entropy

MDPI

Article

Two Types of Geometric Jensen–Shannon Divergences

Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan; frank.nielsen.x@gmail.com

Abstract

The geometric Jensen–Shannon divergence (G-JSD) has gained popularity in machine learning and information sciences thanks to its closed-form expression between Gaussian distributions. In this work, we introduce an alternative definition of the geometric Jensen–Shannon divergence tailored to positive densities which does not normalize geometric mixtures. This novel divergence is termed the extended G-JSD, as it applies to the more general case of positive measures. We explicitly report the gap between the extended G-JSD and the G-JSD when considering probability densities, and show how to express the G-JSD and extended G-JSD using the Jeffreys divergence and the Bhattacharyya distance or Bhattacharyya coefficient. The extended G-JSD is proven to be an f-divergence, which is a separable divergence satisfying information monotonicity and invariance in information geometry. We derive a corresponding closed-form formula for the two types of G-JSDs when considering the case of multivariate Gaussian distributions that is often met in applications. We consider Monte Carlo stochastic estimations and approximations of the two types of G-JSD using the projective γ-divergences. Although the square root of the JSD yields a metric distance, we show that this is no longer the case for the two types of G-JSD. Finally, we explain how these two types of geometric JSDs can be interpreted as regularizations of the ordinary JSD.

Keywords: Jensen–Shannon divergence; quasi-arithmetic means; total variation distance; Bhattacharyya distance; Chernoff information; Jeffreys divergence; Taneja divergence; geometric mixtures; exponential families; projective γ-divergences; f-divergence; separable divergence; information monotonicity

Academic Editor: Nikolai Leonenko

Received: 8 August 2025
Revised: 29 August 2025
Accepted: 9 September 2025
Published: 11 September 2025

Citation: Nielsen, F. Two Types of Geometric Jensen–Shannon Divergences. *Entropy* **2025**, *27*, 947. https://doi.org/10.3390/e27090947

1. Introduction

1.1. Kullback–Leibler and Jensen–Shannon Divergences

Let $(\mathcal{X}, \mathcal{E}, \mu)$ be a measure space on the sample space $\mathcal{X}$, σ-algebra of events $\mathcal{E}$, with μ a prescribed positive measure on the measurable space $(\mathcal{X}, \mathcal{E})$ (e.g., counting measure or Lebesgue measure). Let $M_+(\mathcal{X}) = \{Q\}$ be the set of positive distributions Q and $M_+^1(\mathcal{X}) = \{P\}$ be the subset of probability measures P. We denote by $M_\mu = \{\frac{\mathrm{d}Q}{\mathrm{d}\mu} : Q \in M_+(\mathcal{X})\}$ and $M_\mu^1 = \{\frac{\mathrm{d}P}{\mathrm{d}\mu} : P \in M_+^1(\mathcal{X})\}$ the corresponding sets of Radon–Nikodym positive and probability densities, respectively.

Consider two probability measures P_1 and P_2 of $M_+^1(\mathcal{X})$ with Radon–Nikodym densities with respect to μ $p_1 := \frac{\mathrm{d}P_1}{\mathrm{d}\mu} \in M_\mu^1$ and $p_2 := \frac{\mathrm{d}P_2}{\mathrm{d}\mu} \in M_\mu^1$, respectively. The deviation of P_1 to P_2 (also called distortion, dissimilarity, or deviance) is commonly measured in information theory [1] by the Kullback–Leibler divergence (KLD):

$$\mathrm{KL}(p_1, p_2) := \int p_1 \log \frac{p_1}{p_2} \, \mathrm{d}\mu = E_{p_1}\left[\log \frac{p_1}{p_2}\right]. \tag{1}$$

Informally, the KLD quantifies the information lost when p_2 is used to approximate p_1 by measuring, on average, the surprise when outcomes sampled from p_1 are assumed to emanate from p_2: Shannon entropy $H(p) = \int p \log \frac{1}{p} \, \mathrm{d}\mu$ is the expected surprise $H(p) = E_p[-\log p]$, where $-\log p(x)$ measures the surprise of the outcome x. Logarithms are taken to base 2 when information is measured in bits, and to base e when it is measured in nats. Gibbs' inequality asserts that $\mathrm{KL}(P_1, P_2) \geq 0$ with equality if and only if $P_1 = P_2$ μ-almost everywhere. Since $\mathrm{KL}(p_1, p_2) \neq \mathrm{KL}(p_2, p_1)$, various symmetrization schemes of the KLD have been proposed in the literature [1] (e.g., Jeffreys divergence [1,2], resistor average divergence [3] (harmonic KLD symmetrization), Chernoff information [1], etc.)

An important symmetrization technique of the KLD is the Jensen–Shannon divergence [4,5] (JSD):

$$\mathrm{JS}(p_1, p_2) := \frac{1}{2} \left(\mathrm{KL}(p_1, a) + \mathrm{KL}(p_2, a) \right), \tag{2}$$

where $a = \frac{1}{2} p_1 + \frac{1}{2} p_2$ denotes the statistical mixture of p_1 and p_2. The JSD is guaranteed to be upper-bounded by $\log 2$ even when the support of p_1 and p_2 differ, making it attractive in applications. Furthermore, its square root $\sqrt{\mathrm{JS}}$ yields a metric distance [6,7].

The JSD can be extended to a set of densities to measure the diversity of the set as an information radius [8]. In information theory, the JSD can also be interpreted as an information gain [6] since it can be equivalently written as

$$\mathrm{JS}(p_1, p_2) = H\left(\frac{1}{2} p_1 + \frac{1}{2} p_2 \right) - \frac{H(p_1) + H(p_2)}{2},$$

where $H(p) = -\int p \log p \, \mathrm{d}\mu$ is Shannon entropy (Shannon entropy for discrete measures and differential entropy for continuous measures). The JSD has also been defined in the setting of quantum information [9], where it has also been proven that its square root yields a metric distance [10].

Remark 1. *Both the KLD and the JSD belong to the family of f-divergences [11,12] defined for a convex generator $f(u)$ (strictly convex at 1) by*

$$I_f(p_1, p_2) := \int p_1 \, f\left(\frac{p_2}{p_1} \right) \mathrm{d}\mu.$$

Indeed, we have $\mathrm{KL}(p_1, p_2) = I_{f_{\mathrm{KL}}}(p_1, p_2)$ and $\mathrm{JS}(p_1, p_2) = I_{f_{\mathrm{JS}}}(p_1, p_2)$ for the following generators:

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

The family of f-divergences is the invariant divergences in information geometry [13]. The f-divergences guarantee information monotonicity by coarse graining [13] (also called lumping in information theory [14]). Using Jensen inequality, we get $I_f(p_1, p_2) \geq f(1)$.

Remark 2. *The metrization of f-divergences was studied in [15]. Once a metric distance $D(p_1, p_2)$ is given, we may use the following metric transform [16] to obtain another metric which is guaranteed to be bounded by 1:*

$$0 \leq d(p_1, p_2) = \frac{D(p_1, p_2)}{1 + D(p_1, p_2)} \leq 1.$$

1.2. Jensen–Shannon Symmetrization of Dissimilarities with Generalized Mixtures

In [17], a generalization of the KLD Jensen–Shannon symmetrization scheme was studied for arbitrary statistical dissimilarity $D(\cdot, \cdot)$ using an arbitrary weighted mean [18]

M_α. A generic weighted mean $M_\alpha(a,b) = M_{1-\alpha}(b,a)$ for $a,b \in \mathbb{R}_{>0}$ is a continuous symmetric monotonic map $\alpha \in [0,1] \mapsto M_\alpha(a,b)$ such that $M_0(a,b) = b$ and $M_1(a,b) = 1$. For example, the quasi-arithmetic means [18] are defined according to a monotonous continuous function ϕ as follows:

$$M_\alpha^\phi(a,b) := \phi^{-1}(\alpha\phi(a) + (1-\alpha)\phi(b)).$$

When $\phi_p(u) = u^p$, we get the p-power mean $M_\alpha^{\phi_p}(a,b) = (\alpha a^p + (1-\alpha)b^p)^{\frac{1}{p}}$ for $p \in \mathbb{R}\setminus\{0\}$. We extend ϕ_p for $p = 0$ by defining $\phi_0(u) = \log u$, and get $M_\alpha^{\phi_0}(a,b) = a^\alpha b^{1-\alpha}$, the weighted geometric mean G_α.

