entropy

MDPI

Article

On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid

Frank Nielsen

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

Received: 5 December 2019; Accepted: 14 February 2020; Published: 16 February 2020

check for updates

Abstract: The Jensen–Shannon divergence is a renown bounded symmetrization of the Kullback–Leibler divergence which does not require probability densities to have matching supports. In this paper, we introduce a vector-skew generalization of the scalar α-Jensen–Bregman divergences and derive thereof the vector-skew α-Jensen–Shannon divergences. We prove that the vector-skew α-Jensen–Shannon divergences are f-divergences and study the properties of these novel divergences. Finally, we report an iterative algorithm to numerically compute the Jensen–Shannon-type centroids for a set of probability densities belonging to a mixture family: This includes the case of the Jensen–Shannon centroid of a set of categorical distributions or normalized histograms.

Keywords: Bregman divergence; f-divergence; Jensen–Bregman divergence; Jensen diversity; Jensen–Shannon divergence; capacitory discrimination; Jensen–Shannon centroid; mixture family; information geometry; difference of convex (DC) programming

1. Introduction

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space [1] where $\mathcal{X}$ denotes the sample space, $\mathcal{F}$ the σ-algebra of measurable events, and μ a positive measure; for example, the measure space defined by the Lebesgue measure μ_L with Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ for $\mathcal{X} = \mathbb{R}^d$ or the measure space defined by the counting measure μ_c with the power set σ-algebra $2^{\mathcal{X}}$ on a finite alphabet $\mathcal{X}$. Denote by $L_1(\mathcal{X}, \mathcal{F}, \mu)$ the Lebesgue space of measurable functions, $\mathcal{P}_1$ the subspace of *positive* integrable functions f such that $\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) = 1$ and $f(x) > 0$ for all $x \in \mathcal{X}$, and $\overline{\mathcal{P}}_1$ the subspace of *non-negative* integrable functions f such that $\int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) = 1$ and $f(x) \geq 0$ for all $x \in \mathcal{X}$.

We refer to the book of Deza and Deza [2] and the survey of Basseville [3] for an introduction to the many types of statistical divergences met in information sciences and their justifications. The *Kullback–Leibler Divergence* (KLD) $\mathrm{KL} : \mathcal{P}_1 \times \mathcal{P}_1 \to [0, \infty]$ is an oriented statistical distance (commonly called the relative entropy in information theory [4]) defined between two densities p and q (i.e., the Radon–Nikodym densities of μ-absolutely continuous probability measures P and Q) by

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

Although $\mathrm{KL}(p : q) \geq 0$ with equality iff. $p = q$ μ-a. e. (Gibb's inequality [4]), the KLD may diverge to infinity depending on the underlying densities. Since the KLD is asymmetric, several symmetrizations [5] have been proposed in the literature.

A well-grounded symmetrization of the KLD is the *Jensen–Shannon Divergence* [6] (JSD), also called *capacitory discrimination* in the literature (e.g., see [7]):

$$\begin{aligned} \mathrm{JS}(p,q) &:= \frac{1}{2}\left(\mathrm{KL}\left(p:\frac{p+q}{2}\right)+\mathrm{KL}\left(q:\frac{p+q}{2}\right)\right), &(2)\\ &= \frac{1}{2}\int\left(p\log\frac{2p}{p+q}+q\log\frac{2q}{p+q}\right)\mathrm{d}\mu=\mathrm{JS}(q,p). &(3)\end{aligned}$$

The Jensen–Shannon divergence can be interpreted as the *total KL divergence to the average distribution* $\frac{p+q}{2}$. The Jensen–Shannon divergence was historically implicitly introduced in [8] (Equation (19)) to calculate distances between random graphs. A nice feature of the Jensen–Shannon divergence is that this divergence can be applied to densities with *arbitrary* support (i.e., $p,q\in\overline{\mathcal{P}}_1$ with the convention that $0\log 0=0$ and $\log\frac{0}{0}=0$); moreover, the JSD is *always* upper bounded by $\log 2$. Let $\mathcal{X}_p=\mathrm{supp}(p)$ and $\mathcal{X}_q=\mathrm{supp}(q)$ denote the supports of the densities p and q, respectively, where $\mathrm{supp}(p):=\{x\in\mathcal{X}\ :\ p(x)>0\}$. The JSD saturates to $\log 2$ whenever the supports $\mathcal{X}_p$ and $\mathcal{X}_p$ are disjoints. We can rewrite the JSD as

$$\mathrm{JS}(p,q)=h\left(\frac{p+q}{2}\right)-\frac{h(p)+h(q)}{2}, \tag{4}$$

where $h(p)=-\int p\log p\mathrm{d}\mu$ denotes Shannon's entropy. Thus, the JSD can also be interpreted as the *entropy of the average distribution minus the average of the entropies*.

The square root of the JSD is a metric [9] satisfying the triangle inequality, but the square root of the JD is not a metric (nor any positive power of the Jeffreys divergence, see [10]). In fact, the JSD can be interpreted as a Hilbert metric distance, meaning that there exists some isometric embedding of $(\mathcal{X},\sqrt{\mathrm{JS}})$ into a Hilbert space [11,12]. Other principled symmetrizations of the KLD have been proposed in the literature: For example, Naghshvar et al. [13] proposed the *extrinsic Jensen–Shannon divergence* and demonstrated its use for variable-length coding over a discrete memoryless channel (DMC).

Another symmetrization of the KLD sometimes met in the literature [14–16] is the *Jeffreys divergence* [17,18] (JD) defined by

$$J(p,q):=\mathrm{KL}(p:q)+\mathrm{KL}(q:p)=\int(p-q)\log\frac{p}{q}\mathrm{d}\mu=J(q,p). \tag{5}$$

However, we point out that this Jeffreys divergence lacks sound information-theoretical justifications.

For two positive but not necessarily normalized densities $\tilde{p}$ and $\tilde{q}$, we define the *extended Kullback–Leibler divergence* as follows:

$$\begin{aligned} \mathrm{KL}^+(\tilde{p}:\tilde{q}) &:= \mathrm{KL}(\tilde{p}:\tilde{q})+\int\tilde{q}\mathrm{d}\mu-\int\tilde{p}\mathrm{d}\mu, &(6)\\ &= \int\left(\tilde{p}\log\frac{\tilde{p}}{\tilde{q}}+\tilde{q}-\tilde{p}\right)\mathrm{d}\mu. &(7)\end{aligned}$$

The Jensen–Shannon divergence and the Jeffreys divergence can both be extended to positive (unnormalized) densities without changing their formula expressions:

$$\begin{aligned} \mathrm{JS}^+(\tilde{p},\tilde{q}) &:= \frac{1}{2}\left(\mathrm{KL}^+\left(\tilde{p}:\frac{\tilde{p}+\tilde{q}}{2}\right)+\mathrm{KL}^+\left(\tilde{q}:\frac{\tilde{p}+\tilde{q}}{2}\right)\right), &(8)\\ &= \frac{1}{2}\left(\mathrm{KL}\left(\tilde{p}:\frac{\tilde{p}+\tilde{q}}{2}\right)+\mathrm{KL}\left(\tilde{q}:\frac{\tilde{p}+\tilde{q}}{2}\right)\right)=\mathrm{JS}(\tilde{p},\tilde{q}), &(9)\\ J^+(\tilde{p},\tilde{q}) &:= \mathrm{KL}^+(\tilde{p}:\tilde{q})+\mathrm{KL}^+(\tilde{p}:\tilde{q})=\int(\tilde{p}-\tilde{q})\log\frac{\tilde{p}}{\tilde{q}}\mathrm{d}\mu=J(\tilde{p},\tilde{q}). &(10)\end{aligned}$$

However, the extended JS^{+} divergence is upper-bounded by $(\frac{1}{2}\log 2)(\int(\tilde{p}+\tilde{q})\mathrm{d}\mu) = \frac{1}{2}(\mu(p)+\mu(q))\log 2$ instead of $\log 2$ for normalized densities (i.e., when $\mu(p)+\mu(q)=2$).

Let $(pq)_\alpha(x) := (1-\alpha)p(x)+\alpha q(x)$ denote the statistical weighted mixture with component densities p and q for $\alpha\in[0,1]$. The asymmetric α-skew Jensen–Shannon divergence can be defined for a scalar parameter $\alpha\in(0,1)$ by considering the weighted mixture $(pq)_\alpha$ as follows:

$$\begin{aligned}\mathrm{JS}_a^\alpha(p:q) &:= (1-\alpha)\mathrm{KL}(p:(pq)_\alpha)+\alpha\mathrm{KL}(q:(pq)_\alpha), &(11)\\ &= (1-\alpha)\int p\log\frac{p}{(pq)_\alpha}\mathrm{d}\mu+\alpha\int q\log\frac{q}{(pq)_\alpha}\mathrm{d}\mu. &(12)\end{aligned}$$

Let us introduce the *α-skew K-divergence* [6,19] $K_\alpha(p:q)$ by:

$$K_\alpha(p:q) := \mathrm{KL}(p:(1-\alpha)p+\alpha q) = \mathrm{KL}(p:(pq)_\alpha). \quad (13)$$

Then, both the Jensen–Shannon divergence and the Jeffreys divergence can be rewritten [20] using K_α as follows:

$$\begin{aligned}\mathrm{JS}(p,q) &= \frac{1}{2}\left(K_{\frac{1}{2}}(p:q)+K_{\frac{1}{2}}(q:p)\right), &(14)\\ J(p,q) &= K_1(p:q)+K_1(q:p), &(15)\end{aligned}$$

since $(pq)_1 = q$, $\mathrm{KL}(p:q) = K_1(p:q)$ and $(pq)_{\frac{1}{2}} = (qp)_{\frac{1}{2}}$.

We can thus define the *symmetric α-skew Jensen–Shannon divergence* [20] for $\alpha\in(0,1)$ as follows:

$$\mathrm{JS}^\alpha(p,q) := \frac{1}{2}K_\alpha(p:q)+\frac{1}{2}K_\alpha(q:p) = \mathrm{JS}^\alpha(q,p). \quad (16)$$

The ordinary Jensen–Shannon divergence is recovered for $\alpha=\frac{1}{2}$.

In general, skewing divergences (e.g., using the divergence K_α instead of the KLD) have been experimentally shown to perform better in applications like in some natural language processing (NLP) tasks [21].

The *α-Jensen–Shannon divergences* are Csiszár f-divergences [22–24]. An f-divergence is defined for a convex function f, strictly convex at 1, and satisfies $f(1)=0$ as:

$$I_f(p:q) = \int q(x)f\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x \geq f(1) = 0. \quad (17)$$

We can always symmetrize f-divergences by taking the *conjugate* convex function $f^*(x) = xf(\frac{1}{x})$ (related to the perspective function): $I_{f+f^*}(p,q)$ is a symmetric divergence. The f-divergences are convex statistical distances which are provably the only separable invariant divergences in information geometry [25], except for binary alphabets $\mathcal{X}$ (see [26]).

The Jeffreys divergence is an f-divergence for the generator $f(x) = (x-1)\log x$, and the α-Jensen–Shannon divergences are f-divergences for the generator family $f_\alpha(x) = -\log((1-\alpha)+\alpha x)-x\log((1-\alpha)+\frac{\alpha}{x})$. The f-divergences are upper-bounded by $f(0)+f^*(0)$. Thus, the f-divergences are finite when $f(0)+f^*(0)<\infty$.

The main contributions of this paper are summarized as follows:

- First, we generalize the Jensen–Bregman divergence by skewing a weighted separable Jensen–Bregman divergence with a k-dimensional *vector* $\alpha\in[0,1]^k$ in Section 2. This yields a generalization of the symmetric skew α-Jensen–Shannon divergences to a vector-skew parameter. This extension retains the key properties for being upper-bounded and for application to densities with potentially different supports. The proposed generalization also allows one to grasp a better understanding of the "mechanism" of the Jensen–Shannon divergence itself. We also show how to

directly obtain the weighted vector-skew Jensen–Shannon divergence from the decomposition of the KLD as the difference of the cross-entropy minus the entropy (i.e., KLD as the relative entropy).

