113)\\ &= B_F(\theta' : \theta), && (114)\\ &= B_{F^*}(\eta : \eta'), && (115) \end{aligned}$$

where $\eta' = \nabla F(\theta')$.

Notice that $D_{\mathrm{KL}}[p_\theta : \mathcal{Q}] := \min_{\theta' \in \Theta'} D_{\mathrm{KL}}[p_\theta : q_{\theta'}]$ is unique and can be calculated as $\eta' = \nabla F_{\mathcal{Q}}(\theta') = E_{p_\theta}[t_{\mathcal{Q}}(x)]$.

Let us report two examples of calculations of the KLD between two densities of two exponential families.

Example 8. *For the first exponential family, consider the family of Laplacian distributions:*

$$\mathcal{P} = \mathcal{L} = \left\{ p_\sigma(x) := \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma} \right) \ : \ \sigma > 0 \right\}.$$

The canonical decomposition of the density yields $t_{\mathcal{L}}(x) = |x|$, $\theta = -\frac{1}{\sigma}$, $k_{\mathcal{L}}(x) = 0$, and $F_{\mathcal{L}}(\theta) = \log \frac{2}{-\theta}$. (i.e., $F_{\mathcal{L}}(\theta(\sigma)) = \log 2\sigma$). It follows that $\eta(\theta) = F'_{\mathcal{L}}(\theta) = -\frac{1}{\theta}$ ($\eta(\sigma) = \sigma = E[|x|]$), $\theta(\eta) = -\frac{1}{\eta}$, and $F^*_{\mathcal{L}}(\eta) = -1 - \log(2\eta)$ and, therefore, $F^*_{\mathcal{L}}(\eta(\sigma)) = -1 - \log(2\sigma)$.

For the second family, consider the exponential family of zero-centered Gaussian distributions:

$$\mathcal{Q} = \mathcal{N}_0 = \left\{ q_{\sigma'}(x) = \frac{1}{\sqrt{2\pi(\sigma')^2}} \exp\left(-\frac{x^2}{2(\sigma')^2} \right) \right\}.$$

We have $t_{\mathcal{N}_0}(x) = x^2$, $k_{\mathcal{N}_0}(x) = 0$, $\theta' = -\frac{1}{2(\sigma')^2}$, and $F_{\mathcal{N}_0}(\sigma') = \frac{1}{2}\log(2\pi(\sigma')^2)$.

Moreover, let us calculate $E_{p_\sigma}[t_{\mathcal{N}_0}(x)] = E_{p_\sigma}[x^2] = 2\sigma^2$. Subsequently, we can calculate the Kullback–Leibler divergence between $p_\sigma \sim \mathcal{L}(\sigma)$ and $q_{\sigma'} \sim \mathcal{N}_0(\sigma')$, as follows:

$$\begin{aligned} D_{\mathrm{KL}}[p_\sigma : q_{\sigma'}] &= F_{\mathcal{Q}}(\theta'(\sigma')) + F^*_{\mathcal{P}}(\eta(\sigma)) - E_{p_\sigma}[t_{\mathcal{Q}}(x)]^\top \theta'(\sigma') + E_{p_\sigma}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)], && (116)\\ &= \frac{1}{2}\log(2\pi(\sigma')^2) - 1 - \log(2\sigma) - 2\sigma^2\left(-\frac{1}{2(\sigma')^2} \right), && (117)\\ &= \log\left(\frac{\sigma'}{\sigma}\right) + \left(\frac{\sigma}{\sigma'}\right)^2 + \frac{1}{2}\log\left(\frac{\pi}{2}\right) - 1. && (118) \end{aligned}$$

Notice that $D_{\mathrm{KL}}[p_\sigma : q_{\sigma'}] \geq 0$, but never 0 since the $\mathcal{P} \cap \mathcal{Q} = \emptyset$.

Let us now compute the reverse Kullback–Leibler divergence $D_{\mathrm{KL}}[q_{\sigma'} : p_\sigma]$. We first calculate $E_{q_{\sigma'}}[t_{\mathcal{L}}(x)] = E_{q_{\sigma'}(\sigma')}[|x|] = \sqrt{\frac{2}{\pi}}\sigma'$. Since $F_{\mathcal{Q}}(\theta') = \frac{1}{2}\log(\frac{\pi}{-\theta'})$, we have $\eta'(\theta') = F'_{\mathcal{Q}}(\theta') = -\frac{1}{2\theta'}$. Thus $\eta'(\sigma') = (\sigma')^2$ and $F^*_{\mathcal{Q}}(\eta') = -\frac{1}{2} - \frac{1}{2}\log(2\pi\eta)$. Therefore, we get $F^*_{\mathcal{Q}}(\eta'(\sigma')) = -h[q_{\sigma'}] = -\frac{1}{2}\log(2\pi e(\sigma')^2)$.