Let us recall the generalization of the Jensen–Shannon symmetrization scheme of a dissimilarity measure presented in [17]:

Definition 1 ((α,β) M-JS dissimilarity [17])**.** *The Jensen–Shannon skew symmetrization of a statistical dissimilarity $D(\cdot,\cdot)$ with respect to an arbitrary weighted bivariate mean $M_\alpha(\cdot,\cdot)$ is given by*

$$D^{\mathrm{JS}}_{M_\alpha,\beta}(p_1,p_2) := \beta\, D\Big(p_1, (p_1p_2)_{M_\alpha}\Big) + (1-\beta)\, D\Big(p_2, (p_1p_2)_{M_\alpha}\Big), \quad (\alpha,\beta) \in (0,1)^2, \tag{3}$$

where $(p_1p_2)_{M_\alpha}$ is the statistical normalized weighted M-mixture of p_1 and p_2:

$$(p_1p_2)_{M_\alpha}(x) := \frac{M_\alpha(p_1(x),p_2(x))}{\int M_\alpha(p_1(x),p_2(x))\,\mathrm{d}\mu(x)}. \tag{4}$$

Remark 3. *A more general definition is given in [17] by using another arbitrary weighted mean N_β to average the two dissimilarities in Equation (3):*

$$D^{\mathrm{JS}}_{M_\alpha,N_\beta}(p_1,p_2) := N_\beta\Big(D\Big(p_1,(p_1p_2)_{M_\alpha}\Big), D\Big(p_2,(p_1p_2)_{M_\alpha}\Big)\Big), \quad (\alpha,\beta) \in (0,1)^2. \tag{5}$$

When $N_\beta = A_\alpha$, the weighted arithmetic mean $A_\alpha(a,b) = \alpha a + (1-\alpha)b$, Equation (5) amounts to Equation (3).

When $\alpha = \frac{1}{2}$, we write, for short, $(p_1p_2)_M$ instead of $(p_1p_2)_{M_{\frac{1}{2}}}$ in the reminder.

When $D = \mathrm{KL}$, $M = N = A_{\frac{1}{2}}$, Equation (5) yields the Jensen–Shannon divergence of Equation (2): $\mathrm{JS}(p_1,p_2) = \mathrm{KL}^{\mathrm{JS}}_{A_{\frac{1}{2}},A_{\frac{1}{2}}}(p_1,p_2) = \mathrm{KL}^{\mathrm{JS}}_{A,A}(p_1,p_2)$.

Lower and upper bounds for the skewed α-Jensen–Shannon divergence were reported in [19].

The abstract mixture normalizer of $(p_1p_2)_{M_\alpha}$ shall be denoted by

$$Z_{M_\alpha}(p_1,p_2) := \int M_\alpha(p_1(x),p_2(x))\,\mathrm{d}\mu(x),$$

so that the normalized M-mixture is written as $(p_1p_2)_{M_\alpha}(x) = \frac{M_\alpha(p_1(x),p_2(x))}{Z_{M_\alpha}(p_1,p_2)}$. The normalizer $Z_{M_\alpha}(p_1,p_2)$ is always finite, and thus the weighted M-mixtures $(p_1p_2)_{M_\alpha}$ are well-defined:

Proposition 1. *For any generic weighted mean M_α, we have the normalizer of the weighted M-mixture bounded by 2:*

$$0 \le Z_{M_\alpha}(p_1,p_2) \le 2.$$

Proof. Since M_α is a scalar weighted mean, it satisfies the following in-betweenness property:

$$\min\{p_1(x), p_2(x)\} \leq M_\alpha(p_1(x), p_2(x)) \leq \max\{p_1(x), p_2(x)\}. \tag{6}$$

Hence, by using the following two identities for $a \geq 0$ and $b \geq 0$,

$$\begin{aligned} \min\{a, b\} &= \frac{a+b}{2} - \frac{1}{2}|a-b|, \\ \max\{a, b\} &= \frac{a+b}{2} + \frac{1}{2}|a-b|, \end{aligned}$$

we get

$$\begin{aligned} \int \min\{p_1(x), p_2(x)\}\,\mathrm{d}\mu(x) \leq & \int M_\alpha(p_1(x), p_2(x))\,\mathrm{d}\mu(x) \leq \int \max\{p_1(x), p_2(x)\}\,\mathrm{d}\mu(x), \\ 0 \leq 1 - \mathrm{TV}(p_1, p_2) \leq & \quad Z_{M_\alpha}(p_1, p_2) \quad \leq 1 + \mathrm{TV}(p_1, p_2) \leq 2, \end{aligned} \tag{7}$$

where

$$\mathrm{TV}(p_1, p_2) := \frac{1}{2}\int |p_1 - p_2|\,\mathrm{d}\mu,$$

is the total variation distance, upper-bounded by 1. When the support of the densities p_1 and p_2 intersect (i.e., non-singular probability measures P_1 and P_2), we have $Z_{M_\alpha}(p_1, p_2) > 0$, and therefore the weighted M-mixtures $(p_1 p_2)_{M_\alpha}$ are well-defined. □

The generic Jensen–Shannon symmetrization of dissimilarities given in Definition 1 allows us to re-interpret some well-known statistical dissimilarities:

For example, the Chernoff information [1,20] is defined by

$$C(p_1, p_2) := \max_{\alpha \in (0,1)} B_\alpha(p_1, p_2), \tag{8}$$

where $B_\alpha(p_1, p_2)$ denotes the α-skewed Bhattacharrya distance:

$$B_\alpha(p_1, p_2) := -\log \int p_1^\alpha\, p_2^{1-\alpha}\,\mathrm{d}\mu \tag{9}$$

When $\alpha = \frac{1}{2}$, we note $B(p_1, p_2) = B_{\frac{1}{2}}(p_1, p_2)$, the Bhattacharrya distance. Notice that the Bhattacharrya distance is not a metric distance as it violates the triangle inequality of metrics.

Using the framework of JS-symmetrization of dissimilarities, we can reinterpret the Chernoff information as

$$C(p_1, p_2) = (\mathrm{KL}^*)^{\mathrm{JS}}_{G_{\alpha^*}, A_{\frac{1}{2}}}(p_1, p_2),$$

where α^* is provably the unique optimal skewing factor in Equation (8), such that we have [20]:

$$\begin{aligned} C(p_1, p_2) &= \mathrm{KL}^*(p_1, (p_1 p_2)_{G_{\alpha^*}}) = \mathrm{KL}^*(p_2, (p_1 p_2)_{G_{\alpha^*}}), \\ &= \frac{1}{2}\left(\mathrm{KL}^*(p_1, (p_1 p_2)_{G_{\alpha^*}}) + \mathrm{KL}^*(p_2, (p_1 p_2)_{G_{\alpha^*}})\right), \end{aligned}$$

where KL^* denotes the reverse KLD:

$$\mathrm{KL}^*(p_1, p_2) := \mathrm{KL}(p_2, p_1).$$

Note that the KLD is sometimes called the forward KLD (e.g., [21]), and we have $\mathrm{KL}^{**}(p_1, p_2) = \mathrm{KL}(p_1, p_2)$.

Although arithmetic mixtures are most often used in statistics, geometric mixtures are also encountered, for example in Bayesian statistics [22] or in Markov chain Monte Carlo annealing [23], just to give two examples. In information geometry, statistical power mixtures based on the homogeneous power means are used to perform stochastic integration of statistical models [24].