- Second, we prove that weighted vector-skew Jensen–Shannon divergences are f-divergences (Theorem 1), and show how to build families of symmetric Jensen–Shannon-type divergences which can be controlled by a vector of parameters in Section 2.3, generalizing the work of [20] from scalar skewing to vector skewing. This may prove useful in applications by providing additional tuning parameters (which can be set, for example, by using cross-validation techniques).
- Third, we consider the calculation of the *Jensen–Shannon centroids* in Section 3 for densities belonging to mixture families. Mixture families include the family of categorical distributions and the family of statistical mixtures sharing the same prescribed components. Mixture families are well-studied manifolds in information geometry [25]. We show how to compute the Jensen–Shannon centroid using a concave–convex numerical iterative optimization procedure [27]. The experimental results graphically compare the Jeffreys centroid with the Jensen–Shannon centroid for grey-valued image histograms.

2. Extending the Jensen–Shannon Divergence

2.1. Vector-Skew Jensen–Bregman Divergences and Jensen Diversities

Recall our notational shortcut: $(ab)_\alpha := (1-\alpha)a + \alpha b$. For a k-dimensional vector $\alpha \in [0,1]^k$, a weight vector w belonging to the $(k-1)$-dimensional open simplex Δ_k, and a scalar $\gamma \in (0,1)$, let us define the following vector *skew α-Jensen–Bregman divergence* (α-JBD) following [28]:

$$\mathrm{JB}_F^{\alpha,\gamma,w}(\theta_1 : \theta_2) := \sum_{i=1}^{k} w_i B_F\left((\theta_1\theta_2)_{\alpha_i} : (\theta_1\theta_2)_\gamma\right) \geq 0, \tag{18}$$

where B_F is the *Bregman divergence* [29] induced by a strictly convex and smooth generator F:

$$B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2)\rangle, \tag{19}$$

with $\langle \cdot, \cdot \rangle$ denoting the Euclidean inner product $\langle x, y\rangle = x^\top y$ (dot product). Expanding the Bregman divergence formulas in the expression of the α-JBD and using the fact that

$$(\theta_1\theta_2)_{\alpha_i} - (\theta_1\theta_2)_\gamma = (\gamma - \alpha_i)(\theta_1 - \theta_2), \tag{20}$$

we get the following expression:

$$\mathrm{JB}_F^{\alpha,\gamma,w}(\theta_1 : \theta_2) = \left(\sum_{i=1}^{k} w_i F\left((\theta_1\theta_2)_{\alpha_i}\right)\right) - F\left((\theta_1\theta_2)_\gamma\right) - \left\langle \sum_{i=1}^{k} w_i(\gamma - \alpha_i)(\theta_1 - \theta_2), \nabla F((\theta_1\theta_2)_\gamma)\right\rangle. \tag{21}$$

The inner product term of Equation (21) vanishes when

$$\gamma = \sum_{i=1}^{k} w_i \alpha_i := \bar{\alpha}. \tag{22}$$

Thus, when $\gamma = \bar{\alpha}$ (assuming at least two distinct components in α so that $\gamma \in (0,1)$), we get the simplified formula for the vector-skew α-JBD:

$$\boxed{\mathrm{JB}_F^{\alpha,w}(\theta_1 : \theta_2) = \left(\sum_{i=1}^{k} w_i F\left((\theta_1\theta_2)_{\alpha_i}\right)\right) - F\left((\theta_1\theta_2)_{\bar{\alpha}}\right).} \tag{23}$$

This vector-skew Jensen–Bregman divergence is always finite and amounts to a *Jensen diversity* [30] J_F induced by Jensen's inequality gap:

$$\mathrm{JB}_F^{\alpha,w}(\theta_1 : \theta_2) = J_F((\theta_1\theta_2)_{\alpha_1}, \ldots, (\theta_1\theta_2)_{\alpha_k}; w_1, \ldots, w_k) := \sum_{i=1}^{k} w_i F\left((\theta_1\theta_2)_{\alpha_i}\right) - F\left((\theta_1\theta_2)_{\bar{\alpha}}\right) \geq 0. \quad (24)$$

The Jensen diversity is a quantity which arises as a generalization of the cluster variance when clustering with Bregman divergences instead of the ordinary squared Euclidean distance; see [29,30] for details. In the context of Bregman clustering, the Jensen diversity has been called the *Bregman information* [29] and motivated by rate distortion theory: Bregman information measures the minimum expected loss when encoding a set of points using a single point when the loss is measured using a Bregman divergence. In general, a k-point measure is called a diversity measure (for $k > 2$), while a distance/divergence is the special case of a 2-point measure.

Conversely, in 1D, we may start from Jensen's inequality for a strictly convex function F:

$$\sum_{i=1}^{k} w_i F(\theta_i) \geq F\left(\sum_{i=1}^{k} w_i \theta_i\right). \quad (25)$$

Let us notationally write $[k] := \{1, \ldots, k\}$, and define $\theta_m := \min_{i\in[k]}\{\theta_i\}_i$ and $\theta_M := \max_{i\in[k]}\{\theta_i\}_i > \theta_m$ (i.e., assuming at least two distinct values). We have the barycenter $\bar{\theta} = \sum_i w_i\theta_i =: (\theta_m\theta_M)_\gamma$ which can be interpreted as the linear interpolation of the extremal values for some $\gamma \in (0, 1)$. Let us write $\theta_i = (\theta_m\theta_M)_{\alpha_i}$ for $i \in [k]$ and proper values of the α_is. Then, it comes that

$$\begin{aligned}
\bar{\theta} &= \sum_i w_i\theta_i, & (26)\\
&= \sum_i w_i(\theta_m\theta_M)_{\alpha_i}, & (27)\\
&= \sum_i w_i((1-\alpha_i)\theta_m + \alpha_i\theta_M), & (28)\\
&= \left(1 - \sum_i w_i\alpha_i\right)\theta_m + \sum_i \alpha_i w_i \theta_M, & (29)\\
&= (\theta_m\theta_M)_{\sum_i w_i\alpha_i} = (\theta_m\theta_M)_\gamma, & (30)
\end{aligned}$$

so that $\gamma = \sum_i w_i\alpha_i = \bar{\alpha}$.

2.2. Vector-Skew Jensen–Shannon Divergences

Let $f(x) = x\log x - x$ be a strictly smooth convex function on $(0, \infty)$. Then, the Bregman divergence induced by this univariate generator is

$$B_f(p : q) = p\log\frac{p}{q} + q - p = \mathrm{kl}_+(p : q), \quad (31)$$

the *extended scalar Kullback–Leibler divergence*.

We extend the scalar-skew Jensen–Shannon divergence as follows: $\mathrm{JS}^{\alpha,w}(p : q) := \mathrm{JB}_{-h}^{\alpha,\bar{\alpha},w}(p : q)$ for h, the Shannon's entropy [4] (a strictly concave function [4]).

Definition 1 (Weighted vector-skew (α, w)-Jensen–Shannon divergence)**.** *For a vector $\alpha \in [0, 1]^k$ and a unit positive weight vector $w \in \Delta_k$, the (α, w)-Jensen–Shannon divergence between two densities $p, q \in \bar{\mathcal{P}}_1$ is defined by:*

$$\boxed{\mathrm{JS}^{\alpha,w}(p : q) := \sum_{i=1}^{k} w_i \mathrm{KL}((pq)_{\alpha_i} : (pq)_{\bar{\alpha}}) = h\left((pq)_{\bar{\alpha}}\right) - \sum_{i=1}^{k} w_i h\left((pq)_{\alpha_i}\right),}$$

with $\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i$, *where* $h(p) = -\int p(x) \log p(x) \mathrm{d}\mu(x)$ *denotes the Shannon entropy [4] (i.e.,* $-h$ *is strictly convex).*

This definition generalizes the ordinary JSD; we recover the ordinary Jensen–Shannon divergence when $k = 2$, $\alpha_1 = 0$, $\alpha_2 = 1$, and $w_1 = w_2 = \frac{1}{2}$ with $\bar{\alpha} = \frac{1}{2}$: $\mathrm{JS}(p, q) = \mathrm{JS}^{(0,1),(\frac{1}{2},\frac{1}{2})}(p : q)$.

Let $\mathrm{KL}_{\alpha,\beta}(p : q) := \mathrm{KL}((pq)_\alpha : (pq)_\beta)$. Then, we have $\mathrm{KL}_{\alpha,\beta}(q : p) = \mathrm{KL}_{1-\alpha,1-\beta}(p : q)$. Using this (α, β)-KLD, we have the following identity:

$$\begin{aligned} \mathrm{JS}^{\alpha,w}(p : q) &= \sum_{i=1}^{k} w_i \mathrm{KL}_{\alpha_i,\bar{\alpha}}(p : q), & (32) \\ &= \sum_{i=1}^{k} w_i \mathrm{KL}_{1-\alpha_i,1-\bar{\alpha}}(q : p) = \mathrm{JS}^{1_k-\alpha,w}(q : p), & (33) \end{aligned}$$

since $\sum_{i=1}^{k} w_i(1 - \alpha_i) = \overline{1_k - \alpha} = 1 - \bar{\alpha}$, where $1_k = (1, \ldots, 1)$ is a k-dimensional vector of ones.

A very interesting property is that the vector-skew Jensen–Shannon divergences are f-divergences [22].

Theorem 1. *The vector-skew Jensen–Shannon divergences* $\mathrm{JS}^{\alpha,w}(p : q)$ *are* f*-divergences for the generator* $f_{\alpha,w}(u) = \sum_{i=1}^{k} w_i(\alpha_i u + (1 - \alpha_i)) \log \frac{(1-\alpha_i)+\alpha_i u}{(1-\bar{\alpha})+\bar{\alpha}u}$ *with* $\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i$.

Proof. First, let us observe that the positively weighted sum of f-divergences is an f-divergence: $\sum_{i=1}^{k} w_i I_{f_i}(p : q) = I_f(p : q)$ for the generator $f(u) = \sum_{i=1}^{k} w_i f_i(u)$.

Now, let us express the divergence $\mathrm{KL}_{\alpha,\beta}(p : q)$ as an f-divergence:

$$\mathrm{KL}_{\alpha,\beta}(p : q) = I_{f_{\alpha,\beta}}(p : q), \quad (34)$$

with generator

$$f_{\alpha,\beta}(u) = (\alpha u + 1 - \alpha) \log \frac{(1 - \alpha) + \alpha u}{(1 - \beta) + \beta u}. \quad (35)$$

Thus, it follows that

$$\begin{aligned} \mathrm{JS}^{\alpha,w}(p : q) &= \sum_{i=1}^{k} w_i \mathrm{KL}((pq)_{\alpha_i} : (pq)_{\bar{\alpha}}), & (36) \\ &= \sum_{i=1}^{k} w_i I_{f_{\alpha_i,\bar{\alpha}}}(p : q), & (37) \\ &= I_{\sum_{i=1}^{k} w_i f_{\alpha_i,\bar{\alpha}}}(p : q). & (38) \end{aligned}$$

Therefore, the vector-skew Jensen–Shannon divergence is an f-divergence for the following generator:

$$\boxed{f_{\alpha,w}(u) = \sum_{i=1}^{k} w_i(\alpha_i u + (1 - \alpha_i)) \log \frac{(1 - \alpha_i) + \alpha_i u}{(1 - \bar{\alpha}) + \bar{\alpha} u},} \quad (39)$$

where $\bar{\alpha} = \sum_{i=1}^{k} w_i \alpha_i$.