It follows that

$$\begin{aligned} D_{\mathrm{KL}}[q_{\sigma'} : p_\sigma] &= F_{\mathcal{P}}(\theta(\sigma)) + F^*_{\mathcal{Q}}(\eta'(\sigma')) - E_{q_{\theta'}}[t_{\mathcal{P}}(x)]^\top \theta(\sigma) + E_{q_{\theta'}}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)], && (119)\\ &= \log(2\sigma) - \frac{1}{2}\log(2\pi e(\sigma')^2) - \sqrt{\frac{2}{\pi}}\sigma' \times \left(-\frac{1}{\sigma} \right), && (120)\\ &= \sqrt{\frac{2}{\pi}\frac{\sigma'}{\sigma}} + \log\left(\frac{\sigma}{\sigma'}\right) - \frac{1}{2}\log(\frac{\pi}{2}e). && (121) \end{aligned}$$

Again, we have $D_{\mathrm{KL}}[q_{\sigma'} : p_\sigma] \geq 0$*, but never* 0*, because* $\mathcal{P} \cap \mathcal{Q} = \emptyset$*.*

Example 9. *Let us use the formula of Equation (109) to calculate the KLD between two Weibull distributions [63]. A Weibull distribution of shape* $\kappa > 0$ *and scale* $\sigma > 0$ *has a density defined on* $\mathcal{X} = [0, \infty)$*, as follows:*

$$p^{\mathrm{Wei}}_{\kappa,\sigma}(x) := \frac{\kappa}{\sigma}\left(\frac{x}{\sigma}\right)^{\kappa-1} \exp\left(-\left(\frac{x}{\sigma}\right)^{\kappa}\right).$$

For a fixed shape κ*, the set of Weibull distributions* $\mathcal{W}_\kappa := \{p^{\mathrm{Wei}}_{\kappa,\sigma} : \sigma > 0\}$ *form an exponential family with natural parameter* $\theta = -\frac{1}{\sigma^\kappa}$*, sufficient statistic* $t_\kappa(x) = x^\kappa$*, auxiliary carrier term* $k_\kappa(x) = (\kappa - 1)\log x + \log \kappa$*, and cumulant function* $F_\kappa(\theta) = -\log(-\theta)$ *(so that* $F_\kappa(\theta(\sigma)) = F_\kappa(\sigma) = \kappa \log \sigma$*):*

$$p^{\mathrm{Wei}}_{\kappa,\sigma}(x) := \exp\left(-\frac{1}{\sigma^\kappa} x^k + \log \frac{1}{\sigma^\kappa} + k(x)\right).$$

We recover the exponential family of exponential distributions of rate parameter $\lambda = \frac{1}{\sigma}$ *when* $\kappa = 1$*:*

$$\begin{aligned} p^{\mathrm{Exp}}_\lambda(x) &= p^{\mathrm{Wei}}_{1,\sigma}(x) = \frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right), \\ &= \lambda \exp(-\lambda x), \end{aligned}$$

and the exponential family of Rayleigh distributions when $\kappa = 2$ *with scale parameter* $\sigma_{\mathrm{Ray}} = \frac{\sigma}{\sqrt{2}}$*:*

$$\begin{aligned} p^{\mathrm{Ray}}_{\sigma_{\mathrm{Ray}}}(x) &= p^{\mathrm{Wei}}_{2,\sigma}(x) = \frac{2x}{\sigma^2}\exp\left(-\frac{x^2}{\sigma^2}\right), \\ &= \frac{x}{\sigma^2_{\mathrm{Ray}}}\exp\left(-\frac{x^2}{2\sigma^2_{\mathrm{Ray}}}\right). \end{aligned}$$

Now, assume that we are given the differential entropy of the Weibull distributions [64] (pp. 155–156):

$$h\left[p^{\mathrm{Wei}}_{\kappa_1,\sigma_1}\right] = \gamma\left(1 - \frac{1}{\kappa_1}\right) + \log\frac{\sigma_1}{\kappa_1} + 1,$$

where $\gamma \approx 0.5772156649$ *is the Euler–Mascheroni constant, and the Weibull raw moments [64] (p. 155):*

$$m = E_{p^{\mathrm{Wei}}_{\kappa_1,\sigma}}[x^{\kappa_2}] = \sigma_1^{\kappa_2}\Gamma\left(1 + \frac{\kappa_2}{\kappa_1}\right),$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ *is the gamma function (with* $\Gamma(n) = (n-1)!$ *for integers* n*). Because* $h[p^{\mathrm{Wei}}_{\kappa,\sigma}] = F_\kappa(\theta) - \theta^\top \nabla F_\kappa(\theta) - E_{p^{\mathrm{Wei}}_{\kappa,\sigma}}[k_\kappa(x)] = -F^*_\kappa(\eta) - E_{p^{\mathrm{Wei}}_{\kappa,\sigma}}[k_\kappa(x)]$*, we deduce that*

$$E_{p^{\mathrm{Wei}}_{\kappa,\sigma}}[k_\kappa(x)] = -F^*_\kappa(\eta) - h\left[p^{\mathrm{Wei}}_{\kappa,\sigma}\right],$$

where $F^*_\kappa(\eta)$ *is the Legendre transform of* $F_\kappa(\theta)$ *and* $\eta(\theta) = \nabla F_\kappa(\theta) = -\frac{1}{\theta} = E[t(x)] = E[x^\kappa]$*. We have* $\theta(\eta) = \nabla F^*_\kappa(\eta) = -\frac{1}{\eta}$ *and* $F^*_\kappa(\eta) = \eta^\top \nabla F^*_\kappa(\eta) - F_\kappa(\nabla F^*_\kappa(\eta)) = -1 - \log \eta$*. It follows that*

$$E_{p^{\mathrm{Wei}}_{\kappa,\sigma}}[k_\kappa(x)] = 1 + \log\left(\sigma\Gamma\left(1 + \frac{1}{\kappa}\right)\right) - \gamma\left(1 - \frac{1}{\kappa}\right) - \log\frac{\sigma}{\kappa} + 1.$$

Therefore, we deduce that the logarithmic moment of $p^{\mathrm{Wei}}_{\kappa_1,\sigma}$ *is:*

$$E_{p^{\mathrm{Wei}}_{\kappa_1,\sigma}}[\log x] = -\frac{\gamma}{\kappa_1} + \log \sigma_1.$$

This coincides with the explicit definite integral calculation reported in [63].