Proposition 2 (Bhattacharyya distance as G-JSD)**.** *The Bhattacharyya distance [25] and the α-skewed Bhattacharyya distances can be interpreted as JS-symmetrizations of the reverse KLD with respect to the geometric mean G:*

$$\begin{aligned} B(p_1, p_2) &:= -\log \int \sqrt{p_1\, p_2}\, \mathrm{d}\mu = (\mathrm{KL}^*)_G^{\mathrm{JS}}(p_1, p_2), \\ B_\alpha(p_1, p_2) &:= -\log \int p_1^\alpha\, p_2^{1-\alpha}\, \mathrm{d}\mu = (\mathrm{KL}^*)_{G_\alpha}^{\mathrm{JS}}(p_1, p_2). \end{aligned}$$

Proof. Let $m = (p_1 p_2)_G = \frac{\sqrt{p_1 p_2}}{Z(p_1, p_2)}$ denote the weighted geometric mixture with the normalizer $Z_G(p_1, p_2) = \int \sqrt{p_1 p_2}\, \mathrm{d}\mu$. By definition of the JS-symmetrization of the reverse KLD, we have

$$\begin{aligned} (\mathrm{KL}^*)_G^{\mathrm{JS}}(p_1, p_2) &:= \frac{1}{2}(\mathrm{KL}^*(p_1, (p_1 p_2)_G) + \mathrm{KL}^*(p_2, (p_1 p_2)_G)), \\ &= \frac{1}{2}(\mathrm{KL}((p_1 p_2)_G, p_1) + \mathrm{KL}((p_1 p_2)_G, p_2)), \\ &= \frac{1}{2}\left(\int \left(m\, \log \frac{\sqrt{p_1 p_2}}{p_1\, Z_G(p_1, p_2)} + m\, \log \frac{\sqrt{p_1 p_2}}{p_2\, Z_G(p_1, p_2)} \right) \mathrm{d}\mu \right), \\ &= \frac{1}{2}\left(\int \frac{1}{2} m \log \frac{p_2}{p_1}\, \frac{p_1}{p_2} \mathrm{d}\mu - 2 \log Z_G(p_1, p_2) \int m\, \mathrm{d}\mu \right), \\ &= -\log Z_G(p_1, p_2) =: B(p_1, p_2). \end{aligned}$$

The proof carries on similarly for the α-skewed JS-symmetrization of the reverse KLD: we now let $m_\alpha = (p_1 p_2)_{G_\alpha} = \frac{p_1^\alpha p_2^{1-\alpha}}{Z_{G_\alpha}(p_1, p_2)}$ be the α-weighted geometric mixture with the normalizer $Z_{G_\alpha}(p_1, p_2) = \int p_1^\alpha p_2^{1-\alpha}\, \mathrm{d}\mu$, written as Z_{G_α} for short below:

$$\begin{aligned} {\mathrm{KL}^*}_{G_\alpha, \alpha}^{\mathrm{JS}}(p_1, p_2) &:= \alpha\, \mathrm{KL}^*(p_1, (p_1 p_2)_{G_\alpha}) + (1 - \alpha)\, \mathrm{KL}^*(p_2, (p_1 p_2)_{G_\alpha}), \\ &= \alpha\, \mathrm{KL}(m_\alpha, p_1) + (1 - \alpha)\, \mathrm{KL}(m_\alpha, p_2), \\ &= \int \left(\alpha m_\alpha \log \frac{p_1^\alpha\, p_2^{1-\alpha}}{Z_{G_\alpha}\, p_1} + (1 - \alpha) m_\alpha \log \frac{p_1^\alpha p_2^{1-\alpha}}{Z_{G_\alpha}\, p_2} \right) \mathrm{d}\mu, \\ &= -(\alpha + 1 - \alpha) \log Z_{G_\alpha} \int m_\alpha\, \mathrm{d}\mu + \int m_\alpha \log \left(\frac{p_2}{p_1} \right)^{\alpha(1-\alpha)} \left(\frac{p_1}{p_2} \right)^{\alpha(1-\alpha)} \mathrm{d}\mu, \\ &= -\log Z_{G_\alpha}(p_1, p_2) =: B_\alpha(p_1, p_2). \end{aligned}$$

□

Besides information theory [1], the JSD also plays an important role in machine learning [26–28]. However, one drawback that refrains its use in practice is that the JSD between two Gaussian distributions (normal distributions) is not known in closed form, since no analytic formula is known for the differential entropy of a two-component Gaussian mixture [29], and thus the JSD needs to be numerically approximated in practice by various methods.

To circumvent this problem, the geometric G-JSD was defined in [17] as follows:

Definition 2 (G-JSD [17])**.** *The geometric Jensen–Shannon divergence (G-JSD) between two probability densities, p_1 and p_2, is defined by*

$$\mathrm{JS}_G(p_1, p_2) := \frac{1}{2}\left(\mathrm{KL}(p_1, (p_1 p_2)_G) + \mathrm{KL}(p_2, (p_1 p_2)_G)\right),$$

where $(p_1 p_2)_G(x) = \frac{\sqrt{p_1(x)\, p_2(x)}}{\int \sqrt{p_1(x)\, p_2(x)}\, \mathrm{d}\mu}$ *is the (normalized) geometric mixture of p_1 and p_2.*

We have $\mathrm{JS}_G(p_1, p_2) = \mathrm{KL}_G^{\mathrm{JS}}(p_1, p_2)$. Since, by default, the M- mixture JS-symmetrization of dissimilarities D is performed on the right argument (i.e., D_M^{JS}), we may also consider a dual JS-symmetrization by setting the M-mixtures on the left argument. We denote this left mixture JS-symmetrization with $D_M^{\mathrm{JS}^*}$. We have $D_M^{\mathrm{JS}^*}(p_1, p_2) = (D^*)_M^{\mathrm{JS}}(p_1, p_2)$, i.e., the left-sided JS-symmetrization of D amounts to a right-sided JS-symmetrization of the dual dissimilarity $D^*(p_1, p_2) := D(p_2, p_1)$.

Thus, a left-sided G-JSD divergence JS_G^* was also defined in [17]:

Definition 3. *The left-sided geometric Jensen–Shannon divergence (G-JSD) between two probability densities p_1 and p_2 is defined by*

$$\begin{aligned} \mathrm{JS}_G^*(p_1, p_2) &:= \frac{1}{2}\left(\mathrm{KL}((p_1 p_2)_G, p_1) + \mathrm{KL}((p_1 p_2)_G, p_2)\right), \\ &= \frac{1}{2}\left(\mathrm{KL}^*(p_1, (p_1 p_2)_G) + \mathrm{KL}^*(p_2, (p_1 p_2)_G)\right), \end{aligned}$$

where $(p_1 p_2)_G(x) = \frac{\sqrt{p_1(x)\, p_2(x)}}{\int \sqrt{p_1(x)\, p_2(x)}\, \mathrm{d}\mu}$ *is the (normalized) geometric mixture of p_1 and p_2.*

To contrast with the numerical approximation limitation of the JSD between Gaussians, one advantage of the geometric Jensen–Shannon divergence (G-JSD) is that it admits a closed-form expression between Gaussian distributions [17]. However, the G-JSD is no longer bounded. The G-JSD formula between Gaussian distributions has been used in several scenarios. See [30–38] for a few use cases.

Let us express the G-JSD divergence using other familiar divergences.

Proposition 3. *We have the following expression of the geometric Jensen–Shannon divergence:*

$$\mathrm{JS}_G(p_1, p_2) = \frac{1}{4} J(p_1, p_2) - B(p_1, p_2),$$

where $J(p_1, p_2) := \int (p_1 - p_2) \log \frac{p_1}{p_2}\, \mathrm{d}\mu$ *is Jeffreys' divergence [2], and*

$$B(p_1, p_2) = -\log \int \sqrt{p_1 p_2}\, \mathrm{d}\mu = -\log Z_G(p_1, p_2),$$

is the Bhattacharyya distance.

Proof. We have the following:

$$\begin{aligned} \mathrm{JS}_G(p_1, p_2) &:= \frac{1}{2}(\mathrm{KL}(p_1, (p_1 p_2)_G) + \mathrm{KL}(p_2, (p_1 p_2)_G)), \\ &= \frac{1}{2}\left(\int \left(p_1(x) \log \frac{p_1(x)\, Z_G(p_1, p_2)}{\sqrt{p_1(x)\, p_2(x)}} + p_2(x) \log \frac{p_2(x)\, Z_G(p_1, p_2)}{\sqrt{p_1(x)\, p_2(x)}}\right) \mathrm{d}\mu(x)\right), \\ &= \frac{1}{2}\left(\int (p_1(x) + p_2(x)) \log Z_G(p_1, p_2)\, \mathrm{d}\mu(x) + \frac{1}{2}\mathrm{KL}(p_1, p_2) + \frac{1}{2}\mathrm{KL}(p_2, p_1)\right), \\ &= \log Z_G(p_1, p_2) + \frac{1}{4} J(p_1, p_2), \\ &= \frac{1}{4} J(p_1, p_2) - B(p_1, p_2). \end{aligned}$$

□

Corollary 1 (G-JSD upper bound). *We have the upper bound* $\mathrm{JS}_G(p,q) \leq \frac{1}{4} J(p,q)$.