When $\alpha = (0, 1)$ and $w = (\frac{1}{2}, \frac{1}{2})$, we recover the f-divergence generator for the JSD:

$$\begin{aligned} f_{\mathrm{JS}}(u) &= \frac{1}{2} \log \frac{1}{\frac{1}{2} + \frac{1}{2}u} + \frac{1}{2} u \log \frac{u}{\frac{1}{2} + \frac{1}{2}u}, & (40) \\ &= \frac{1}{2} \left(\log \frac{2}{1 + u} + u \log \frac{2u}{1 + u} \right). & (41) \end{aligned}$$

Observe that $f^*_{\alpha,w}(u) = uf_{\alpha,w}(1/u) = f_{1-\alpha,w}(u)$, where $1-\alpha := (1-\alpha_1, \ldots, 1-\alpha_k)$.

We also refer the reader to Theorem 4.1 of [31], which defines skew f-divergences from any f-divergence. □

Remark 1. *Since the vector-skew Jensen divergence is an f-divergence, we easily obtain Fano and Pinsker inequalities following [32], or reverse Pinsker inequalities following [33,34] (i.e., upper bounds for the vector-skew Jensen divergences using the total variation metric distance), data processing inequalities using [35], etc.*

Next, we show that $\mathrm{KL}_{\alpha,\beta}$ (and $\mathrm{JS}^{\alpha,w}$) are separable convex divergences. Since the f-divergences are separable convex, the $\mathrm{KL}_{\alpha,\beta}$ divergences and the $\mathrm{JS}^{\alpha,w}$ divergences are separable convex. For the sake of completeness, we report a simplex explicit proof below.

Theorem 2 (Separable convexity)**.** *The divergence $\mathrm{KL}_{\alpha,\beta}(p:q)$ is strictly separable convex for $\alpha \neq \beta$ and $x \in \mathcal{X}_p \cap \mathcal{X}_q$.*

Proof. Let us calculate the second partial derivative of $\mathrm{KL}_{\alpha,\beta}(x:y)$ with respect to x, and show that it is strictly positive:

$$\frac{\partial^2}{\partial x^2}\mathrm{KL}_{\alpha,\beta}(x:y) = \frac{(\beta-\alpha)^2 y^2}{(xy)_\alpha (xy)^2_\beta} > 0, \tag{42}$$

for $x, y > 0$. Thus, $\mathrm{KL}_{\alpha,\beta}$ is strictly convex on the left argument. Similarly, since $\mathrm{KL}_{\alpha,\beta}(y:x) = \mathrm{KL}_{1-\alpha,1-\beta}(x:y)$, we deduce that $\mathrm{KL}_{\alpha,\beta}$ is strictly convex on the right argument. Therefore, the divergence $\mathrm{KL}_{\alpha,\beta}$ is separable convex. □

It follows that the divergence $\mathrm{JS}^{\alpha,w}(p:q)$ is strictly separable convex, since it is a convex combination of weighted $\mathrm{KL}_{\alpha_i,\bar{\alpha}}$ divergences.

Another way to derive the vector-skew JSD is to decompose the KLD as the difference of the cross-entropy $h^\times$ minus the entropy h (i.e., KLD is also called the relative entropy):

$$\mathrm{KL}(p:q) = h^\times(p:q) - h(p), \tag{43}$$

where $h^\times(p:q) := -\int p \log q \mathrm{d}\mu$ and $h(p) := h^\times(p:p)$ (self cross-entropy). Since $\alpha_1 h^\times(p_1:q) + \alpha_2 h^\times(p_2:q) = h^\times(\alpha_1 p_1 + \alpha_2 p_2 : q)$ (for $\alpha_2 = 1-\alpha_1$), it follows that

$$\begin{aligned}
\mathrm{JS}^{\alpha,w}(p:q) &:= \sum_{i=1}^k w_i \mathrm{KL}((pq)_{\alpha_i} : (pq)_\gamma), && (44)\\
&= \sum_{i=1}^k w_i \left(h^\times((pq)_{\alpha_i} : (pq)_\gamma) - h((pq)_{\alpha_i}) \right), && (45)\\
&= h^\times\left(\sum_{i=1}^k w_i (pq)_{\alpha_i} : (pq)_\gamma \right) - \sum_{i=1}^k w_i h\left((pq)_{\alpha_i}\right). && (46)
\end{aligned}$$

Here, the "trick" is to choose $\gamma = \bar{\alpha}$ in order to "convert" the cross-entropy into an entropy: $h^\times(\sum_{i=1}^k w_i (pq)_{\alpha_i} : (pq)_\gamma) = h((pq)_{\bar{\alpha}})$ when $\gamma = \bar{\alpha}$. Then, we end up with

$$\boxed{\mathrm{JS}^{\alpha,w}(p:q) = h\left((pq)_{\bar{\alpha}}\right) - \sum_{i=1}^k w_i h\left((pq)_{\alpha_i}\right).} \tag{47}$$

When $\alpha = (\alpha_1, \alpha_2)$ with $\alpha_1 = 0$ and $\alpha_2 = 0$ and $w = (w_1, w_2) = (\frac{1}{2}, \frac{1}{2})$, we have $\bar{\alpha} = \frac{1}{2}$, and we recover the Jensen–Shannon divergence:

$$\mathrm{JS}(p : q) = h\left(\frac{p+q}{2}\right) - \frac{h(p)+h(q)}{2}. \tag{48}$$

Notice that Equation (13) is the usual definition of the Jensen–Shannon divergence, while Equation (48) is the reduced formula of the JSD, which can be interpreted as a Jensen gap for Shannon entropy, hence its name: The *Jensen–Shannon divergence*.

Moreover, if we consider the cross-entropy/entropy extended to positive densities $\tilde{p}$ and $\tilde{q}$:

$$h_+^\times(\tilde{p} : \tilde{q}) = -\int (\tilde{p} \log \tilde{q} + \tilde{q}) \mathrm{d}\mu, \quad h_+(\tilde{p}) = h_+^\times(\tilde{p} : \tilde{p}) = -\int (\tilde{p} \log \tilde{p} + \tilde{p}) \mathrm{d}\mu, \tag{49}$$

we get:

$$\mathrm{JS}_+^{\alpha,w}(\tilde{p} : \tilde{q}) = \sum_{i=1}^k w_i \mathrm{KL}_+((\tilde{p}\tilde{q})_{\alpha_i} : (\tilde{p}\tilde{q})_\gamma) = h_+((\tilde{p}\tilde{q})_{\bar{\alpha}}) - \sum_{i=1}^k w_i h_+((\tilde{p}\tilde{q})_{\alpha_i}). \tag{50}$$

Next, we shall prove that our generalization of the skew Jensen–Shannon divergence to vector-skewing is always bounded. We first start by a lemma bounding the KLD between two mixtures sharing the same components:

Lemma 1 (KLD between two w-mixtures)**.** *For $\alpha \in [0,1]$ and $\beta \in (0,1)$, we have:*

$$\mathrm{KL}_{\alpha,\beta}(p : q) = \mathrm{KL}\left((pq)_\alpha : (pq)_\beta\right) \leq \log \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\}.$$

Proof. For $p(x), q(x) > 0$, we have

$$\frac{(1-\alpha)p(x) + \alpha q(x)}{(1-\beta)p(x) + \beta q(x)} \leq \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\}. \tag{51}$$

Indeed, by considering the two cases $\alpha \geq \beta$ (or equivalently, $1-\alpha \leq 1-\beta$) and $\alpha \leq \beta$ (or equivalently, $1-\alpha \geq 1-\beta$), we check that $(1-\alpha)p(x) \leq \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\}(1-\beta)p(x)$ and $\alpha q(x) \leq \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\}\beta q(x)$. Thus, we have $\frac{(1-\alpha)p(x)+\alpha q(x)}{(1-\beta)p(x)+\beta q(x)} \leq \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\}$. Therefore, it follows that:

$$\mathrm{KL}\left((pq)_\alpha : (pq)_\beta\right) \leq \int (pq)_\alpha \log \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\} \mathrm{d}\mu = \log \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\}. \tag{52}$$

Notice that we can interpret $\log \max\left\{\frac{1-\alpha}{1-\beta}, \frac{\alpha}{\beta}\right\} = \max\{\log \frac{1-\alpha}{1-\beta}, \log \frac{\alpha}{\beta}\}$ as the ∞-Rényi divergence [36,37] between the following two two-point distributions: $(\alpha, 1-\alpha)$ and $(\beta, 1-\beta)$. See Theorem 6 of [36].

A weaker upper bound is $\mathrm{KL}((pq)_\alpha : (pq)_\beta) \leq \log \frac{1}{\beta(1-\beta)}$. Indeed, let us form a partition of the sample space $\mathcal{X}$ into two dominance regions:

- $R_p := \{x \in \mathcal{X} \ : \ q(x) \leq p(x)\}$ and
- $R_q := \{x \in \mathcal{X} \ : \ q(x) > p(x)\}$.

We have $(pq)_\alpha(x) = (1-\alpha)p(x) + \alpha q(x) \leq p(x)$ for $x \in R_p$ and $(pq)_\alpha(x) \leq q(x)$ for $x \in R_q$. It follows that

$$\mathrm{KL}\left((pq)_\alpha : (pq)_\beta\right) \leq \int_{R_p} (pq)_\alpha(x) \log \frac{p(x)}{(1-\beta)p(x)} \mathrm{d}\mu(x) + \int_{R_q} (pq)_\alpha(x) \log \frac{q(x)}{\beta q(x)} \mathrm{d}\mu(x).$$

That is, $\mathrm{KL}((pq)_\alpha : (pq)_\beta) \leq -\log(1-\beta) - \log\beta = \log\frac{1}{\beta(1-\beta)}$. Notice that we allow $\alpha \in \{0,1\}$ but not β to take the extreme values (i.e., $\beta \in (0,1)$). □

In fact, it is known that for both $\alpha, \beta \in (0,1)$, $\mathrm{KL}\left((pq)_\alpha : (pq)_\beta\right)$ amount to compute a Bregman divergence for the Shannon negentropy generator, since $\{(pq)_\gamma \ : \ \gamma \in (0,1)\}$ defines a *mixture family* [38] of order 1 in information geometry. Hence, it is always finite, as Bregman divergences are always finite (but not necessarily bounded).

By using the fact that

$$\mathrm{JS}^{\alpha,w}(p:q) = \sum_{i=1}^{k} w_i \mathrm{KL}\left((pq)_{\alpha_i} : (pq)_{\bar{\alpha}}\right), \tag{53}$$

we conclude that the vector-skew Jensen–Shannon divergence is upper-bounded:

Lemma 2 (Bounded (w,α)-Jensen–Shannon divergence)**.** $\mathrm{JS}^{\alpha,w}$ *is bounded by* $\log\frac{1}{\bar{\alpha}(1-\bar{\alpha})}$ *where* $\bar{\alpha} = \sum_{i=1}^{k} w_i\alpha_i \in (0,1)$.

Proof. We have $\mathrm{JS}^{\alpha,w}(p:q) = \sum_i w_i \mathrm{KL}\left((pq)_{\alpha_i} : (pq)_{\bar{\alpha}}\right)$. Since $0 \leq \mathrm{KL}\left((pq)_{\alpha_i} : (pq)_{\bar{\alpha}}\right) \leq \log\frac{1}{\bar{\alpha}(1-\bar{\alpha})}$, it follows that we have

$$0 \leq \mathrm{JS}^{\alpha,w}(p:q) \leq \log\frac{1}{\bar{\alpha}(1-\bar{\alpha})}.$$

Notice that we also have

$$\mathrm{JS}^{\alpha,w}(p:q) \leq \sum_i w_i \log\max\left\{\frac{1-\alpha_i}{1-\bar{\alpha}}, \frac{\alpha_i}{\bar{\alpha}}\right\}.$$

□

The vector-skew Jensen–Shannon divergence is symmetric if and only if for each index $i \in [k]$ there exists a matching index $\sigma(i)$ such that $\alpha_{\sigma(i)} = 1-\alpha_i$ and $w_{\sigma(i)} = w_i$.