Subsequently, we calculate the KLD between two Weibull distributions using Equation (109), as follows:

$$
\begin{aligned}
D_{\mathrm{KL}}\left[p^{\mathrm{Wei}}_{\kappa_1,\sigma_1} : p^{\mathrm{Wei}}_{\kappa_2,\sigma_2}\right] &= F_{\kappa_2}(\theta') + F^*_{\kappa_1}(\eta) - E_{p_{\kappa_1,\sigma_1}}[x^{\kappa_2}]^\top \theta' + E_{p_{\kappa_1,\sigma_1}}[k_{\kappa_1}(x) - k_{\kappa_2}(x)] \qquad (122)\\
&= \log\frac{\kappa_1}{\sigma_1^{\kappa_1}} - \log\frac{\kappa_2}{\sigma_2^{\kappa_2}} + (\kappa_1 - \kappa_2)\left[\log\sigma_1 - \frac{\gamma}{\kappa_1}\right] + \left(\frac{\sigma_1}{\sigma_2}\right)^{\kappa_2}\Gamma\left(\frac{\kappa_2}{\kappa_1} + 1\right) - 1 \qquad (123)
\end{aligned}
$$

since we have the following terms:

$$
\begin{aligned}
F_{\kappa_2}(\theta') &= \log\sigma_2^{\kappa_2},\\
F^*_{\kappa_1}(\eta) &= -1 - \log\sigma_1^{\kappa_1},\\
-E_{p_{\kappa_1,\sigma_1}}[x^{\kappa_2}]^\top\theta' &= \frac{1}{\sigma_2^{\kappa_2}}\sigma_1^{\kappa_2}\Gamma\left(1 + \frac{\kappa_2}{\kappa_1}\right)\\
E_{p_{\kappa_1,\sigma_1}}[k_{\kappa_1}(x) - k_{\kappa_2}(x)] &= (\kappa_1 - \kappa_2)E_{p_{\kappa_1,\sigma_1}}[\log x] + \log\frac{\kappa_1}{\kappa_2},\\
&= \log\frac{\kappa_1}{\kappa_2} + (\kappa_1 - \kappa_2)\left(\log\sigma_1 - \frac{\gamma}{\kappa_1}\right).
\end{aligned}
$$

This formula matches the formula reported in [63].

When $\kappa_1 = \kappa_2 = 1$, we recover the ordinary KLD formula between two exponential distributions [60] with $\lambda_i = \frac{1}{\sigma_i}$ since $\Gamma(2) = (2-1)! = 1$:

$$
\begin{aligned}
D_{\mathrm{KL}}\left[p^{\mathrm{Wei}}_{1,\sigma_1} : p^{\mathrm{Wei}}_{1,\sigma_2}\right] &= \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1}{\sigma_2} - 1, \qquad (124)\\
&= \frac{\lambda_2}{\lambda_1} - \log\frac{\lambda_2}{\lambda_1} - 1. \qquad (125)
\end{aligned}
$$

When $\kappa_1 = \kappa_2 = 2$, we recover the ordinary KLD formula between two Rayleigh distributions [60], with $\sigma_{\mathrm{Ray}} = \frac{\sigma}{\sqrt{2}}$:

$$
\begin{aligned}
D_{\mathrm{KL}}\left[p^{\mathrm{Wei}}_{2,\sigma_1} : p^{\mathrm{Wei}}_{2,\sigma_2}\right] &= \log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} - 1, \qquad (126)\\
&= \log\left(\frac{{\sigma_{\mathrm{Ray}}}_2^2}{{\sigma_{\mathrm{Ray}}}_1^2}\right) + \frac{{\sigma_{\mathrm{Ray}}}_1^2}{{\sigma_{\mathrm{Ray}}}_2^2} - 1. \qquad (127)
\end{aligned}
$$

The formulae of Equations (125) and (127) are linked by the fact that if $X \sim \mathrm{Exp}(\lambda)$ and $Y = \sqrt{X}$ then $Y \sim \mathrm{Ray}\left(\frac{1}{\sqrt{2\lambda}}\right)$, and f-divergences [65], including the Kullback–Leibler divergence are invariant by a differentiable transformation [66].

Jeffreys' divergence symmetrizes the KLD divergence, as follows:

$$D_J[p:q] := D_{\mathrm{KL}}[p:q] + D_{\mathrm{KL}}[q:p] = 2A(D_{\mathrm{KL}}[p:q], D_{\mathrm{KL}}[q:p]). \qquad (128)$$

The Jeffreys divergence between two densities of different exponential families $\mathcal{P}$ and $\mathcal{Q}$ is

$$D_J[p_\theta : q_{\theta'}] = \theta'^\top(\eta' - E_{p_\theta}[t_{\mathcal{Q}}(x)]) + \theta^\top(\eta - E_{q_{\theta'}}[t_{\mathcal{P}}(x)]) + E_{p_\theta}[k_{\mathcal{P}}(x) - k_{\mathcal{Q}}(x)] + E_{q_{\theta'}}[k_{\mathcal{Q}}(x) - k_{\mathcal{P}}(x)]. \qquad (129)$$

When $\mathcal{P} = \mathcal{Q}$, we have $E_{p_\theta}[t_{\mathcal{Q}}(x)] = \eta$ and $E_{q_{\theta'}}[t_{\mathcal{P}}(x)]) = \eta'$, so that we find the usual expression of the Jeffreys divergence between two densities of an exponential family:

$$D_J[p_\theta : p_{\theta'}] = (\theta' - \theta)^\top(\eta' - \eta). \qquad (130)$$