Proof. Since $B(p_1,p_2) \geq 0$ and $\mathrm{JS}_G(p_1,p_2) = \frac{1}{4} J(p_1,p_2) - B(p_1,p_2)$, we have $\mathrm{JS}_G(p,q) \leq \frac{1}{4} J(p,q)$. □

Remark 4. *Although the KLD and JSD are separable divergences (i.e., f-divergences expressed as integrals of scalar divergences), the M-JSD divergence is, in general, not separable, because it requires mixtures to be normalized inside the log terms. Notice that the Bhattacharyya distance is, similarly, not a separable divergence, but the Bhattacharyya similarity coefficient* $\mathrm{BC}(p_1,p_2) = \exp(-B(p_1,p_2)) = \int \sqrt{p_1\, p_2}\, \mathrm{d}\mu$ *is a separable "f-divergence"/f-coefficient for* $f_{\mathrm{BC}}(u) = \sqrt{u}$ *(here, a concave generator):* $\mathrm{BC}(p_1,p_2) = I_{f_{\mathrm{BC}}}(p_1,p_2)$. *Notice that* $f_{\mathrm{BC}}(1) = 1$, *and because of the concavity of* f_{BC}, *we have* $I_{f_{\mathrm{BC}}}(p_1,p_2) \leq f_{\mathrm{BC}}(1) = 1$ *(hence, the term f-coefficient to reflect the notion of a similarity measure).*

1.3. Paper Outline

The paper is organized as follows: We first give an alternative definition of the M-JSD in Section 2 (Definition 4) which extends to positive measures and does not require normalization of geometric mixtures. We call this new divergence the extended M-JSD, and we compare the two types of geometric JSDs when dealing with probability measures. In Section 4, we show that these normalized/extended M-JSD divergences can be interpreted as regularizations of the Jensen–Shannon divergence, and exhibit several bounds. We discuss Monte Carlo stochastic approximations and approximations using γ-divergences [39] in Section 5. For the case of geometric mixtures, although the G-JSD is not an f-divergence, we show that the extended G-JSD is an f-divergence (Proposition 5), and we express both the G-JSD and the extended G-JSD using both the Jeffreys divergence and the Bhattacharyya divergence or coefficient. We report a related closed-form formula for the G-JSD and extended G-JSD between two Gaussian distributions in Section 3. Finally, we summarize the main results in the concluding section, Section 6.

A list of notations is provided in Nomenclature.

2. A Novel Definition: The G-JSD, Extended to Positive Measures

2.1. Definition and Properties

We may consider the following two modifications of the G-JSD:

- First, we replace the KLD with the extended KLD between positive densities $q_1 \in M_\mu^+$ and $q_2 \in M_\mu^+$ instead of normalized densities:

$$\mathrm{KL}^+(q_1,q_2) := \int \left(q_1 \log \frac{q_1}{q_2} + q_2 - q_1 \right) \mathrm{d}\mu, \tag{10}$$

  (with $\mathrm{KL}^+(p_1,p_2) = \mathrm{KL}(p_1,p_2)$);
- Second, we consider unnormalized M-mixture densities:

$$(q_1 q_2)_{\tilde{M}_\alpha}(x) := M_\alpha(q_1(x), q_2(x)),$$

  where we use the $\tilde{M}$ tilde notation to indicate that the M-mixture is not normalized, instead of normalized densities $(q_1 q_2)_{M_\alpha}(x)$.

The extended KLD can be interpreted as a pointwise integral of a scalar Bregman divergence obtained for the negative Shannon entropy generator [40]. This proves that $\mathrm{KL}^+(q_1,q_2) \geq 0$ with equality if and only if $q_1 = q_2$ μ-almost everywhere. Notice that $\mathrm{KL}(q_1,q_2)$ may be negative when q_1 and/or q_2 are not normalized to probability densities, but we always have $\mathrm{KL}^+(q_1,q_2) \geq 0$.

The extended KLD is an extended f-divergence [41]: $\mathrm{KL}^+(q_1, q_2) = I^+_{f_{\mathrm{KL}^+}}(q_1, q_2)$ for $f_{\mathrm{KL}^+}(u) = -\log(u) + u - 1$, where $I^+_f(q_1, q_2)$ denotes the f-divergence extended to positive densities q_1 and q_2:

$$I^+_f(q_1, q_2) = \int q_1 \, f\left(\frac{q_2}{q_1}\right) \mathrm{d}\mu.$$

Remark 5. *As a side remark, it is preferable in practice to estimate the KLD between p_1 and p_2 by Monte Carlo methods using Equation (10) instead of Equation (1) in order to guarantee the non-negativeness of the KLD (Gibbs' inequality). Indeed, the sampling of s samples $x_1, \ldots, x_s$, defining two unnormalized distributions $q_1(x) = \frac{1}{s}\sum_{i=1}^s p_1(x)\delta_{x_i}(x)$ and $q_2(x) = \frac{1}{s}\sum_{i=1}^s p_2(x)\delta_{x_i}(x)$, where*

$$\delta_{x_i}(x) = \begin{cases} 1, & \text{if } x = x_i \\ 0, & \text{otherwise} \end{cases}.$$

Remark 6. *For an arbitrary distortion measure $D^+(q_1, q_2)$ between positive measures q_1 and q_2, we can build a corresponding projective divergence $\tilde{D}(q_1, q_2)$ as follows:*

$$\tilde{D}(q_1, q_2) := D^+\left(\frac{q_1}{Z(q_1)}, \frac{q_1}{Z(q_2)}\right),$$

where $Z(q) := \int q \, \mathrm{d}\mu$ is the normalization factor of the positive density q. The divergence $\tilde{D}$ is said to be projective because we have, for all $\lambda_1 > 0, \lambda_2 > 0$, the property that $\tilde{D}(\lambda_1 q_1, \lambda_2 q_2) = \tilde{D}(q_1, q_2) = D^+(p_1, p_2)$, where $p_i = \frac{q_i}{Z(q_i)}$ are the normalized densities. The projective Kullback–Leibler divergence $\widetilde{\mathrm{KL}}$ is thus another projective extension of the KLD to non-normalized densities which coincide with the KLD for probability densities. But the projective KLD is different from the extended KLD of Equation (10), and furthermore, we have $\widetilde{\mathrm{KL}}(q_1, q_2) = 0$ if and only if $q_1 = \lambda \, q_2$ μ-almost everywhere for some $\lambda > 0$.

Let us now define the Jensen–Shannon symmetrization of an extended statistical divergence D^+ with respect to an arbitrary weighted mean M_α as follows:

Definition 4 (Extended M-JSD)**.** *A Jensen–Shannon skew symmetrization of a statistical divergence $D^+(\cdot, \cdot)$ between two positive measures q_1 and q_2 with respect to a weighted mean M_α is defined by*

$$D^{\mathrm{JS}^+}_{\tilde{M}_\alpha, \beta}(q_1, q_2) := \beta \, D^+\left(q_1, (q_1 q_2)_{\tilde{M}_\alpha}\right) + (1 - \beta) \, D^+\left(q_1, (q_1 q_2)_{\tilde{M}_\alpha}\right), \tag{11}$$

When $\beta = \frac{1}{2}$, we write, for short, $D^{\mathrm{JS}^+}_{\tilde{M}_\alpha}(q_1, q_2)$, and furthermore, when $\alpha = \frac{1}{2}$, we simplify the notation to $D^{\mathrm{JS}^+}_{\tilde{M}}(q_1, q_2)$.