For example, we may define the *symmetric scalar α-skew Jensen–Shannon divergence* as

$$\begin{aligned}
\mathrm{JS}_s^\alpha(p,q) &= \frac{1}{2}\mathrm{KL}((pq)_\alpha : (pq)_{\frac{1}{2}}) + \frac{1}{2}\mathrm{KL}((pq)_{1-\alpha} : (pq)_{\frac{1}{2}}), && (54)\\
&= \frac{1}{2}\int (pq)_\alpha \log\frac{(pq)_\alpha}{(pq)_{\frac{1}{2}}}\mathrm{d}\mu + \frac{1}{2}\int (pq)_{1-\alpha}\log\frac{(pq)_{1-\alpha}}{(pq)_{\frac{1}{2}}}\mathrm{d}\mu, && (55)\\
&= \frac{1}{2}\int (qp)_{1-\alpha}\log\frac{(qp)_{1-\alpha}}{(qp)_{\frac{1}{2}}}\mathrm{d}\mu + +\frac{1}{2}\int (qp)_\alpha\log\frac{(qp)_\alpha}{(qp)_{\frac{1}{2}}}\mathrm{d}\mu, && (56)\\
&= h((pq)_{\frac{1}{2}}) - \frac{h((pq)_\alpha) + h((pq)_{1-\alpha})}{2}, && (57)\\
&=: \mathrm{JS}_s^\alpha(q,p), && (58)
\end{aligned}$$

since it holds that $(ab)_c = (ba)_{1-c}$ for any $a,b,c \in \mathbb{R}$. Note that $\mathrm{JS}_s^\alpha(p,q) \neq \mathrm{JS}^\alpha(p,q)$.

Remark 2. *We can always symmetrize a vector-skew Jensen–Shannon divergence by doubling the dimension of the skewing vector. Let* $\alpha = (\alpha_1, \ldots, \alpha_k)$ *and* w *be the vector parameters of an asymmetric vector-skew JSD, and consider* $\alpha' = (1-\alpha_1, \ldots, 1-\alpha_k)$ *and* w *to be the parameters of* $\mathrm{JS}^{\alpha',w}$*. Then,* $\mathrm{JS}^{(\alpha,\alpha'),(\frac{w}{2},\frac{w}{2})}$ *is a symmetric skew-vector JSD:*

$$\begin{aligned}
\mathrm{JS}^{(\alpha,\alpha'),(\frac{w}{2},\frac{w}{2})}(p:q) &:= \frac{1}{2}\mathrm{JS}^{\alpha,w}(p:q) + \frac{1}{2}\mathrm{JS}^{\alpha',w}(p:q), && (59)\\
&= \frac{1}{2}\mathrm{JS}^{\alpha,w}(p:q) + \frac{1}{2}\mathrm{JS}^{\alpha,w}(q:p) = \mathrm{JS}^{(\alpha,\alpha'),(\frac{w}{2},\frac{w}{2})}(q:p). && (60)
\end{aligned}$$

Since the vector-skew Jensen–Shannon divergence is an f-divergence for the generator $f_{\alpha,w}$ (Theorem 1), we can take generator $f^{s}_{w,\alpha}(u) = \frac{f_{w,\alpha}(u)+f^{*}_{w,\alpha}(u)}{2}$ to define the symmetrized f-divergence, where $f^{*}_{w,\alpha}(u) = uf_{w,\alpha}(\frac{1}{u})$ denotes the convex conjugate function. When $f_{\alpha,w}$ yields a symmetric f-divergence $I_{f_{\alpha,w}}$, we can apply the generic upper bound of f-divergences (i.e., $I_f \leq f(0) + f^{*}(0)$) to get the upper bound on the symmetric vector-skew Jensen–Shannon divergences:

$$\begin{aligned} I_{f_{\alpha,w}}(p:q) &\leq f_{\alpha,w}(0) + f^{*}_{\alpha,w}(0), && (61)\\ &\leq \sum_{i=1}^{k} w_i \left((1-\alpha_i) \log \frac{1-\alpha_i}{1-\bar{\alpha}} + \alpha_i \log \frac{\alpha_i}{\bar{\alpha}} \right), && (62) \end{aligned}$$

since

$$\begin{aligned} f^{*}_{\alpha,w}(u) &= u f_{\alpha,w}\left(\frac{1}{u}\right), && (63)\\ &= \sum_{i=1}^{k} w_i ((1-\alpha_i)u + \alpha_i) \log \frac{(1-\alpha_i)u+\alpha_i}{(1-\bar{\alpha})u+\bar{\alpha}}. && (64) \end{aligned}$$

For example, consider the ordinary Jensen–Shannon divergence with $w = \left(\frac{1}{2}, \frac{1}{2}\right)$ and $\alpha = (0, 1)$. Then, we find $\mathrm{JS}(p:w) = I_{f_{(0,1),\left(\frac{1}{2},\frac{1}{2}\right)}}(p:q) \leq \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = \log 2$, the usual upper bound of the JSD.

As a side note, let us notice that our notation $(pq)_\alpha$ allows one to compactly write the following property:

Property 1. *We have $q = (qq)_\lambda$ for any $\lambda \in [0,1]$, and $((p_1p_2)_\lambda(q_1q_2)_\lambda)_\alpha = ((p_1q_1)_\alpha(p_2q_2)_\alpha)_\lambda$ for any $\alpha, \lambda \in [0,1]$.*

Proof. Clearly, $q = (1-\lambda)q + \lambda q =: ((qq)_\lambda)$ for any $\lambda \in [0,1]$. Now, we have

$$\begin{aligned} ((p_1p_2)_\lambda(q_1q_2)_\lambda)_\alpha &= (1-\alpha)(p_1p_2)_\lambda + \alpha(q_1q_2)_\lambda, && (65)\\ &= (1-\alpha)((1-\lambda)p_1 + \lambda p_2) + \alpha((1-\lambda)q_1 + \lambda q_2), && (66)\\ &= (1-\lambda)((1-\alpha)p_1 + \alpha q_1) + \lambda((1-\alpha)p_2 + \alpha q_2), && (67)\\ &= (1-\lambda)(p_1q_1)_\alpha + \lambda(p_2q_2)_\alpha, && (68)\\ &= ((p_1q_1)_\alpha(p_2q_2)_\alpha)_\lambda. && (69) \end{aligned}$$

□

2.3. Building Symmetric Families of Vector-Skewed Jensen–Shannon Divergences

We can build infinitely many vector-skew Jensen–Shannon divergences. For example, consider $\alpha = \left(0, 1, \frac{1}{3}\right)$ and $w = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. Then, $\bar{\alpha} = \frac{1}{3} + \frac{1}{9} = \frac{4}{9}$, and

$$\mathrm{JS}^{\alpha,w}(p:q) = h\left((pq)_{\frac{4}{9}}\right) - \frac{h(p) + h(q) + h\left((pq)_{\frac{1}{3}}\right)}{3} \neq \mathrm{JS}^{\alpha,w}(q:p). \quad (70)$$

Interestingly, we can also build infinitely many families of *symmetric* vector-skew Jensen–Shannon divergences. For example, consider these two examples that illustrate the construction process:

- Consider $k = 2$. Let $(w, 1-w)$ denote the weight vector, and $\alpha = (\alpha_1, \alpha_2)$ the skewing vector. We have $\bar{\alpha} = w\alpha_1 + (1-w)\alpha_2 = \alpha_2 + w(\alpha_1 - \alpha_2)$. The vector-skew JSD is symmetric iff. $w =$

$1-w=\frac{1}{2}$ (with $\bar{\alpha}=\frac{\alpha_1+\alpha_2}{2}$) and $\alpha_2=1-\alpha_1$. In that case, we have $\bar{\alpha}=\frac{1}{2}$, and we obtain the following family of symmetric Jensen–Shannon divergences:

$$\mathrm{JS}^{(\alpha,1-\alpha),(\frac{1}{2},\frac{1}{2})}(p,q) = h\left((pq)_{\frac{1}{2}}\right)-\frac{h((pq)_\alpha)+h((pq)_{1-\alpha})}{2}, \quad (71)$$

$$= h\left((pq)_{\frac{1}{2}}\right)-\frac{h((pq)_\alpha)+h((qp)_\alpha)}{2}=\mathrm{JS}^{(\alpha,1-\alpha),(\frac{1}{2},\frac{1}{2})}(q,p). \quad (72)$$

- Consider $k=4$, weight vector $w=\left(\frac{1}{3},\frac{1}{3},\frac{1}{6},\frac{1}{6}\right)$, and skewing vector $\alpha=(\alpha_1,1-\alpha_1,\alpha_2,1-\alpha_2)$ for $\alpha_1,\alpha_2\in(0,1)$. Then, $\bar{\alpha}=\frac{1}{2}$, and we get the following family of symmetric vector-skew JSDs:

$$\mathrm{JS}^{(\alpha_1,\alpha_2)}(p,q) = h\left((pq)_{\frac{1}{2}}\right)-\frac{2h((pq)_{\alpha_1})+2h((pq)_{1-\alpha_1})+h((pq)_{\alpha_2})+h((pq)_{1-\alpha_2})}{6}, \quad (73)$$

$$= h\left((pq)_{\frac{1}{2}}\right)-\frac{2h((pq)_{\alpha_1})+2h((qp)_{\alpha_1})+h((pq)_{\alpha_2})+h((qp)_{\alpha_2})}{6}, \quad (74)$$

$$= \mathrm{JS}^{(\alpha_1,\alpha_2)}(q,p). \quad (75)$$

- We can similarly carry on the construction of such symmetric JSDs by increasing the dimensionality of the skewing vector.

In fact, we can define

$$\mathrm{JS}_s^{\alpha,w}(p,q):=h\left((pq)_{\frac{1}{2}}\right)-\sum_{i=1}^k w_i\frac{h((pq)_{\alpha_i})+h((pq)_{1-\alpha_i})}{2}=\sum_{i=1}^k w_i\mathrm{JS}_s^{\alpha_i}(p,q), \quad (76)$$

with

$$\mathrm{JS}_s^{\alpha}(p,q):=h\left((pq)_{\frac{1}{2}}\right)-\frac{h((pq)_\alpha)+h((pq)_{1-\alpha})}{2}. \quad (77)$$

3. Jensen–Shannon Centroids on Mixture Families

3.1. Mixture Families and Jensen–Shannon Divergences

Consider a mixture family in information geometry [25]. That is, let us give a prescribed set of $D+1$ linearly independent probability densities $p_0(x),\ldots,p_D(x)$ defined on the sample space $\mathcal{X}$. A *mixture family* $\mathcal{M}$ of order D consists of all *strictly* convex combinations of these component densities:

$$\mathcal{M}:=\left\{m(x;\theta):=\sum_{i=1}^D\theta^i p_i(x)+\left(1-\sum_{i=1}^D\theta^i\right)p_0(x)\ :\ \theta^i>0,\ \sum_{i=1}^D\theta^i<1\right\}. \quad (78)$$

For example, the family of categorical distributions (sometimes called "multinouilli" distributions) is a mixture family [25]:

$$\mathcal{M}=\left\{m_\theta(x)=\sum_{i=1}^D\theta_i\delta(x-x_i)+\left(1-\sum_{i=1}^D\theta_i\right)\delta(x-x_0)\right\}, \quad (79)$$

where $\delta(x)$ is the Dirac distribution (i.e., $\delta(x)=1$ for $x=0$ and $\delta(x)=0$ for $x\neq 0$). Note that the mixture family of categorical distributions can also be interpreted as an exponential family.

Notice that the linearly independent assumption on probability densities is to ensure to have an identifiable model: $\theta\leftrightarrow m(x;\theta)$.

The KL divergence between two densities of a mixture family $\mathcal{M}$ amounts to a Bregman divergence for the Shannon negentropy generator $F(\theta)=-h(m_\theta)$ (see [38]):

$$\mathrm{KL}(m_{\theta_1}:m_{\theta_2})=B_F(\theta_1:\theta_2)=B_{-h(m_\theta)}(\theta_1:\theta_2). \quad (80)$$

On a mixture manifold $\mathcal{M}$, the mixture density $(1-\alpha)m_{\theta_1} + \alpha m_{\theta_2}$ of two mixtures m_{θ_1} and m_{θ_2} of $\mathcal{M}$ also belongs to $\mathcal{M}$:

$$(1-\alpha)m_{\theta_1} + \alpha m_{\theta_2} = m_{(\theta_1\theta_2)_\alpha} \in \mathcal{M}, \tag{81}$$

where we extend the notation $(\theta_1\theta_2)_\alpha := (1-\alpha)\theta_1 + \alpha\theta_2$ to vectors θ_1 and θ_2: $(\theta_1\theta_2)^i_\alpha = (\theta_1^i\theta_2^i)_\alpha$.