To find the best density q_θ approximating p by minimizing $\min_\theta D_{\mathrm{KL}}[p : q_\theta]$, we solve $\nabla F(\theta) = \eta = m$ and, therefore, $\theta = \nabla F^*(m) = (\nabla F)^{-1}(m)$, where $F^*(\eta) = E_{q_\eta}[\log q_\eta(m)]$, with F^* denoting the Legendre–Fenchel convex conjugate [2]. In particular, when $p = \sum w_i p_{\theta_i}$ is a mixture of EFs (with $m = E_p[t(x)] = \sum w_i \eta_i$ with $\eta_i = E_{p_{\theta_i}}[t(x)]$ thanks to the linearity of the expectation), then the best density of the EF simplifying p is

$$\begin{aligned}\min_\theta D_{\mathrm{KL}}[p : q_\theta] &= \min_\theta F(\theta) - m^\top \theta, &(131)\\ &= \min_\theta F(\theta) - \sum w_i \eta_i^\top \theta. &(132)\end{aligned}$$

Taking the gradient with respect to θ, we have $\nabla F(\theta) = \eta = \sum w_i \eta_i$. This yields another proof without the Pythagoras theorem [67,68].

Proposition 6. *Let $m(x) = \sum w_i p_{\theta_i}(x)$ be a mixture with components that belong to an exponential family with cumulant function F. Subsequently, $\theta^* = \arg_\theta \min_\theta D_{\mathrm{KL}}[p : q_\theta]$ is $\nabla F^*(\sum_{i=1}^n w_i \eta_i)$, where the $\eta_i = \nabla F(\theta_i)$ are the moment parameters of the mixture components.*

Consider the following two problems:

Problem 1 (Density clustering). *Given a set of n weighted densities $(w_1, p_1), \ldots, (w_n, p_n)$, partition them into k clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ in order to minimize the k-centroid objective function with respect to a statistical divergence D: $\sum_{i=1}^n w_i \min_{l \in \{1,\ldots,k\}} D[p_i : c_l]$, where c_l denotes the centroid of cluster $\mathcal{C}_l$ for $l \in \{1, \ldots, k\}$.*

For example, when all the densities p_i's are isotropic Gaussians, we recover the k-means objective function [69].

Problem 2 (Mixture component quantization). *Given a statistical mixture $m(x) = \sum_{i=1}^n w_i\, p_i(x)$, quantize the mixture components into k densities $q_1, \ldots, q_k$ in order to minimize $\sum_i w_i \min_{l \in \{1,\ldots,k\}} D[p_i : q_l]$.*

Notice that, in Problem 1, the input densities p_i's may be mixtures, i.e., $p_i(x) = \sum_{j=1}^{n_i} w_{i,j} p_{i,j}(x)$. Using the relative information radius, we can cluster a set of distributions (potentially mixtures) into an exponential family mixture, or quantize an exponential family mixture. Indeed, we can implement an extension of k-means [69] with k-centers q_{θ_i}, to assign density p_i to cluster C_j (with center q_j), we need to perform basic comparison tests $D_{\mathrm{KL}}[p_i : q_{\theta_l}] \geq D_{\mathrm{KL}}[p_i : q_{\theta_j}]$. Provided that the cumulant F of the exponential family is in closed-form, we do not need formula for the entropies $h(p_i)$.

Clustering and quantization of densities/mixtures have been widely studied in the literature, see, for example, [70–76].

5. Conclusions

To summarize, the ordinary Jensen-Shannon divergence has been defined in three equivalent ways in the literature:

$$\begin{aligned}D_{\mathrm{JS}}[p, q] &:= \min_{c \in \mathcal{D}} \frac{1}{2}(D_{\mathrm{KL}}[p : c] + D_{\mathrm{KL}}[q : c]), &(133)\\ &= \frac{1}{2}\left(D_{\mathrm{KL}}\left[p : \frac{p+q}{2}\right] + D_{\mathrm{KL}}\left[q : \frac{p+q}{2}\right]\right), &(134)\\ &= h\left[\frac{p+q}{2}\right] - \frac{h[p] + h[q]}{2}. &(135)\end{aligned}$$

The JSD Equation (133) was studied by Sibson in 1969 within the wider scope of information radius [1]: Sibson relied on the Rényi α-divergences (relative Rényi α-entropies [77]) and recovered the ordinary Jensen-Shannon divergence as a particular case of the α-

information radius when $\alpha = 1$ and $n = 2$ points. The α-information radii are related to generalized Bhattacharyya distances with respect to power means and the total variation distance in the limit case of $\alpha = \infty$.

Lin [4] investigated the JSD Equation (134) in 1991 with its connection to the JSD defined in Equation (134)). In Lin [4], the JSD is interpreted as the arithmetic symmetrization of the K-divergence [24]. Generalizations of the JSD based on Equation (134) were proposed in [23] using a generic mean instead of the arithmetic mean. One motivation was to obtain a closed-form formula for the geometric JSD between multivariate Gaussian distributions, which relies on the geometric mixture (see [30] for a use case of that formula in deep learning). Indeed, the ordinary JSD between Gaussians is not available in closed-form (not analytic). However, the JSD between Cauchy distributions admit a closed-form formula [78], despite the calculation of a definite integral of a log-sum term. Instead of using an abstract mean to define a mid-distribution of two densities, one may also consider the mid-point of a geodesic linking these two densities (the arithmetic means $\frac{p+q}{2}$ is interpreted as a geodesic midpoint). Recently, Li [79] investigated the transport Jensen-Shannon divergence as a symmetrization of the Kullback–Leibler divergence in the L^2-Wasserstein space. See Section 5.4 of [79] and the closed-form formula of Equation (18) obtained for the transport Jensen-Shannon divergence between two multivariate Gaussian distributions.