When $D^+ = \mathrm{KL}^+$, we obtain the extended geometric Jensen–Shannon divergence, $\mathrm{JS}^+_{\tilde{G}}(q_1, q_2) = \mathrm{KL}^{\mathrm{JS}^+}_{\tilde{G}}(q_1, q_2)$:

Definition 5 (Extended G-JSD)**.** *The extended geometric Jensen–Shannon divergence between two positive densities q_1 and q_2 is*

$$\mathrm{JS}^+_{\tilde{G}}(q_1, q_2) = \frac{1}{2}\left(\mathrm{KL}^+(q_1, (q_1 q_2)_{\tilde{G}}) + \mathrm{KL}^+(q_2, (q_1 q_2)_{\tilde{G}})\right), \tag{12}$$

The extended G-JSD between two normalized densities p_1 and p_2 is thus

$$\mathrm{JS}^{+}_{\tilde{G}}(p_1,p_2) = \frac{1}{2}\left(\int\left(p_1\log\frac{p_1}{\sqrt{p_1\,p_2}}+p_2\log\frac{p_2}{\sqrt{p_1\,p_2}}\right)\mathrm{d}\mu+\int\sqrt{p_1\,p_2}\,\mathrm{d}\mu)\right)-1, \quad (13)$$

$$= \frac{1}{2}\left(\int\left(p_1\log\sqrt{\frac{p_1}{p_2}}+p_2\log\sqrt{\frac{p_2}{p_1}}\right)\mathrm{d}\mu+Z_G(p_1,p_2)\right)-1, \quad (14)$$

with $Z_G(p_1,p_2)=\exp(-B(p_1,p_2))$.

Thus, we get the following propositions:

Proposition 4. *The extended geometric Jensen–Shannon divergence (G-JSD) can be expressed as follows:*

$$\mathrm{JS}^{+}_{\tilde{G}}(p_1,p_2)=\frac{1}{4}J(p_1,p_2)+\exp(-B(p_1,p_2))-1.$$

Proof. We have

$$\begin{aligned}\mathrm{JS}^{+}_{\tilde{G}}(p_1,p_2) &= \frac{1}{2}\left(\mathrm{KL}^{+}(p_1,(p_1p_2)_{\tilde{G}})+\mathrm{KL}^{+}(p_2,(p_1p_2)_{\tilde{G}})\right),\\ &= \frac{1}{2}\left(\int\left(p_1\log\sqrt{\frac{p_1}{p_2}}+p_2\log\sqrt{\frac{p_2}{p_1}}+2\sqrt{p_1\,p_2}-(p_1+p_2)\right)\mathrm{d}\mu\right),\\ &= \int\frac{1}{4}(p_1-p_2)\log\frac{p_1}{p_2}\,\mathrm{d}\mu+\int\sqrt{p_1\,p_2}\,\mathrm{d}\mu-1,\\ &= \frac{1}{4}J(p_1,p_2)+\exp(-B(p_1,p_2))-1.\end{aligned}$$

□

Thus, we can express the gap between $\mathrm{JS}^{+}_{\tilde{G}}(p_1,p_2)$ and $\mathrm{JS}_G(p_1,p_2)$:

$$\Delta_G(p_1,p_2)=\mathrm{JS}^{+}_{\tilde{G}}(p_1,p_2)-\mathrm{JS}_G(p_1,p_2)=\exp(-B(p_1,p_2))+B(p_1,p_2)-1.$$

Since $Z_G(p_1,p_2)=\exp(-B(p_1,p_2))$, we have:

$$\Delta_G(p_1,p_2)=Z_G(p_1,p_2)-\log Z_G(p_1,p_2)-1.$$

Proposition 5. *The extended G-JSD is an f-divergence for the generator*

$$f_{\tilde{G}}(u)=\frac{1}{4}(u-1)\log u+\sqrt{u}-1.$$

That is, we have $\mathrm{JS}^{+}_{\tilde{G}}(p_1,p_2)=I_{f_{\tilde{G}}}(p_1,p_2)$.

Proof. We proved that $\mathrm{JS}^{+}_{\tilde{G}}(p_1,p_2)=\frac{1}{4}J(p_1,p_2)+\mathrm{BC}(p_1,p_2)-1$. The Jeffreys divergence is an *f*-divergence for the generator $f_J(u)=(u-1)\log u$, and the Bhattacharyya coefficient is an *f*-coefficient for $f_{\mathrm{BC}}(u)=\sqrt{u}$ (a "*f*-divergence" for a concave generator). Thus, we have

$$f_{\tilde{G}}(u)=\frac{1}{4}(u-1)\log u+\sqrt{u}-1,$$

such that $\mathrm{JS}^{+}_{\tilde{G}}(p_1,p_2)=I_{f_{\tilde{G}}}(p_1,p_2)$. We check that $f_{\tilde{G}}(u)$ is convex, since $f''_{\tilde{G}}(u)=\frac{\sqrt{u}(u+1)-u}{4u^{\frac{5}{2}}}$ (and by a change of variable $t=\sqrt{u}$, the numerator $t(t^2-t+1)$ is shown to be positive, since the discriminant of t^2-t+1 is negative), and we have $f_{\tilde{G}}(1)=0$. Thus, the extended G-JSD is a proper *f*-divergence. □

It follows that the extended G-JSD satisfies the information monotonicity of invariant divergences in information geometry [13].

By abuse of notations, we have

$$\mathrm{KL}^{+}(q_1, q_2) := \mathrm{KL}(q_1, q_2) + \int (q_2 - q_1)\, \mathrm{d}\mu,$$

although q_1 and q_2 may not need to be normalized in the KL term (which can then yield a potentially negative value). Letting $Z(q_i) := \int q_i \, \mathrm{d}\mu$ be the total mass of positive density q_i, we have

$$\mathrm{KL}^{+}(q_1, q_2) = \mathrm{KL}(q_1, q_2) + Z(q_2) - Z(q_1). \tag{15}$$

Let $\tilde{m}_\alpha = M_\alpha(q_1, q_2)$ be the unnormalized M-mixture of positive densities q_1 and q_2, and set $Z_{M_\alpha} = \int \tilde{m}_\alpha \, \mathrm{d}\mu$ be the normalization term so that we have $m_\alpha = \frac{\tilde{m}_\alpha}{Z_{M_\alpha}}$ and $\tilde{m}_\alpha = Z_{M_\alpha}\, m_\alpha$. When clear from context, we write Z_α instead of Z_{M_α}.

We get, after elementary calculus, the following identity:

$$\mathrm{JS}^{+}_{\tilde{M}_\alpha,\beta}(q_1, q_2) = \mathrm{JS}_{M_\alpha,\beta}(q_1, q_2) - (\beta Z(q_1) + (1-\beta) Z(q_2)) \log Z_\alpha + Z_\alpha - (\beta Z(q_1) + (1-\beta) Z(q_2)). \tag{16}$$

Therefore, the difference gap $\Delta_{M_\alpha,\beta}(p_1, p_2)$ (written for short as $\Delta(p_1, p_2)$) between the normalized JSD and the unnormalized M-JSD between two normalized densities p_1 and p_2 (i.e., with $Z_1 = Z(p_1) = 1$ and $Z_2 = Z(p_2) = 1$) is

$$\Delta(p_1, p_2) := \mathrm{JS}^{+}{}_{\tilde{M}_\alpha,\beta}(p_1, p_2) - \mathrm{JS}^{M_\alpha,\beta}(p_1, p_2) = Z_\alpha - \log(Z_\alpha) - 1. \tag{17}$$

Proposition 6 (Extended/normalized M-JSD Gap)**.** *The following identity holds:*

$$\mathrm{JS}^{+}{}_{\tilde{M}_\alpha,\beta}(p_1, p_2) = \mathrm{JS}_{M_\alpha,\beta}(p_1, p_2) + Z_\alpha - \log(Z_\alpha) - 1.$$

Thus, $\mathrm{JS}^{+}{}_{\tilde{M}_\alpha,\beta}(p_1, p_2) \geq \mathrm{JS}_{M_\alpha,\beta}(p_1, p_2)$ when $\Delta(p_1, p_2) \geq 0$, and $\mathrm{JS}^{+}{}_{\tilde{M}_\alpha,\beta}(p_1, p_2) \leq \mathrm{JS}_{M_\alpha,\beta}(p_1, p_2)$ when $\Delta(p_1, p_2) \leq 0$.