Thus, the vector-skew JSD amounts to a vector-skew Jensen diversity for the Shannon negentropy convex function $F(\theta) = -h(m_\theta)$:

$$\begin{aligned}
\mathrm{JS}^{\alpha,w}(m_{\theta_1} : m_{\theta_2}) &= \sum_{i=1}^k w_i \mathrm{KL}\left((m_{\theta_1}m_{\theta_2})_{\alpha_i} : (m_{\theta_1}m_{\theta_2})_{\bar{\alpha}}\right), && (82)\\
&= \sum_{i=1}^k w_i \mathrm{KL}\left(m_{(\theta_1\theta_2)_{\alpha_i}} : m_{(\theta_1\theta_2)_{\bar{\alpha}}}\right), && (83)\\
&= \sum_{i=1}^k w_i B_F\left((\theta_1\theta_2)_{\alpha_i} : (\theta_1\theta_2)_{\bar{\alpha}}\right), && (84)\\
&= \mathrm{JB}_F^{\alpha,\bar{\alpha},w}(\theta_1 : \theta_2), && (85)\\
&= \sum_{i=1}^k w_i F\left((\theta_1\theta_2)_{\alpha_i}\right) - F\left((\theta_1\theta_2)_{\bar{\alpha}}\right), && (86)\\
&= h(m_{(\theta_1\theta_2)_{\bar{\alpha}}}) - \sum_{i=1}^k w_i h\left(m_{(\theta_1\theta_2)_{\alpha_i}}\right). && (87)
\end{aligned}$$

3.2. Jensen–Shannon Centroids

Given a set of n mixture densities $m_{\theta_1}, \ldots, m_{\theta_n}$ of $\mathcal{M}$, we seek to calculate the *skew-vector Jensen–Shannon centroid* (or barycenter for non-uniform weights) defined as m_{θ^*}, where θ^* is the minimizer of the following objective function (or loss function):

$$L(\theta) := \sum_{j=1}^n \omega_j \mathrm{JS}^{\alpha,w}(m_{\theta_k} : m_\theta), \tag{88}$$

where $\omega \in \Delta_n$ is the weight vector of densities (uniform weight for the centroid and non-uniform weight for a barycenter). This definition of the skew-vector Jensen–Shannon centroid is a generalization of the *Fréchet mean* (the Fréchet mean may not be unique, as it is the case on the sphere for two antipodal points for which their Fréchet means with respect to the geodesic metric distance form a great circle) [39] to non-metric spaces. Since the divergence $\mathrm{JS}^{\alpha,w}$ is strictly separable convex, it follows that the Jensen–Shannon-type centroids are unique when they exist.

Plugging Equation (86) into Equation (88), we get that the calculation of the Jensen–Shannon centroid amounts to the following minimization problem:

$$L(\theta) = \sum_{j=1}^n \omega_j \left(\sum_{i=1}^k w_i F((\theta_j\theta)_{\alpha_i}) - F\left((\theta_j\theta)_{\bar{\alpha}}\right)\right). \tag{89}$$

This optimization is a *Difference of Convex* (DC) programming optimization, for which we can use the ConCave–Convex procedure [27,40] (CCCP). Indeed, let us define the following two convex functions:

$$\begin{aligned}
A(\theta) &= \sum_{j=1}^n \sum_{i=1}^k \omega_j w_i F((\theta_j\theta)_{\alpha_i}), && (90)\\
B(\theta) &= \sum_{j=1}^n \omega_j F\left((\theta_j\theta)_{\bar{\alpha}}\right). && (91)
\end{aligned}$$

Both functions $A(\theta)$ and $B(\theta)$ are convex since F is convex. Then, the minimization problem of Equation (89) to solve can be rewritten as:

$$\min_{\theta} A(\theta) - B(\theta). \tag{92}$$

This is a DC programming optimization problem which can be solved iteratively by initializing θ to an arbitrary value $\theta^{(0)}$ (say, the centroid of the θ_is), and then by updating the parameter at step t using the CCCP [27] as follows:

$$\theta^{(t+1)} = (\nabla B)^{-1}(\nabla A(\theta^{(t)})). \tag{93}$$

Compared to a gradient descent local optimization, there is no required step size (also called "learning" rate) in CCCP.

We have $\nabla A(\theta) = \sum_{j=1}^{n} \sum_{i=1}^{k} \omega_j w_i \alpha_i \nabla F((\theta_j \theta)_{\alpha_i})$ and $\nabla B(\theta) = \sum_{j=1}^{n} \omega_j \bar{\alpha} \nabla F\left((\theta_j \theta)_{\bar{\alpha}}\right)$.

The CCCP converges to a local optimum θ^* where the support hyperplanes of the function graphs of A and B at θ^* are parallel to each other, as depicted in Figure 1. The set of stationary points is $\{\theta : \nabla A(\theta) = \nabla B(\theta)\}$. In practice, the delicate step is to invert ∇B. Next, we show how to implement this algorithm for the Jensen–Shannon centroid of a set of categorical distributions (i.e., normalized histograms with all non-empty bins).

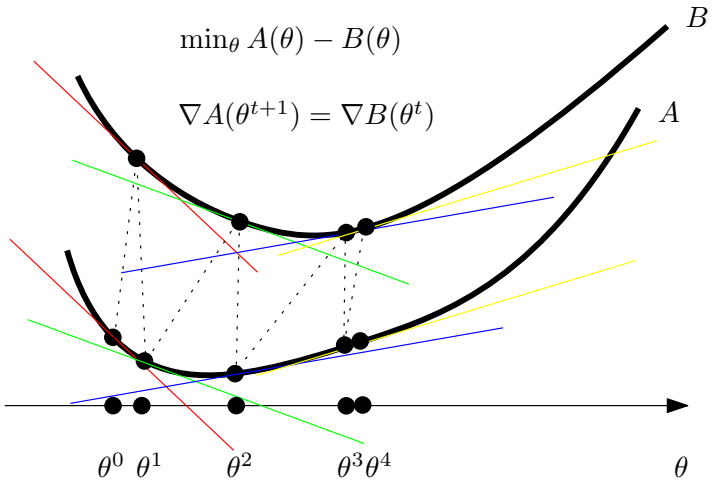

Figure 1. The Convex–ConCave Procedure (CCCP) iteratively updates the parameter θ by aligning the support hyperplanes at θ. In the limit case of convergence to θ^*, the support hyperplanes at θ^* are parallel to each other. CCCP finds a local minimum.

3.2.1. Jensen–Shannon Centroids of Categorical Distributions

To illustrate the method, let us consider the mixture family of categorical distributions [25]:

$$\mathcal{M} = \left\{ m_\theta(x) = \sum_{i=1}^{D} \theta_i \delta(x - x_i) + \left(1 - \sum_{i=1}^{D} \theta_i\right) \delta(x - x_0) \right\}. \tag{94}$$

The Shannon negentropy is

$$F(\theta) = -h(m_\theta) = \sum_{i=1}^{D} \theta_i \log \theta_i + \left(1 - \sum_{i=1}^{D} \theta_i\right) \log \left(1 - \sum_{i=1}^{D} \theta_i\right). \tag{95}$$

We have the partial derivatives

$$\nabla F(\theta) = \left[\frac{\partial}{\partial \theta_i}\right]_i, \quad \frac{\partial}{\partial \theta_i} F(\theta) = \log \frac{\theta_i}{1 - \sum_{j=1}^{D} \theta_j}. \tag{96}$$

Inverting the gradient ∇F requires us to solve the equation $\nabla F(\theta) = \eta$ so that we get $\theta = (\nabla F)^{-1}(\eta)$. We find that

$$\nabla F^*(\eta) = (\nabla F)^{-1}(\eta) = \frac{1}{1+\sum_{j=1}^{D}\exp(\eta_j)}[\exp(\eta_i)]_i, \quad \theta_i = (\nabla F^{-1}(\eta))_i = \frac{\exp(\eta_i)}{1+\sum_{j=1}^{D}\exp(\eta_j)}, \ \forall i \in [D]. \quad (97)$$

Table 1 summarizes the dual view of the family of categorical distributions, either interpreted as an exponential family or as a mixture family.

We have $\mathrm{JS}(p_1, p_2) = J_F(\theta_1, \theta_2)$ for $p_1 = m_{\theta_1}$ and $p_2 = m_{\theta_2}$, where

$$J_F(\theta_1 : \theta_2) = \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right), \quad (98)$$

is the Jensen divergence [40]. Thus, to compute the Jensen–Shannon centroid of a set of n densities $p_1, \ldots, p_n$ of a mixture family (with $p_i = m_{\theta_i}$), we need to solve the following optimization problem for a density $p = m_\theta$:

$$\min_p \sum_i \mathrm{JS}(p_i, p), \quad (99)$$

$$\min_\theta \sum_i J_F(\theta_i, \theta), \quad (100)$$

$$\min_\theta \sum_i \frac{F(\theta_i) + F(\theta)}{2} - F\left(\frac{\theta_i + \theta}{2}\right), \quad (101)$$

$$\equiv \min_\theta \frac{1}{2}F(\theta) - \frac{1}{n}\sum_i F\left(\frac{\theta_i + \theta}{2}\right) := E(\theta). \quad (102)$$

The CCCP algorithm for the Jensen–Shannon centroid proceeds by initializing $\theta^{(0)} = \frac{1}{n}\sum_i \theta_i$ (center of mass of the natural parameters), and iteratively updates as follows:

$$\theta^{(t+1)} = (\nabla F)^{-1}\left(\frac{1}{n}\sum_i \nabla F\left(\frac{\theta_i + \theta^{(t)}}{2}\right)\right). \quad (103)$$

We iterate until the absolute difference $|E(\theta^{(t)}) - E(\theta^{(t+1)})|$ between two successive $\theta^{(t)}$ and $\theta^{(t+1)}$ goes below a prescribed threshold value. The convergence of the CCCP algorithm is linear [41] to a local minimum that is a fixed point of the equation

$$\theta = M_H\left(\frac{\theta_1 + \theta}{2}, \ldots, \frac{\theta_n + \theta}{2}\right), \quad (104)$$

where $M_H(v_1, \ldots, v_n) := H^{-1}(\sum_{i=1}^{n} H(v_i))$ is a vector generalization of the formula of the quasi-arithmetic means [30,40] obtained for the generator $H = \nabla F$. Algorithm 1 summarizes the method for approximating the Jensen–Shannon centroid of a given set of categorical distributions (given a prescribed number of iterations). In the pseudo-code, we used the notation ${}^{(t+1)}\theta$ instead of $\theta^{(t+1)}$ in order to highlight the conversion procedures of the natural parameters to/from the mixture weight parameters by using superscript notations for coordinates.

Table 1. Two views of the family of categorical distributions with d choices: An exponential family or a mixture family of order $D = d - 1$. Note that the Bregman divergence associated to the exponential family view corresponds to the reverse Kullback–Leibler (KL) divergence, while the Bregman divergence associated to the mixture family view corresponds to the KL divergence.