The generalization of the identity between the JSD of Equation (134) and the JSD of Equation (135) was studied while using a skewing vector in [18]. Although the JSD is a f-divergence [8,18], the Sibson-M Jensen-Shannon symmetrization of a distance does not belong, in general, to the class of f-divergences. The variational JSD definition of Equation (133) is implicit, while the definitions of Equations (134) and (135) are explicit because the unique optimal centroid $c^* = \frac{p+q}{2}$ has been plugged into the objective function that was minimized by Equation (133).

In this paper, we proposed a generalization of the Jensen-Shannon divergence based on the variational definition of the ordinary Jensen-Shannon divergence based on the variational JSD definition of Equation (133): $D_{\mathrm{vJS}}[p:q] = \min_c \frac{1}{2}(D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c])$. We introduced the Jensen-Shannon symmetrization of an arbitrary divergence D by considering a generalization of the information radius with respect to an abstract weighted mean M_β: $D_M^{\mathrm{vJS}}[p:q] := \min_c M_\beta(D[p:c], D[q:c])$. Notice that, in the variational JSD, the mean M_β is used for averaging divergence values, while the mean M_α in the (M_α, N_β) JSD is used to define generic statistical mixtures. We also consider relative variational JS symmetrization when the centroid has to belong to a prescribed family of densities. For the case of exponential family, we showed how to compute the relative centroid in closed form, thus extending the pioneering work of Sibson, who considered the relative normal centroid used to calculate the relative normal information radius. Figure 2 illustrates the three generalizations of the ordinary skewed Jensen-Shannon divergence. Notice that, in general, the (M, N)-JSDs and the variational JDSs are not f-divergences (except in the ordinary case).

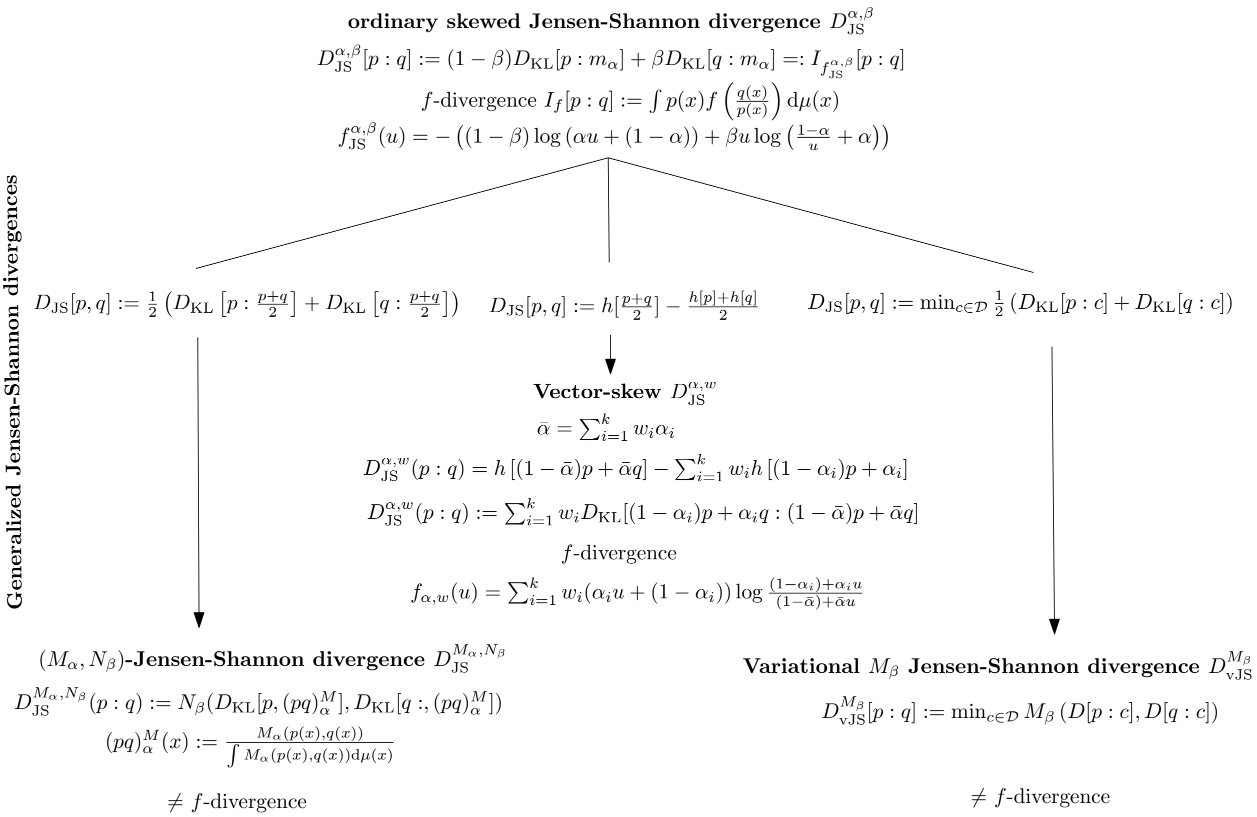
ordinary skewed Jensen-Shannon divergence $D_{\mathrm{JS}}^{\alpha,\beta}$
$D_{\mathrm{JS}}^{\alpha,\beta}[p:q] := (1-\beta)D_{\mathrm{KL}}[p:m_\alpha] + \beta D_{\mathrm{KL}}[q:m_\alpha] =: I_{f_{\mathrm{JS}}^{\alpha,\beta}}[p:q]$
f-divergence $I_f[p:q] := \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)$
$f_{\mathrm{JS}}^{\alpha,\beta}(u) = -\left((1-\beta)\log(\alpha u + (1-\alpha)) + \beta u \log\left(\frac{1-\alpha}{u} + \alpha\right)\right)$
Generalized Jensen-Shannon divergences
$D_{\mathrm{JS}}[p,q] := \frac{1}{2}\left(D_{\mathrm{KL}}\left[p:\frac{p+q}{2}\right] + D_{\mathrm{KL}}\left[q:\frac{p+q}{2}\right]\right)$
$D_{\mathrm{JS}}[p,q] := h[\frac{p+q}{2}] - \frac{h[p]+h[q]}{2}$
$D_{\mathrm{JS}}[p,q] := \min_{c\in\mathcal{D}} \frac{1}{2}\left(D_{\mathrm{KL}}[p:c] + D_{\mathrm{KL}}[q:c]\right)$
Vector-skew $D_{\mathrm{JS}}^{\alpha,w}$
$\bar{\alpha} = \sum_{i=1}^k w_i \alpha_i$
$D_{\mathrm{JS}}^{\alpha,w}(p:q) = h\left[(1-\bar{\alpha})p + \bar{\alpha}q\right] - \sum_{i=1}^k w_i h\left[(1-\alpha_i)p + \alpha_i\right]$
$D_{\mathrm{JS}}^{\alpha,w}(p:q) := \sum_{i=1}^k w_i D_{\mathrm{KL}}[(1-\alpha_i)p + \alpha_i q : (1-\bar{\alpha})p + \bar{\alpha}q]$
f-divergence
$f_{\alpha,w}(u) = \sum_{i=1}^k w_i(\alpha_i u + (1-\alpha_i)) \log \frac{(1-\alpha_i)+\alpha_i u}{(1-\bar{\alpha})+\bar{\alpha}u}$
(M_α, N_β)**-Jensen-Shannon divergence** $D_{\mathrm{JS}}^{M_\alpha,N_\beta}$
$D_{\mathrm{JS}}^{M_\alpha,N_\beta}(p:q) := N_\beta(D_{\mathrm{KL}}[p,(pq)_\alpha^M], D_{\mathrm{KL}}[q:,(pq)_\alpha^M])$
$(pq)_\alpha^M(x) := \frac{M_\alpha(p(x),q(x))}{\int M_\alpha(p(x),q(x))\mathrm{d}\mu(x)}$
$\neq f$-divergence
Variational M_β **Jensen-Shannon divergence** $D_{\mathrm{vJS}}^{M_\beta}$
$D_{\mathrm{vJS}}^{M_\beta}[p:q] := \min_{c\in\mathcal{D}} M_\beta\left(D[p:c], D[q:c]\right)$
$\neq f$-divergence