When we consider the weighted arithmetic mean A_α, we always have $Z_\alpha = 1$ for $\alpha \in (0, 1)$, and thus the two definitions (Definition 1 and Definition 4) of the A-JSD coincide (i.e., $Z_\alpha^A - \log(Z_\alpha^A) - 1 = 0$):

$$\mathrm{JS}_A(p_1, p_2) = \mathrm{JS}_{\tilde{A}}(p_1, p_2).$$

However, when the weighted mean M_α differs from the weighted arithmetic mean (i.e., $M_\alpha \neq A_\alpha$), the two definitions of the M-JSD JS_M and extended M-JSD $\mathrm{JS}_{\tilde{M}}$ differ by the gap expressed in Equation (17).

Remark 7. *When information is measured in bits, logarithms are taken to base* 2*, and when information is measured in nats, base e is considered. Thus, we shall generally consider the gap* $\Delta_b = Z_\alpha - \log_b(Z_\alpha) - 1$*, where b denotes the base of the logarithm. When* $b = e$*, we have* $\Delta_e \geq 0$ *for all* $Z_\alpha > 0$*. When* $b = 2$*, we have* $\Delta_2 = Z_\alpha - \log_2(Z_\alpha) - 1 \geq 0$ *when* $0 < Z_\alpha \leq 1$ *or* $Z_\alpha \geq 2$*. But since* $Z_\alpha \leq 2$ *(see Equation (7)), the condition simplifies to* $\Delta_2 \geq 0$ *if and only if* $Z_\alpha \leq 1$*.*

Remark 8. *Although* $\sqrt{\mathrm{JS}}$ *is a metric distance [5],* $\sqrt{\mathrm{JS}_G}$ *is not a metric distance, as the triangle inequality is not satisfied. It suffices to report a counterexample of the triangle inequality for a triple of points* p_1*,* p_2*, and* p_3*: Consider* $p_1 = (0.55, 0.45)$*,* $p_2 = (0.002, 0.998)$*, and* $p_3 = (0.045, 0.955)$*. Then we have* $\sqrt{\mathrm{JS}_G}(p_1, p_2) \approx 1.0263227\ldots$*,* $\sqrt{\mathrm{JS}_G}(p_1, p_3) \approx 0.63852342\ldots$*, and* $\sqrt{\mathrm{JS}_G}(p_3, p_2) \approx 0.19794622\ldots$*. The triangle inequality fails with an error of*

$$\sqrt{\mathrm{JS}_G}(p_1, p_2) - (\sqrt{\mathrm{JS}_G}(p_1, p_3) + \sqrt{\mathrm{JS}_G}(p_3, p_2)) \approx 0.1898531\ldots.$$

Similarly, the triangle inequality also fails for the extended G-JSD: we have $\sqrt{\mathrm{JS}_G^+(p_1,p_2)} \approx 1.0788275\ldots$, $\sqrt{\mathrm{JS}_G^+(p_1,p_3)} \approx 0.6691922\ldots$, *and* $\sqrt{\mathrm{JS}_G^+(p_3,p_2)} \approx 0.1984633\ldots$ *with a triangle inequality defect value of*

$$\sqrt{\mathrm{JS}_G^+(p_1,p_2)} - (\sqrt{\mathrm{JS}_G^+(p_1,p_3)} + \sqrt{\mathrm{JS}_G^+(p_3,p_2)}) \approx 0.2111719\ldots.$$

2.2. Power JSDs and (Extended) Min-JSD and Max-JSD

Let $P_{\gamma,\alpha}(a,b) := (\alpha a^\gamma + (1-\alpha)b^\gamma)^{\frac{1}{\gamma}}$ be the γ-power mean for $\gamma \neq 0$ (with $A_\alpha = P_{1,\alpha}$). Further define $P_{0,\alpha}(a,b) = G_\alpha(a,b)$ so that $P_{\gamma,\alpha}$ defines the weighted power means for $\gamma \in \mathbb{R}$ and $\alpha \in (0,1)$ in the reminder. Since $P_{\gamma,\alpha}(a,b) \leq P_{\gamma',\alpha}(a,b)$ when $\gamma' \geq \gamma$ for any $a,b>0$, we have

$$Z_\alpha^{P_\gamma}(p_1,p_2) = \int P_{\gamma,\alpha}(p_1(x),p_2(x))\,\mathrm{d}\mu \leq Z_\alpha^{P_{\gamma'}}(p_1,p_2) = \int P_{\gamma',\alpha}(p_1(x),p_2(x))\,\mathrm{d}\mu. \quad (18)$$

Let $P_\gamma(a,b) = P_{\gamma,\frac{1}{2}}(a,b)$. We have $\lim_{\gamma\to-\infty} P_\gamma(a,b) = \min(a,b)$ and $\lim_{\gamma\to+\infty} P_\gamma(a,b) = \max(a,b)$. Thus, we can define both (extended) min-JSD and (extended) max-JSD. Using the fact that $\min(a,b) = \frac{a+b}{2} - \frac{1}{2}|a-b|$ and $\max(a,b) = \frac{a+b}{2} + \frac{1}{2}|a-b|$, we obtain the extremal mixture normalization terms as follows:

$$Z_{\min}(p_1,p_2) = \int \min(p_1,p_2)\,\mathrm{d}\mu = 1 - \mathrm{TV}(p_1,p_2), \quad (19)$$

$$Z_{\max}(p_1,p_2) = \int \max(p_1,p_2)\,\mathrm{d}\mu = 1 + \mathrm{TV}(p_1,p_2), \quad (20)$$

where $\mathrm{TV}(p_1,p_2) = \frac{1}{2}\int |p_1-p_2|\,\mathrm{d}\mu$ is the total variation distance.

Proposition 7 (max-JSD)**.** *The following upper bound holds for* max*-JSD:*

$$0 \leq \mathrm{JS}^+{}_{\widehat{\max}}(p_1,p_2) \leq \mathrm{TV}(p_1,p_2). \quad (21)$$

Furthermore, the following identity relates the two types of max*-JSDs:*

$$\mathrm{JS}^+{}_{\widehat{\max}}(p_1,p_2) = \mathrm{JS}_{\widehat{\max}}(p_1,p_2) + \mathrm{TV}(p_1,p_2) - \log(1+\mathrm{TV}(p_1,p_2)). \quad (22)$$

Proof. We have

$$\mathrm{JS}^+{}_{\widehat{\max}}(p_1,p_2) := \frac{1}{2}\int \left(p_1 \log \frac{p_1}{\max(p_1,p_2)} + p_2 \log \frac{p_2}{\max(p_1,p_2)} + 2\max(p_1,p_2) - (p_1+p_2) \right) \mathrm{d}\mu.$$

Since both $\log \frac{p_1}{\max(p_1,p_2)} \leq 0$ and $\log \frac{p_2}{\max(p_1,p_2)} \leq 0$, and $\max(a,b) = \frac{a+b}{2} + \frac{1}{2}|b-a|$, we have

$$\mathrm{JS}^+{}_{\widehat{\max}}(p_1,p_2) \leq \int \left(\frac{p_1+p_2}{2} + \frac{1}{2}|p_2-p_1| - \frac{p_1+p_2}{2} \right) \mathrm{d}\mu.$$

That is, $\mathrm{JS}^+{}_{\widehat{\max}}(p_1,p_2) \leq \mathrm{TV}(p_1,p_2)$.

We characterize the gap as follows:

$$\begin{aligned} \Delta_{\max}(p_1,p_2) &= Z_{\max}(p_1,p_2) - \log Z_{\max}(p_1,p_2) - 1, \\ &= \mathrm{TV}(p_1,p_2) - \log(1+\mathrm{TV}(p_1,p_2)) \geq 0, \end{aligned}$$

since $0 \leq \mathrm{TV} \leq 1$. Thus $\mathrm{JS}^+{}_{\widehat{\max}}(p_1,p_2) \geq \mathrm{JS}_{\max}(p_1,p_2)$. □

Proposition 8 (min-JSD)**.** *We have the following lower bound on the extended min-JSD:*

$$\mathrm{JS}^{+}_{\widetilde{\min}}(p_1, p_2) \geq \frac{1}{4} J(p_1, p_2) - \mathrm{TV}(p_1, p_2),$$

where $J(p_1, p_2) := \mathrm{KL}(p_1, p_2) + \mathrm{KL}(p_2, p_1) = \int (p_1 - p_2) \log \frac{p_1}{p_2} \mathrm{d}\mu$ *is the Jeffreys' divergence [2] and*

$$\mathrm{JS}^{+}_{\widetilde{\min}}(p_1, p_2) = \mathrm{JS}_{\min}(p_1, p_2) - \mathrm{TV}(p_1, p_2) + \log(1 - \mathrm{TV}(p_1, p_2)).$$

Proof. We have $Z_{\min}(p_1, p_2) = \int \min\{p_1, p_2\} \mathrm{d}\mu = 1 - \mathrm{TV}(p_1, p_2) \leq 1$ and

$$\begin{aligned} \Delta_{\min}(p_1, p_2) &= Z_{\min}(p_1, p_2) - \log Z_{\min}(p_1, p_2) - 1, \\ &= -\mathrm{TV}(p_1, p_2) - \log(1 - \mathrm{TV}(p_1, p_2)) \geq 0, \end{aligned}$$

since $-x - \log(1 - x) \geq 0$ for $x \leq 1$. Note that the gap can be arbitrarily large when $\mathrm{TV}(p_1, p_2) \to 1^{-}$.