	Exponential Family	**Mixture Family**
pdf	$p_\theta(x) = \prod_{i=1}^{d} p_i^{t_i(x)}, p_i = \Pr(x = e_i), t_i(x) \in \{0,1\}, \sum_{i=1}^{d} t_i(x) = 1$	$m_\theta(x) = \sum_{i=1}^{d} p_i \delta_{e_i}(x)$
primal θ	$\theta_i = \log \frac{p_i}{p_d}$	$\theta_i = p_i$
$F(\theta)$	$\log(1 + \sum_{i=1}^{D} \exp(\theta_i))$	$\theta_i \log \theta_i + (1 - \sum_{i=1}^{D} \theta_i) \log(1 - \sum_{i=1}^{D} \theta_i)$
dual $\eta = \nabla F(\theta)$	$\frac{e^{\theta_i}}{1+\sum_{j=1}^{D} \exp(\theta_j)}$	$\log \frac{\theta_i}{1-\sum_{j=1}^{D} \theta_j}$
primal $\theta = \nabla F^*(\eta)$	$\log \frac{\eta_i}{1-\sum_{j=1}^{D} \eta_j}$	$\frac{e^{\theta_i}}{1+\sum_{j=1}^{D} \exp(\theta_j)}$
$F^*(\eta)$	$\sum_{i=1}^{D} \eta_i \log \eta_i + (1 - \sum_{j=1}^{D} \eta_j) \log(1 - \sum_{j=1}^{D} \eta_j)$	$\log(1 + \sum_{i=1}^{D} \exp(\eta_i))$
Bregman divergence	$B_F(\theta : \theta') = \mathrm{KL}^*(p_\theta : p_{\theta'})$ $= \mathrm{KL}(p_{\theta'} : p_\theta)$	$B_F(\theta : \theta') = \mathrm{KL}(m_\theta : m_{\theta'})$

Algorithm 1: The CCCP algorithm for computing the Jensen–Shannon centroid of a set of categorical distributions.

Input: A set $\{p_i = (p_i^1, \ldots, p_i^d)\}_{i \in [n]}$ of n categorical distributions belonging to the $(d-1)$-dimensional probability simplex Δ_{d-1}

Input: T: The number of CCCP iterations

Output: An approximation ${}^{(T)}\bar{p}$ of the Jensen–Shannon centroid $\bar{p}$

`/* Convert the categorical distributions to their natural parameters by dropping the last coordinate */`

$\theta_i^j = p_i^j$ for $j \in \{1, \ldots, d-1\}$; `/* Initialize the JS centroid */`

$t \leftarrow 0$; ${}^{(0)}\bar{\theta} = \frac{1}{n} \sum_{i=1} \theta_i$; `/* Convert the initial natural parameter of the JS centroid to a categorical distribution */`

${}^{(0)}\bar{p}^j = {}^{(0)}\bar{\theta}^j$ for $j \in \{1, \ldots, d-1\}$; ${}^{(0)}\bar{p}^d = 1 - \sum_{i=1}^{d} {}^{(0)}\bar{p}^j$;

`/* Perform the ConCave-Convex Procedure (CCCP) */`

while $t \leq T$ **do**

  `/* Use Equation (96) for` ∇F `and Equation (97) for` $\nabla F^* = (\nabla F)^{-1}$ `*/`

  ${}^{(t+1)}\theta = (\nabla F)^{-1} \left(\frac{1}{n} \sum_i \nabla F \left(\frac{\theta_i + {}^{(t)}\theta}{2} \right) \right)$; $t \leftarrow t + 1$;

end

${}^{(T)}\bar{p}^j = {}^{(T)}\bar{\theta}^j$ for $j \in \{1, \ldots, d-1\}$; ${}^{(T)}\bar{p}^d = 1 - \sum_{i=1}^{d} {}^{(T)}\bar{p}^j$; **return** ${}^{(T)}\bar{p}$;

Figure 2 displays the results of the calculations of the Jeffreys centroid [18] and the Jensen–Shannon centroid for two normalized histograms obtained from grey-valued images of `Lena` and `Barbara`. Figure 3 show the Jeffreys centroid and the Jensen–Shannon centroid for the `Barbara` image and its negative image. Figure 4 demonstrates that the Jensen–Shannon centroid is well defined even if the input histograms do not have coinciding supports. Notice that on the parts of the support where only one distribution is defined, the JS centroid is a scaled copy of that defined distribution.

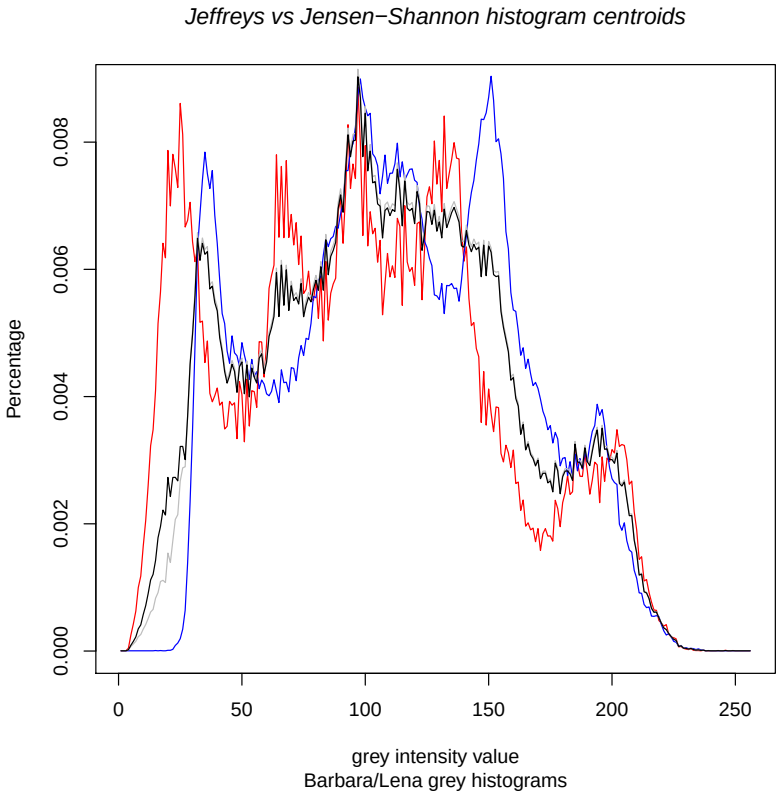

Figure 2. The Jeffreys centroid (grey histogram) and the Jensen–Shannon centroid (black histogram) for two grey normalized histograms of the `Lena` image (red histogram) and the `Barbara` image (blue histogram). Although these Jeffreys and Jensen–Shannon centroids look quite similar, observe that there is a major difference between them in the range $[0, 20]$ where the blue histogram is zero.

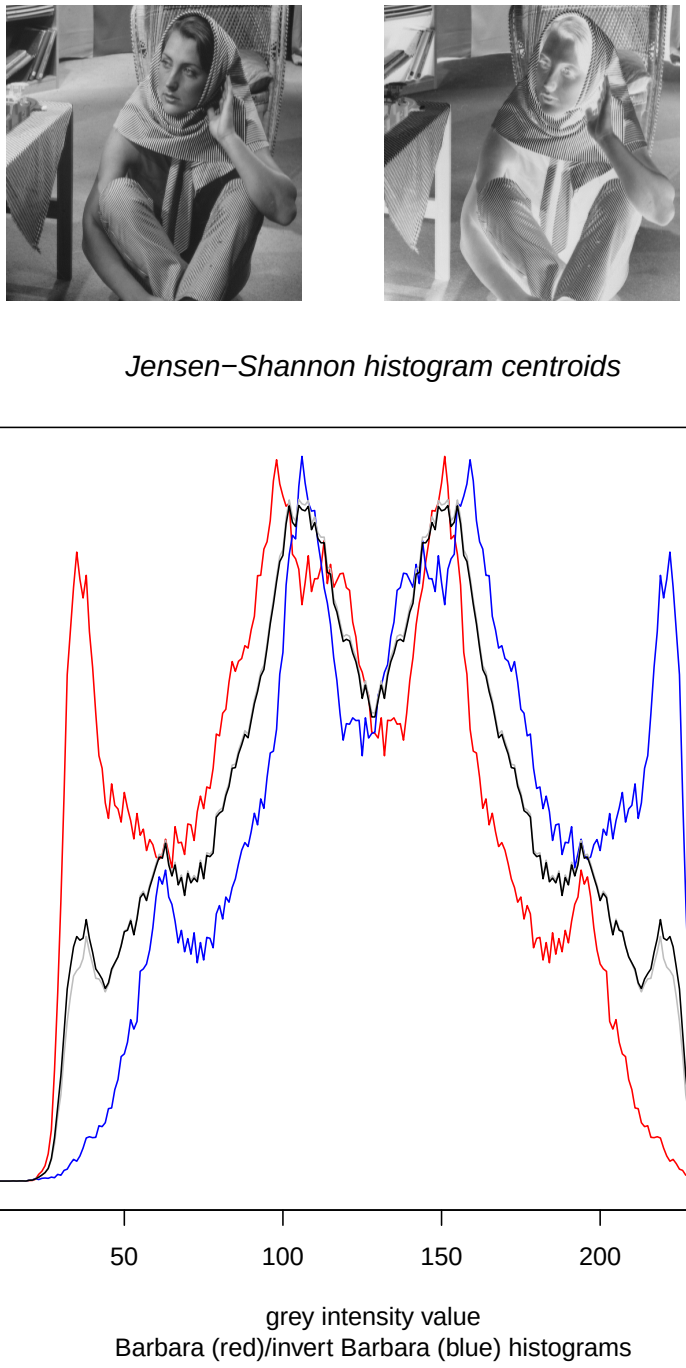

Figure 3. The Jeffreys centroid (grey histogram) and the Jensen–Shannon centroid (black histogram) for the grey normalized histogram of the `Barbara` image (red histogram) and its negative image (blue histogram which corresponds to the reflection around the vertical axis $x = 128$ of the red histogram).

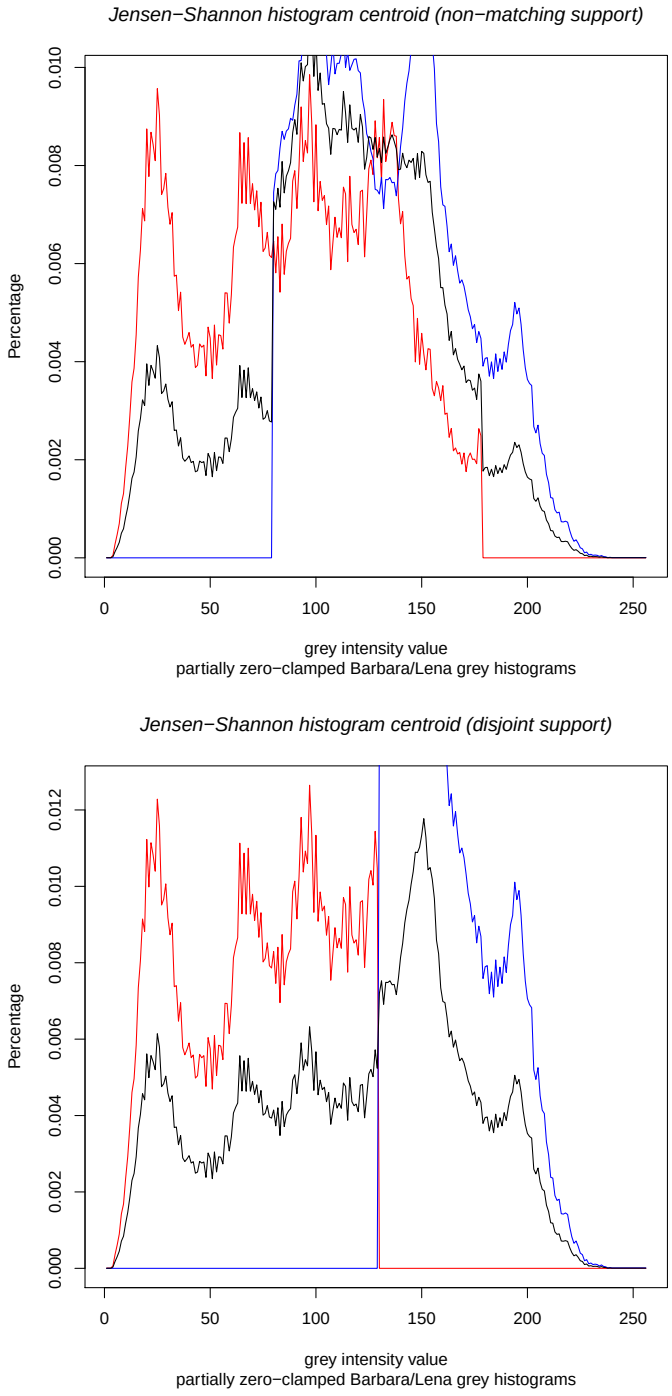

Figure 4. Jensen–Shannon centroid (black histogram) for the clamped grey normalized histogram of the `Lena` image (red histograms) and the clamped gray normalized histogram of `Barbara` image (blue histograms). Notice that on the part of the sample space where only one distribution is non-zero, the JS centroid scales that histogram portion.