Figure 2. Three equivalent expressions of the ordinary (skewed) Jensen-Shannon divergence which yield three different generalizations.

In a similar vein, Chen et al. [80] considered the following minimax symmetrization of the scalar Bregman divergence [81]:

$$
\begin{aligned}
B_f^{\mathrm{minmax}}(p,q) &:= \min_{c} \max_{\lambda\in[0,1]} \lambda B_f(p:c) + (1-\lambda)B_f(q:c), && (136)\\
&= \max_{\lambda\in[0,1]} \lambda B_f(p:\lambda p + (1-\lambda)q) + (1-\lambda)B_f(q:\lambda p + (1-\lambda)), && (137)\\
&= \lambda f(p) + (1-\lambda)f(q) - f(\lambda p + (1-\lambda)) && (138)
\end{aligned}
$$

where B_f denotes the scalar Bregman divergence induced by a strictly convex and smooth function f:

$$
B_f(p:q) = f(p) - f(q) - (p-q)f'(q). \tag{139}
$$

They proved that $\sqrt{B_f^{\mathrm{minmax}}(p,q)}$ yields a metric when $3(\log f'')'' \geq ((\log f'')')^2$, and extend the definition to the vector case and conjecture that the square-root metrization still holds in the multivariate case. In a sense, this definition geometrically highlights the notion of radius, since the minmax optimization amount to find a smallest enclosing ball enclosing [82] the source distributions. The circumcenter, also called the Chebyshev center [83], is then the mid-distribution instead of the centroid for the information radius. The term "information radius" is well-suited to measure the distance between two points for an arbitrary distance D. Indeed, the JS-symmetrization of D is defined by $D^{\mathrm{JS}}[p:q] := \min_c\{\frac{1}{2}D[p:c] + \frac{1}{2}D[q:c]\}$. When $D[p:q] = D_E[p:q] = \|p-q\|$ is the Euclidean distance, we have $c = \frac{p+q}{2}$, and $D[p:c] = D[q:c] = \frac{1}{2}\|p-q\| =: r$ (i.e., the radius being half of the diameter $\|p-q\|$). Thus, $D_E^{\mathrm{JS}}[p:q] = r$; hence, the term chosen by Sibson [1] for D^{JS}: information radius. Besides providing another viewpoint, variational definitions of divergences have proven to be useful in practice (e.g., for estimation). For example, a variational definition of the Rényi divergence generalizing the Donsker–