Thus, we have $\mathrm{JS}^{+}_{\widetilde{\min}}(p_1, p_2) \geq \mathrm{JS}_{\min}(p_1, p_2)$.

To get the lower bound, we use the fact that $\min(p_1, p_2) \leq \sqrt{p_1 p_2}$. Indeed, we have

$$\begin{aligned} \mathrm{JS}^{+}_{\widetilde{\min}}(p_1, p_2) &= \frac{1}{2}\left(\int (p_1 \log \frac{p_1}{\min(p_1, p_2)} + p_2 \log \frac{p_2}{\min(p_1, p_2)} + 2\min(p_1, p_2) - (p_1 + p_2) \right) \mathrm{d}\mu, \\ &\geq \frac{1}{2} \int \left(\frac{1}{2} p_1 \log \frac{p_1}{p_2} + \frac{1}{2} p_2 \log \frac{p_2}{p_1} + 2\min(p_1, p_2) - (p_1 + p_2) \right) \mathrm{d}\mu, \\ &= \frac{1}{4} J(p_1, p_2) - \mathrm{TV}(p_1, p_2). \end{aligned}$$

□

Remark 9. *Let us report the total variation distance between two univariate Gaussian distributions* p_{μ_1,σ_1} *and* p_{μ_2,σ_2} *in closed form using the error function [42]* $\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} \mathrm{d}t$.

- *When* $\sigma_1 = \sigma_2 = \sigma$, *we have*

$$\mathrm{TV}(p_1, p_2) = \frac{1}{2} |\Phi(x^*; \mu_2, \sigma) - \Phi(x^*; \mu_1, \sigma)|, \tag{23}$$

where $\Phi(x; \mu, \sigma) = \frac{1}{2}(1 + \mathrm{erf}(\frac{x - \mu}{\sigma\sqrt{2}}))$ *is the cumulative distribution, and*

$$x^* = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)}. \tag{24}$$

- *When* $\sigma_1 \neq \sigma_2$, *we let* $x_1 = \frac{-b - \sqrt{\Delta}}{2a}$ *and* $x_2 = \frac{-b + \sqrt{\Delta}}{2a}$, *where* $\Delta = b^2 - 4ac \geq 0$ *and*

$$a = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}, \tag{25}$$

$$b = 2\left(\frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1} \right), \tag{26}$$

$$c = \left(\frac{\mu_1}{\sigma_1} \right)^2 - \left(\frac{\mu_2}{\sigma_2} \right)^2 - 2 \log \frac{\sigma_2}{\sigma_1}. \tag{27}$$

The total variation is given by

$$\mathrm{TV}(p_1, p_2) = \frac{1}{2}\left(\left| \mathrm{erf}\left(\frac{x_1 - \mu_1}{\sigma_1 \sqrt{2}} \right) - \mathrm{erf}\left(\frac{x_1 - \mu_2}{\sigma_2 \sqrt{2}} \right) \right| + \left| \mathrm{erf}\left(\frac{x_2 - \mu_1}{\sigma_1 \sqrt{2}} \right) - \mathrm{erf}\left(\frac{x_2 - \mu_2}{\sigma_2 \sqrt{2}} \right) \right| \right) \tag{28}$$

Next, we shall consider the important case of p_1 and p_2 belonging to the family of multivariate normal distributions, commonly called Gaussian distributions.

3. Geometric JSDs Between Gaussian Distributions

3.1. Exponential Families

The formula for the G-JSD between two Gaussian distributions was reported in [17] using the more general framework of exponential families. An exponential family [43] is a family of probability measures $\{P_\lambda\}$ with Radon–Nikodym densities p_λ with respect to μ expressed canonically as

$$\begin{aligned} p_\lambda(x) &:= \exp(\langle \theta(\lambda), t(x) \rangle - F(\theta) + k(x)), \\ &= \frac{1}{Z(\theta)} \exp(\langle \theta(\lambda), t(x) \rangle + k(x)), \end{aligned}$$

where $\theta(\lambda)$ is the natural parameter, $t(x)$ the sufficient statistic, $k(x)$ an auxiliary carrier term with respect to μ, and $F(\theta)$ the cumulant function. The partition function $Z(\theta)$ is the normalizer denominator: $Z(\theta) = \exp(F(\theta))$. The cumulant function $F(\theta) = \log Z(\theta)$ is strictly convex and analytic [43], and the partition function $Z(\theta) = \exp(F(\theta))$ is strictly log-convex (and hence also strictly convex).

We consider the exponential family of multivariate Gaussian distributions

$$\mathcal{N} = \{N(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathrm{PD}(d)\},$$

where $\mathrm{PD}(d)$ denotes the set of symmetric positive–definite matrices of size $d \times d$. Let $\lambda{:=}(\lambda_v, \lambda_M) = (\mu, \Sigma)$ denote the compound (vector, matrix) parameter of a Gaussian. The d-variate Gaussian density is given by

$$p_\lambda(x; \lambda) := \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\lambda_M|}} \exp\left(-\frac{1}{2} (x - \lambda_v)^\top \lambda_M^{-1} (x - \lambda_v) \right), \tag{29}$$

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

$$\langle \theta, \theta' \rangle {:=} \theta_v^\top \theta_v' + \mathrm{tr}\left({\theta_M'}^\top \theta_M \right), \tag{30}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace, we rewrite the Gaussian density of Equation (29) in the canonical form of an exponential family:

$$p_\theta(x; \theta) := \exp(\langle t(x), \theta \rangle - F_\theta(\theta)) = p_\lambda(x), \tag{31}$$

where $\theta = \theta(\lambda)$ with

$$\theta = (\theta_v, \theta_M) = \left(\Sigma^{-1} \mu, -\frac{1}{2} \Sigma^{-1} \right) = \theta(\lambda) = \left(\lambda_M^{-1} \lambda_v, -\frac{1}{2} \lambda_M^{-1} \right), \tag{32}$$

is the compound vector-matrix natural parameter and

$$t(x) = (x, -xx^\top), \tag{33}$$

is the compound vector-matrix sufficient statistic. There is no auxiliary carrier term (i.e., $k(x) = 0$). The function F_θ is given by

$$F_\theta(\theta) {:=} \frac{1}{2} \left(d \log \pi - \log |\theta_M| + \frac{1}{2} \theta_v^\top \theta_M^{-1} \theta_v \right), \tag{34}$$

Remark 10. *Beware that when the cumulant function is expressed using the ordinary parameter $\lambda = (\mu, \Sigma)$, the cumulant function $F_\theta(\theta(\lambda))$ is no longer convex:*

$$\begin{aligned} F_\lambda(\lambda) &= \frac{1}{2}\left(\lambda_v^\top \lambda_M^{-1}\lambda_v + \log|\lambda_M| + d\log 2\pi\right), && (35)\\ &= \frac{1}{2}\left(\mu^\top \Sigma^{-1}\mu + \log|\Sigma| + d\log 2\pi\right). && (36)\end{aligned}$$