3.2.2. Special Cases

Let us now consider two special cases:

- For the special case of $D = 1$, the categorical family is the Bernoulli family, and we have $F(\theta) = \theta \log \theta + (1-\theta)\log(1-\theta)$ (binary negentropy), $F'(\theta) = \log \frac{\theta}{1-\theta}$ (and $F''(\theta) = \frac{1}{\theta(1-\theta)} > 0$) and

$(F')^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$. The CCCP update rule to compute the binary Jensen–Shannon centroid becomes

$$\theta^{(t+1)} = (F')^{-1}\left(\sum_i w_i F'\left(\frac{\theta^{(t)}+\theta_i}{2}\right)\right). \tag{105}$$

- Since the skew-vector Jensen–Shannon divergence formula holds for positive densities:

$$\begin{aligned}
\mathrm{JS}^{+\alpha,w}(\tilde{p}:\tilde{q}) &= \sum_{i=1}^{k} w_i \mathrm{KL}^{+}((\tilde{p}\tilde{q})_{\alpha_i}:((\tilde{p}\tilde{q})_{\bar{\alpha}}), && (106)\\
&= \sum_{i=1}^{k} w_i \left(\mathrm{KL}((\tilde{p}\tilde{q})_{\alpha_i}:((\tilde{p}\tilde{q})_{\bar{\alpha}}) + \int (\tilde{p}\tilde{q})_{\bar{\alpha}}\mathrm{d}\mu - \underbrace{\sum_{i=1}^{k} w_i \int (\tilde{p}\tilde{q})_{\alpha_i}\mathrm{d}\mu}_{=\int (\tilde{p}\tilde{q})_{\bar{\alpha}}\mathrm{d}\mu}\right), && (107)\\
&= \mathrm{JS}^{\alpha,w}(\tilde{p}:\tilde{q}), && (108)
\end{aligned}$$

we can *relax* the computation of the Jensen–Shannon centroid by considering 1D separable minimization problems. We then normalize the positive JS centroids to get an approximation of the probability JS centroids. This approach was also considered when dealing with the Jeffreys' centroid [18]. In 1D, we have $F(\theta) = \theta\log\theta - \theta$, $F'(\theta) = \log\theta$ and $(F')^{-1}(\eta) = e^{\eta}$.

In general, calculating the negentropy for a mixture family with continuous densities sharing the same support is not tractable because of the log-sum term of the differential entropy. However, the following remark emphasizes an extension of the mixture family of categorical distributions:

3.2.3. Some Remarks and Properties

Remark 3. *Consider a mixture family* $m(\theta) = \sum_{i=1}^{D}\theta_i p_i(x) + \left(1-\sum_{i=1}^{D}\theta_i\right)p_0(x)$ *(for a parameter* θ *belonging to the D-dimensional standard simplex) of probability densities* $p_0(x),\ldots,p_D(x)$ *defined respectively on the supports* $\mathcal{X}_0,\mathcal{X}_1,\ldots,\mathcal{X}_D$*. Let* $\theta_0 := 1-\sum_{i=1}^{D}\theta_i$*. Assume that the support* $\mathcal{X}_i$*s of the* p_i*s are* mutually non-intersecting *(*$\mathcal{X}_i\cap\mathcal{X}_j = \emptyset$ *for all* $i\neq j$ *implying that the* $D+1$ *densities are linearly independent) so that* $m_\theta(x) = \theta_i p_i(x)$ *for all* $x\in\mathcal{X}_i$*, and let* $\mathcal{X} = \cup_i\mathcal{X}_i$*. Consider Shannon negative entropy* $F(\theta) = -h(m_\theta)$ *as a strictly convex function. Then, we have*

$$\begin{aligned}
F(\theta) &= -h(m_\theta) = \int_{\mathcal{X}} m_\theta(x)\log m_\theta(x), && (109)\\
&= \sum_{i=0}^{D}\theta_i\int_{\mathcal{X}_i} p_i(x)\log(\theta_i p_i(x))\mathrm{d}\mu(x), && (110)\\
&= \sum_{i=0}^{D}\theta_i\log\theta_i - \sum_{i=0}^{D}\theta_i h(p_i). && (111)
\end{aligned}$$

Note that the term $\sum_i\theta_i h(p_i)$ *is affine in* θ*, and Bregman divergences are defined up to affine terms so that the Bregman generator F is equivalent to the Bregman generator of the family of categorical distributions. This example generalizes the ordinary mixture family of categorical distributions where the* p_i*s are distinct Dirac distributions. Note that when the support of the component distributions are not pairwise disjoint, the (neg)entropy may not be analytic [42] (e.g., mixture of the convex weighting of two prescribed distinct Gaussian distributions). This contrasts with the fact that the cumulant function of an exponential family is always real-analytic [43]. Observe that the term* $\sum_i\theta_i h(p_i)$ *can be interpreted as a conditional entropy:* $\sum_i\theta_i h(p_i) = h(X|\Theta)$ *where* $\Pr(\Theta=i) = \theta_i$ *and* $\Pr(X\in S|\Theta=i) = \int_S p_i(x)\mathrm{d}\mu(x)$.

Notice that we can truncate an exponential family [25] to get a (potentially non-regular [44]) exponential family for defining the p_is on mutually non-intersecting domains $\mathcal{X}_i$s. The entropy of a natural exponential family $\{e(x:\theta) = \exp(x^\top\theta - \psi(\theta)) \;:\; \theta \in \Theta\}$ with cumulant function $\psi(\theta)$ and natural parameter space Θ is $-\psi^*(\eta)$, where $\eta = \nabla\psi(\theta)$, and ψ^* is the Legendre convex conjugate [45]: $h(e(x:\theta)) = -\psi^*(\nabla\psi(\theta))$.

In general, the entropy and cross-entropy between densities of a mixture family (whether the distributions have disjoint supports or not) can be calculated in closed-form.

Property 2. *The entropy of a density belonging to a mixture family $\mathcal{M}$ is $h(m_\theta) = -F(\theta)$, and the cross-entropy between two mixture densities m_{θ_1} and m_{θ_2} is $h^\times(m_{\theta_1} : m_{\theta_2}) = -F(\theta_2) - (\theta_1 - \theta_2)^\top\eta_2 = F^*(\eta_2) - \theta_1^\top\eta_2$.*

Proof. Let us write the KLD as the difference between the cross-entropy minus the entropy [4]:

$$\begin{aligned} \mathrm{KL}(m_{\theta_1} : m_{\theta_2}) &= h^\times(m_{\theta_1} : m_{\theta_2}) - h(m_{\theta_1}), && (112)\\ &= B_F(\theta_1 : \theta_2), && (113)\\ &= F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top\nabla F(\theta_2). && (114) \end{aligned}$$

Following [45], we deduce that $h(m_\theta) = -F(\theta) + c$ and $h^\times(m_{\theta_1} : m_{\theta_2}) = -F(\theta_2) - (\theta_1 - \theta_2)^\top\eta_2 - c$ for a constant c. Since $F(\theta) = -h(m_\theta)$ by definition, it follows that $c = 0$ and that $h^\times(m_{\theta_1} : m_{\theta_2}) = -F(\theta_2) - (\theta_1 - \theta_2)^\top\eta_2 = F^*(\eta_2) - \theta_1^\top\eta_2$ where $\eta = \nabla F(\theta)$. □

Thus, we can numerically compute the Jensen–Shannon centroids (or barycenters) of a set of densities belonging to a mixture family. This includes the case of categorical distributions and the case of Gaussian Mixture Models (GMMs) with prescribed Gaussian components [38] (although in this case, the negentropy needs to be stochastically approximated using Monte Carlo techniques [46]). When the densities do not belong to a mixture family (say, the Gaussian family, which is an exponential family [25]), we face the problem that the mixture of two densities does not belong to the family anymore. One way to tackle this problem is to project the mixture onto the Gaussian family. This corresponds to an m-projection (mixture projection) which can be interpreted as a Maximum Entropy projection of the mixture [25,47]).

Notice that we can perform fast k-means clustering without centroid calculations using a generalization of the k-means++ probabilistic initialization [48,49]. See [50] for details of the generalized k-means++ probabilistic initialization defined according to an arbitrary divergence.

Finally, let us notice some decompositions of the Jensen–Shannon divergence and the skew Jensen divergences.

Remark 4. *We have the following decomposition for the Jensen–Shannon divergence:*

$$\begin{aligned} \mathrm{JS}(p_1, p_2) &= h\left(\frac{p_1 + p_2}{2}\right) - \frac{h(p_1) + h(p_2)}{2}, && (115)\\ &= h^\times_{\mathrm{JS}}(p_1 : p_2) - h_{\mathrm{JS}}(p_2) \geq 0, && (116) \end{aligned}$$

where

$$h^\times_{\mathrm{JS}}(p_1 : p_2) = h\left(\frac{p_1 + p_2}{2}\right) - \frac{1}{2}h(p_1), \quad (117)$$

and $h_{\mathrm{JS}}(p_2) = h^\times_{\mathrm{JS}}(p_2 : p_2) = h(p_2) - \frac{1}{2}h(p_2) = \frac{1}{2}h(p_2)$. This decomposition bears some similarity with the KLD decomposition viewed as the cross-entropy minus the entropy (with the cross-entropy always upper-bounding the entropy).

Similarly, the α-skew Jensen divergence

$$J_F^\alpha(\theta_1 : \theta_2) := (F(\theta_1)F(\theta_2))_\alpha - F\left((\theta_1\theta_2)_\alpha\right), \quad \alpha \in (0,1) \tag{118}$$

can be decomposed as the sum of the information $I_F^\alpha(\theta_1) = (1-\alpha)F(\theta_1)$ *minus the cross-information* $C_F^\alpha(\theta_1 : \theta_2) := F\left((\theta_1\theta_2)_\alpha\right) - \alpha F(\theta_2)$*:*

$$J_F^\alpha(\theta_1 : \theta_2) = I_F^\alpha(\theta_1) - C_F^\alpha(\theta_1 : \theta_2) \geq 0. \tag{119}$$

Notice that the information $I_F^\alpha(\theta_1)$ *is the self cross-information:* $I_F^\alpha(\theta_1) = C_F^\alpha(\theta_1 : \theta_1) = (1-\alpha)F(\theta_1)$*. Recall that the convex information is the negentropy where the entropy is concave. For the Jensen–Shannon divergence on the mixture family of categorical distributions, the convex generator* $F(\theta) = -h(m_\theta) = \sum_{i=1}^D \theta^i \log \theta^i$ *is the Shannon negentropy.*

Finally, let us briefly mention the *Jensen–Shannon diversity* [30] which extends the Jensen–Shannon divergence to a weighted set of densities as follows:

$$\mathrm{JS}(p_1, \ldots, p_k; w_1, \ldots, w_k) := \sum_{i=1}^k w_i \mathrm{KL}(p_i : \bar{p}), \tag{120}$$

where $\bar{p} = \sum_{i=1}^k w_i p_i$. The Jensen–Shannon diversity plays the role of the variance of a cluster with respect to the KLD. Indeed, let us state the compensation identity [51]: For any q, we have

$$\sum_{i=1}^k w_i \mathrm{KL}(p_i : q) = \sum_{i=1}^k w_i \mathrm{KL}(p_i : \bar{p}) + \mathrm{KL}(\bar{p} : q). \tag{121}$$

Thus, the cluster center defined as the minimizer of $\sum_{i=1}^k w_i \mathrm{KL}(p_i : q)$ is the centroid $\bar{p}$, and

$$\sum_{i=1}^k w_i \mathrm{KL}(p_i : \bar{p}) = \mathrm{JS}(p_1, \ldots, p_k; w_1, \ldots, w_k). \tag{122}$$

4. Conclusions and Discussion

The Jensen–Shannon divergence [6] is a renown symmetrization of the Kullback–Leibler oriented divergence that enjoys the following three essential properties:

1. It is always bounded,
2. it applies to densities with potentially different supports, and
3. it extends to unnormalized densities while enjoying the same formula expression.