Varadhan variational formula of the KLD is given in [84], which is used to estimate the Rényi Divergences.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: We warmly thank Rob Brekelmans (Information Sciences Institute, University of Southern California, USA) for discussions and feedback related to the contents of this work. The author thanks the reviewers for valuable feedback, comments, and suggestions, and Gaëtan Hadjeres (Sony CSL Paris) for his careful reading of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sibson, R. Information radius. *Z. Wahrscheinlichkeitstheorie Verwandte Geb.* **1969**, *14*, 149–160. [CrossRef]
2. Barndorff-Nielsen, O. *Information and Exponential Families: In Statistical Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
3. Billingsley, P. *Probability and Measure*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
4. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
5. Kullback, S. *Information Theory and Statistics*; Courier Corporation: Chelmsford, MA, USA, 1997.
6. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
7. Morimoto, T. Markov processes and the H-theorem. *J. Phys. Soc. Jpn.* **1963**, *18*, 328–331. [CrossRef]
8. Csiszár, I. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kut. Int. Koezl.* **1964**, *8*, 85–108.
9. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B (Methodological)* **1966**, *28*, 131–142. [CrossRef]
10. Amari, S.i. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Tokyo, Japan, 2016.
11. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
12. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [CrossRef]
13. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [CrossRef]
14. Fuglede, B.; Topsoe, F. Jensen-Shannon divergence and Hilbert space embedding. In Proceedings of the International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings, Chicago, IL, USA, 27 June–2 July 2004; IEEE: Piscataway, NJ, USA, 2004; p. 31.
15. Virosztek, D. The metric property of the quantum Jensen-Shannon divergence. *Adv. Math.* **2021**, *380*, 107595. [CrossRef]
16. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661.
17. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
18. Nielsen, F. On a generalization of the Jensen-Shannon divergence and the Jensen-Shannon centroid. *Entropy* **2020**, *22*, 221. [CrossRef]
19. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
20. Csiszár, I. Axiomatic characterizations of information measures. *Entropy* **2008**, *10*, 261–273. [CrossRef]
21. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
22. Antolín, J.; Angulo, J.; López-Rosa, S. Fisher and Jensen-Shannon divergences: Quantitative comparisons among distributions. application to position and momentum atomic densities. *J. Chem. Phys.* **2009**, *130*, 074110. [CrossRef] [PubMed]
23. Nielsen, F. On the Jensen-Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485. [CrossRef]
24. Nielsen, F. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv* **2010**, arXiv:1009.4004.
25. Nielsen, F.; Nock, R. Generalizing skew Jensen divergences and Bregman divergences with comparative convexity. *IEEE Signal Process. Lett.* **2017**, *24*, 1123–1127. [CrossRef]
26. de Carvalho, M. Mean, what do you Mean? *Am. Stat.* **2016**, *70*, 270–274. [CrossRef]
27. Bullen, P.S. *Handbook of Means and Their Inequalities*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 560.
28. Niculescu, C.P.; Persson, L.E. *Convex Functions and Their Applications: A Contemporary Approach*; Springer: Berlin/Heidelberg, Germany, 2018.
29. Nielsen, F. Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. *Pattern Recognit. Lett.* **2014**, *42*, 25–34. [CrossRef]
30. Deasy, J.; Simidjievski, N.; Liò, P. Constraining Variational Inference with Geometric Jensen-Shannon Divergence. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020.
31. Amari, S.I. Integration of stochastic models by minimizing α-divergence. *Neural Comput.* **2007**, *19*, 2780–2796. [CrossRef]

32. Calin, O.; Udriste, C. *Geometric Modeling in Probability and Statistics*; Mathematics and Statistics; Springer International Publishing: Berlin/Heidelberg, Germany, 2014.
33. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1961; Volume 1: Contributions to the Theory of Statistics; The Regents of the University of California: Oakland, CA, USA, 1961.
34. Blondel, M.; Martins, A.F.; Niculae, V. Learning with Fenchel-Young losses. *J. Mach. Learn. Res.* **2020**, *21*, 1–69.
35. Faddeev, D.K. Zum Begriff der Entropie einer endlichen Wahrscheinlichkeitsschemas. In *Arbeiten zur Informationstheorie I*; Deutscher Verlag der Wissenschaften: Berlin, Germany, 1957; pp. 85–90.
36. Kolmogorov, A.N.; Castelnuovo, G. *Sur la Notion de la Moyenne*; Bardi, G., Ed.; Atti della Academia Nazionale dei Lincei: Rome, Italy, 1930; Volume 12, pp. 323–343.
37. Nagumo, M. Über eine klasse der mittelwerte. In *Japanese Journal of Mathematics: Transactions and Abstracts*; The Mathematical Society of Japan: Tokyo, Japan, 1930; Volume 7, pp. 71–79.
38. De Finetti, B. *Sul Concetto di Media*; Istituto Italiano Degli Attuari: Roma, Italy, 1931.
39. Van Erven, T.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
40. Sibson, R. A brief description of natural neighbour interpolation. In *Interpreting Multivariate Data*; Barnett, V., Ed.; John Wiley & Sons: Hoboken, NJ, USA, 1981; pp. 21–36. .
41. Boyd, S.; Boyd, S.P.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
42. Nielsen, F.; Sun, K. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy* **2016**, *18*, 442. [CrossRef]
43. Nielsen, F. Chernoff information of exponential families. *arXiv* **2011**, arXiv:1102.2684.
44. Nielsen, F. An information-geometric characterization of Chernoff information. *IEEE Signal Process. Lett.* **2013**, *20*, 269–272. [CrossRef]
45. Nielsen, F.; Yvinec, M. An output-sensitive convex hull algorithm for planar objects. *Int. J. Comput. Geom. Appl.* **1998**, *8*, 39–65. [CrossRef]
46. Nielsen, F.; Nock, R. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Process. Lett.* **2013**, *21*, 10–13. [CrossRef]
47. Nielsen, F. The statistical Minkowski distances: Closed-form formula for Gaussian mixture models. In *International Conference on Geometric Science of Information*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 359–367.
48. Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. L'Institut Henri Poincaré* **1948**, *10*, 215–310.
49. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904. [CrossRef]
50. Naudts, J. *Generalised Thermostatistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
51. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. [CrossRef]
52. Nielsen, F. On Voronoi diagrams on the information-geometric Cauchy manifolds. *Entropy* **2020**, *22*, 713. [CrossRef] [PubMed]
53. Nock, R.; Nielsen, F.; Amari, S.i. On conformal divergences and their population minimizers. *IEEE Trans. Inf. Theory* **2015**, *62*, 527–538. [CrossRef]
54. Brekelmans, R.; Nielsen, F.; Makhzani, A.; Galstyan, A.; Steeg, G.V. Likelihood Ratio Exponential Families. *arXiv* **2020**, arXiv:2012.15480.
55. Brekelmans, R.; Masrani, V.; Bui, T.; Wood, F.; Galstyan, A.; Steeg, G.V.; Nielsen, F. Annealed Importance Sampling with q-Paths. *arXiv* **2020**, arXiv:2012.07823.
56. Nielsen, F. A generalization of the α-divergences based on comparable and distinct weighted means. *arXiv* **2020**, arXiv:2001.09660.
57. Amari, S.i.; Ohara, A. Geometry of q-exponential family of probability distributions. *Entropy* **2011**, *13*, 1170–1185. [CrossRef]
58. Grosse, R.; Maddison, C.J.; Salakhutdinov, R. Annealing between distributions by averaging moments. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 2769–2777.
59. Nielsen, F. What is an information projection? *Not. AMS* **2018**, *65*, 321–324. [CrossRef]
60. Nielsen, F.; Garcia, V. Statistical exponential families: A digest with flash cards. *arXiv* **2009**, arXiv:0911.4863.
61. Nielsen, F.; Nock, R. Entropies and cross-entropies of exponential families. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 3621–3624.
62. Nielsen, F. On Geodesic Triangles with Right Angles in a Dually Flat Space. In *Progress in Information Geometry: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 153–190.
63. Bauckhage, C. Computing the Kullback-Leibler divergence between two Weibull distributions. *arXiv* **2013**, arXiv:1310.3713.
64. Michalowicz, J.V.; Nichols, J.M.; Bucholtz, F. *Handbook of Differential Entropy*; CRC Press: Boca Raton, FL, USA, 2013.
65. Csiszár, I. On topological properties of f-divergences. *Stud. Math. Hungar.* **1967**, *2*, 329–339.
66. Nielsen, F. On information projections between multivariate elliptical and location-scale families. *arXiv* **2021**, arXiv:2101.03839.
67. Pelletier, B. Informative barycentres in statistics. *Ann. Inst. Stat. Math.* **2005**, *57*, 767–780. [CrossRef]
68. Schwander, O.; Nielsen, F. Learning mixtures by simplifying kernel density estimators. In *Matrix Information Geometry*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 403–426.
69. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]

70. Davis, J.V.; Dhillon, I. Differential entropic clustering of multivariate Gaussians. In Proceedings of the 19th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; pp. 337–344.
71. Nielsen, F.; Nock, R. Clustering multivariate normal distributions. In *Emerging Trends in Visual Computing*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 164–174.
72. Fischer, A. Quantization and clustering with Bregman divergences. *J. Multivar. Anal.* **2010**, *101*, 2207–2221. [CrossRef]
73. Zhang, K.; Kwok, J.T. Simplifying mixture models through function approximation. *IEEE Trans. Neural Netw.* **2010**, *21*, 644–658. [CrossRef] [PubMed]
74. Duan, J.; Wang, Y. Information-Theoretic Clustering for Gaussian Mixture Model via Divergence Factorization. In Proceedings of the 2013 Chinese Intelligent Automation Conference, Yangzhou, China, 23–25 August 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 565–573.
75. Wang, J.C.; Yang, Y.H.; Wang, H.M.; Jeng, S.K. Modeling the affective content of music with a Gaussian mixture model. *IEEE Trans. Affect. Comput.* **2015**, *6*, 56–68. [CrossRef]
76. Spurek, P.; Pałka, W. Clustering of Gaussian distributions. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, USA, 24–29 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 3346–3353.
77. Esteban, M.D.; Morales, D. A summary on entropy statistics. *Kybernetika* **1995**, *31*, 337–346.
78. Nielsen, F.; Okamura, K. On f-divergences between Cauchy distributions. *arXiv* **2021**, arXiv:2101.12459.
79. Li, W. Transport information Bregman divergences. *arXiv* **2021**, arXiv:2101.01162.
80. Chen, P.; Chen, Y.; Rao, M. Metrics defined by Bregman divergences: Part 2. *Commun. Math. Sci.* **2008**, *6*, 927–948. [CrossRef]
81. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217. [CrossRef]
82. Arnaudon, M.; Nielsen, F. On approximating the Riemannian 1-center. *Comput. Geom.* **2013**, *46*, 93–104. [CrossRef]
83. Candan, Ç. Chebyshev Center Computation on Probability Simplex With α-Divergence Measure. *IEEE Signal Process. Lett.* **2020**, *27*, 1515–1519. [CrossRef]
84. Birrell, J.; Dupuis, P.; Katsoulakis, M.A.; Rey-Bellet, L.; Wang, J. Variational Representations and Neural Network Estimation for Rényi Divergences. *arXiv* **2020**, arXiv:2007.03814.