This JSD plays an important role in machine learning and in deep learning for studying Generative Adversarial Networks (GANs) [52]. Traditionally, the JSD has been skewed with a scalar parameter [19,53] $\alpha \in (0,1)$. In practice, it has been experimentally demonstrated that skewing divergences may significantly improve the performance of some tasks (e.g., [21,54]).

In general, we can symmetrize the KLD $\mathrm{KL}(p : q)$ by taking an *abstract mean* (we require a symmetric mean $M(x,y) = M(y,x)$ with the in-betweenness property: $\min\{x,y\} \leq M(x,y) \leq \max\{x,y\}$) M between the two orientations $\mathrm{KL}(p : q)$ and $\mathrm{KL}(q : p)$:

$$\mathrm{KL}_M(p,q) := M(\mathrm{KL}(p : q), \mathrm{KL}(q : p)). \tag{123}$$

We recover the Jeffreys divergence by taking the arithmetic mean twice (i.e., $J(p,q) = 2A(\mathrm{KL}(p : q), \mathrm{KL}(q : p))$ where $A(x,y) = \frac{x+y}{2}$), and the resistor average divergence [55] by taking the harmonic

mean (i.e., $R_{\mathrm{KL}}(p,q) = H(\mathrm{KL}(p:q),\mathrm{KL}(q:p)) = \frac{2\mathrm{KL}(p:q)\mathrm{KL}(q:p)}{\mathrm{KL}(p:q)+\mathrm{KL}(q:p)}$ where $H(x,y) = \frac{2}{\frac{1}{x}+\frac{1}{y}}$). When we take the limit of Hölder power means, we get the following extremal symmetrizations of the KLD:

$$\mathrm{KL}^{\min}(p:q) = \min\{\mathrm{KL}(p:q),\mathrm{KL}(q:p)\} = \mathrm{KL}^{\min}(q:p), \quad (124)$$

$$\mathrm{KL}^{\max}(p:q) = \max\{\mathrm{KL}(p:q),\mathrm{KL}(q:p)\} = \mathrm{KL}^{\max}(q:p). \quad (125)$$

In this work, we showed how to *vector-skew* the JSD while preserving the above three properties. These new families of *weighted vector-skew Jensen–Shannon divergences* may allow one to fine-tune the dissimilarity in applications by replacing the skewing scalar parameter of the JSD by a vector parameter (informally, adding some "knobs" for tuning a divergence). We then considered computing the Jensen–Shannon centroids of a set of densities belonging to a mixture family [25] by using the convex–concave procedure [27].

In general, we can vector-skew any arbitrary divergence D by using two k-dimensional vectors $\alpha \in [0,1]^k$ and $\beta \in [0,1]^k$ (with $\alpha \neq \beta$) by building a weighted separable divergence as follows:

$$D^{\alpha,\beta,w}(p:q) := \sum_{i=1}^{k} w_i D\left((pq)_{\alpha_i} : (pq)_{\beta_i}\right) = D^{1_k-\alpha,1_k-\beta,w}(q:p), \quad \alpha \neq \beta. \quad (126)$$

This bi-vector-skew divergence unifies the Jeffreys divergence with the Jensen–Shannon α-skew divergence by setting the following parameters:

$$\mathrm{KL}^{(0,1),(1,0),(1,1)}(p:q) = \mathrm{KL}(p:q)+\mathrm{KL}(q:p) = J(p,q), \quad (127)$$

$$\mathrm{KL}^{(0,\alpha),(1,1-\alpha),(\frac{1}{2},\frac{1}{2})}(p:q) = \frac{1}{2}\mathrm{KL}(p:(pq)_\alpha)+\frac{1}{2}\mathrm{KL}(q:(pq)_\alpha). \quad (128)$$

We have shown in this paper that interesting properties may occur when the skewing vector β is purposely correlated to the skewing vector α: Namely, for the bi-vector-skew Bregman divergences with $\beta = (\bar{\alpha},\ldots,\bar{\alpha})$ and $\bar{\alpha} = \sum_i w_i\alpha_i$, we obtain an equivalent Jensen diversity for the Jensen–Bregman divergence, and, as a byproduct, a vector-skew generalization of the Jensen–Shannon divergence.

Funding: This research received no external funding.

Acknowledgments: The author is very grateful to the two Reviewers and the Academic Editor for their careful reading, helpful comments, and suggestions which led to this improved manuscript. In particular, Reviewer 2 kindly suggested the stronger bound of Lemma 1 and hinted at Theorem 1. .

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

References

1. Billingsley, P. *Probability and Measure*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
2. Deza, M.M.; Deza, E. *Encyclopedia of Distances*; Springer: Berlin/Heidelberg, Germany, 2009.
3. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633. [CrossRef]
4. Cover, T.M.; Thomas, J.A. *Elements of information theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
5. Nielsen, F. On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy* **2019**, *21*, 485. [CrossRef]
6. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
7. Sason, I. Tight bounds for symmetric divergence measures and a new inequality relating f-divergences. In Proceedings of the 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015; pp. 1–5.
8. Wong, A.K.; You, M. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1985**, *7*, 599–609. [CrossRef]

9. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [CrossRef]
10. Kafka, P.; Österreicher, F.; Vincze, I. On powers of f-divergences defining a distance. *Stud. Sci. Math. Hung.* **1991**, *26*, 415–422.
11. Fuglede, B. Spirals in Hilbert space: With an application in information theory. *Expo. Math.* **2005**, *23*, 23–45. [CrossRef]
12. Acharyya, S.; Banerjee, A.; Boley, D. Bregman divergences and triangle inequality. In Proceedings of the 2013 SIAM International Conference on Data Mining, Austin, TX , USA, 2–4 May 2013; pp. 476–484.
13. Naghshvar, M.; Javidi, T.; Wigger, M. Extrinsic Jensen–Shannon divergence: Applications to variable-length coding. *IEEE Trans. Inf. Theory* **2015**, *61*, 2148–2164. [CrossRef]
14. Bigi, B. Using Kullback-Leibler distance for text categorization. In *European Conference on Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 305–319.
15. Chatzisavvas, K.C.; Moustakidis, C.C.; Panos, C. Information entropy, information distances, and complexity in atoms. *J. Chem. Phys.* **2005**, *123*, 174111. [CrossRef]
16. Yurdakul, B. Statistical Properties of Population Stability Index. Ph.D. Thesis, Western Michigan University, Kalamazoo, MI, USA, 2018.
17. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A* **1946**, *186*, 453–461.
18. Nielsen, F. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Process. Lett.* **2013**, *20*, 657–660. [CrossRef]
19. Lee, L. Measures of Distributional Similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99*; Association for Computational Linguistics: Stroudsburg, PA, USA, 1999; pp. 25–32. doi:10.3115/1034678.1034693. [CrossRef]
20. Nielsen, F. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv* **2010**, arXiv:1009.4004.
21. Lee, L. On the effectiveness of the skew divergence for statistical language analysis. In Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics (AISTATS 2001), Key West, FL, USA, 4–7 January 2001.
22. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
23. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B (Methodol.)* **1966**, *28*, 131–142. [CrossRef]
24. Sason, I. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy* **2018**, *20*, 383. [CrossRef]
25. Amari, S.I. *Information Geometry and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2016.
26. Jiao, J.; Courtade, T.A.; No, A.; Venkat, K.; Weissman, T. Information measures: The curious case of the binary alphabet. *IEEE Trans. Inf. Theory* **2014**, *60*, 7616–7626. [CrossRef]
27. Yuille, A.L.; Rangarajan, A. The concave-convex procedure (CCCP). In Proceedings of the Neural Information Processing Systems 2002, Vancouver, BC, Canada, 9–14 December 2002; pp. 1033–1040.
28. Nielsen, F.; Nock, R. Skew Jensen-Bregman Voronoi diagrams. In *Transactions on Computational Science XIV*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 102–128.
29. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
30. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904. [CrossRef]
31. Melbourne, J.; Talukdar, S.; Bhaban, S.; Madiman, M.; Salapaka, M.V. On the Entropy of Mixture distributions. Availabline online: http://box5779.temp.domains/~jamesmel/publications/ (accesse on 16 February 2020).
32. Guntuboyina, A. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Trans. Inf. Theory* **2011**, *57*, 2386–2399. [CrossRef]
33. Sason, I.; Verdu, S. f-divergence Inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006. [CrossRef]
34. Melbourne, J.; Madiman, M.; Salapaka, M.V. Relationships between certain f-divergences. In Proceeding of the 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 24–27 September 2019; pp. 1068–1073.

35. Sason, I. On Data-Processing and Majorization Inequalities for f-Divergences with Applications. *Entropy* **2019**, *21*, 1022. [CrossRef]
36. Van Erven, T.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
37. Xu, P.; Melbourne, J.; Madiman, M. Infinity-Rényi entropy power inequalities. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 2985–2989.
38. Nielsen, F.; Nock, R. On the geometry of mixtures of prescribed distributions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2861–2865.
39. Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. De L'institut Henri PoincarÉ* **1948**, *10*, 215–310.
40. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [CrossRef]
41. Lanckriet, G.R.; Sriperumbudur, B.K. On the convergence of the concave-convex procedure. In Proceedings of the Advances in Neural Information Processing Systems 22 (NIPS 2009), Vancouver, BC, Canada, 7–10 December 2009; pp. 1759–1767.
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. Springer Verlag GmbH, European Mathematical Society. Encyclopedia of Mathematics. Available online: https://www.encyclopediaofmath.org/ (accessed on 19 December 2019).
44. Del Castillo, J. The singly truncated normal distribution: A non-steep exponential family. *Ann. Inst. Stat. Math.* **1994**, *46*, 57–66. [CrossRef]
45. 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; pp. 3621–3624.
46. Nielsen, F.; Hadjeres, G. Monte Carlo information geometry: The dually flat case. *arXiv* **2018**, arXiv:1803.07225.
47. Schwander, O.; Nielsen, F. Learning mixtures by simplifying kernel density estimators. In *Matrix Information Geometry*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 403–426.
48. Arthur, D.; Vassilvitskii, S. k-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'07), New Orleans LA, USA, 7–9 January 2007; pp. 1027–1035.
49. Nielsen, F.; Nock, R.; Amari, S.i. On clustering histograms with k-means by using mixed α-divergences. *Entropy* **2014**, *16*, 3273–3301. [CrossRef]
50. Nielsen, F.; Nock, R. Total Jensen divergences: Definition, properties and clustering. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, Australia, 19–24 April 2015; pp. 2016–2020.
51. Topsøe, F. Basic concepts, identities and inequalities-the toolkit of information theory. *Entropy* **2001**, *3*, 162–190. [CrossRef]
52. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada 8–13 December 2014; pp. 2672–2680.
53. Yamano, T. Some bounds for skewed α-Jensen-Shannon divergence. *Results Appl. Math.* **2019**, *3*, 100064. [CrossRef]
54. Kotlerman, L.; Dagan, I.; Szpektor, I.; Zhitomirsky-Geffet, M. Directional distributional similarity for lexical inference. *Nat. Lang. Eng.* **2010**, *16*, 359–389. [CrossRef]
55. Johnson, D.; Sinanovic, S. Symmetrizing the Kullback-Leibler distance. *IEEE Trans. Inf. Theory* **2001**, 1–8.