We convert between the ordinary parameterization $\lambda = (\mu, \Sigma)$ and the natural parameterization θ using these formulas:

$$\theta = (\theta_v, \theta_M) = \begin{cases} \theta_v(\lambda) = \lambda_M^{-1}\lambda_v = \Sigma^{-1}\mu \\ \theta_M(\lambda) = \frac{1}{2}\lambda_M^{-1} = \frac{1}{2}\Sigma^{-1} \end{cases} \quad \Leftrightarrow \quad \lambda = (\lambda_v, \lambda_M) = \begin{cases} \lambda_v(\theta) = \frac{1}{2}\theta_M^{-1}\theta_v = \mu \\ \lambda_M(\theta) = \frac{1}{2}\theta_M^{-1} = \Sigma \end{cases}$$

The geometric mixture $p_{\theta_1}^\alpha p_{\theta_2}^{1-\alpha}$ of two densities of an exponential family is a density $p_{\alpha\theta_1+(1-\alpha)\theta_2}$ of the exponential family with the partition function $Z_\alpha(\theta_1, \theta_2) = \exp(-J_{F,\alpha}(\theta_1, \theta_2))$, where $J_{F,\alpha}(\theta_1, \theta_2)$ denotes the skew Jensen divergence [44,45]:

$$J_{F,\alpha}(\theta_1, \theta_2) := \alpha F(\theta_1) + (1-\alpha)F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2).$$

Therefore, the difference gap of Equation (17) between the G-JSD and the extended G-JSD between exponential family densities is given by

$$\begin{aligned} \Delta(\theta_1, \theta_2) &= \exp(-J_{F,\alpha}(\theta_1, \theta_2)) + J_{F,\alpha}(\theta_1, \theta_2) - 1, && (37)\\ &= Z_\alpha(\theta_1, \theta_2) - \log Z_\alpha(\theta_1, \theta_2) - 1, && (38)\\ &= Z_\alpha(\theta_1, \theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2) - 1. && (39)\end{aligned}$$

Since $Z_\alpha = \exp(-J_{F,\alpha}(\theta_1, \theta_2)) \leq 1$, the gap Δ is negative, and we have

$$\mathrm{JS}^+_{\tilde{G}_\alpha,\beta}(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}) \leq \mathrm{JS}_{G_\alpha,\beta}(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}).$$

Corollary 2. *When $p_1 = p_{\theta_1}$ and $p_2 = p_{\theta_2}$ belongs to a same exponential family with the cumulant function $F(\theta)$, we have*

$$\mathrm{JS}_G(p_{\theta_1}, p_{\theta_2}) = \frac{1}{4}(\theta_2 - \theta_1)^\top(\nabla F(\theta_2) - \nabla F(\theta_1)) - \left(\frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right)\right), \quad (40)$$

since $J(p_{\theta_1}, p_{\theta_2}) = \langle \theta_2 - \theta_1, \nabla F(\theta_2) - \nabla F(\theta_1)\rangle$ amounts to a symmetrized Bregman divergence.

Proof. We have $J(p_{\theta_1}, p_{\theta_2}) = (\theta_2 - \theta_1)^\top(\nabla F(\theta_2) - \nabla F(\theta_1))$ and $J(p_{\theta_1}, p_{\theta_2}) = J_F(\theta_1, \theta_2)$. □

The extended geometric Jensen–Shannon divergence and geometric Jensen–Shannon divergence between two densities of an exponential family are given by

$$\begin{aligned} \mathrm{JS}_G(p_{\theta_1}, p_{\theta_2}) &= \frac{1}{4}(\theta_2 - \theta_1)^\top(\nabla F(\theta_2) - \nabla F(\theta_1)) - \left(\frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right)\right),\\ \mathrm{JS}_{\tilde{G}}(p_{\theta_1}, p_{\theta_2}) &= \frac{1}{4}\langle \theta_2 - \theta_1, \nabla F(\theta_2) - \nabla F(\theta_1)\rangle - \exp(-J_F(\theta_1, \theta_2)) - 1,\\ \mathrm{JS}^*{}_G(p_{\theta_1}, p_{\theta_2}) &= J_F(\theta_1, \theta_2)\end{aligned}$$

Remark 11. *Given two densities p_1 and p_2, the family $\mathcal{G}$ of geometric mixtures $\{(p_1 p_2)_{G_\alpha} \propto p_1^\alpha p_2^{1-\alpha} : \alpha \in (0,1)\}$ forms a 1D exponential family that has been termed the likelihood ratio exponential family [46] (LREF). The cumulant function of this LREF is $F(\alpha) = -B_\alpha(p_1, p_2)$. Hence, $\mathcal{G}$ has also been called a Bhattacharyya arc or Hellinger arc in the literature [47]. However,*

notice that $\mathrm{KL}(p_i : (p_1 p_2)_{G_\alpha})$ *does not necessarily amount to a Bregman divergence, because neither* p_1 *nor* p_2 *belongs to* $\mathcal{G}$.

3.2. Closed-Form Formula for Gaussian Distributions

Let us report the corresponding closed-form formula for d-variate Gaussian distributions.

When $\alpha = \frac{1}{2}$, we proved that $\mathrm{JS}_G(p_1, p_2) = \frac{1}{4} J(p_1, p_2) - B(p_1, p_2)$ and $\mathrm{JS}^+_{\bar{G}}(p_1, p_2) = \frac{1}{4} J(p_1, p_2) + \exp(-B(p_1, p_2)) - 1$ where $\mathrm{BC}(p_1, p_2) = \exp(-B(p_1, p_2))$. Thus, for the case of balanced geometric mixtures, we need to report the closed form for the Jeffreys and Bhattacharyya distances:

$$\begin{aligned} J(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}) &= \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1\Sigma_2^{-1} + \Sigma_2\Sigma_1^{-1}\right) + (\mu_1 - \mu_2)^\top(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) - 2d\right), \\ B(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}) &= \frac{1}{8}(\mu_1 - \mu_2)^\top \bar{\Sigma}^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\log\left(\frac{\det \bar{\Sigma}}{\sqrt{\det\Sigma_1 \det\Sigma_2}}\right), \end{aligned}$$

where $\bar{\Sigma} = \frac{1}{2}(\Sigma_1 + \Sigma_2)$.

Otherwise, for the arbitrary weighted geometric mixture G_α, define $(\theta_1\theta_2)_\alpha = \alpha\theta_1 + (1-\alpha)\theta_2$, the weighted linear interpolation of the natural parameters θ_1 and θ_2.

Corollary 3. *The skew G-Jensen–Shannon divergence* JS^G_α *and the dual skew G-Jensen–Shannon divergence* JS^{*G}_α *between two d-variate Gaussian distributions* $N(\mu_1, \Sigma_1)$ *and* $N(\mu_2, \Sigma_2)$ *is*

$$\begin{aligned} \mathrm{JS}_{G_\alpha}(p_{(\mu_1,\Sigma_1)}, p_{(\mu_2,\Sigma_2)}) &= \alpha\,\mathrm{KL}(p_{(\mu_1,\Sigma_1)}, p_{(\mu_\alpha,\Sigma_\alpha)}) + (1-\alpha)\,\mathrm{KL}(p_{(\mu_2,\Sigma_2)}, p_{(\mu_\alpha,\Sigma_\alpha)}), \\ &= \alpha\, B_F((\theta_1\theta_2)_\alpha, \theta_1) + (1-\alpha)\, B_F((\theta_1\theta_2)_\alpha, \theta_2), \\ &= \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_\alpha^{-1}(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)\right) + \log\left(\frac{|\Sigma_\alpha|}{|\Sigma_1|^\alpha|\Sigma_2|^{1-\alpha}}\right)\right. \\ &= \left. +\alpha(\mu_\alpha - \mu_1)^\top\Sigma_\alpha^{-1}(\mu_\alpha - \mu_1) + (1-\alpha)(\mu_\alpha - \mu_2)^\top\Sigma_\alpha^{-1}(\mu_\alpha - \mu_2) - d\right) \\ \mathrm{JS}^*_{G_\alpha}(p_{(\mu_1,\Sigma_1)}, p_{(\mu_2,\Sigma_2)}) &= (1-\alpha)\,\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)}, p_{(\mu_1,\Sigma_1)}) + \alpha\mathrm{KL}(p_{(\mu_\alpha,\Sigma_\alpha)}, p_{(\mu_2,\Sigma_2)}), \\ &= \alpha\, B_F(\theta_1, (\theta_1\theta_2)_\alpha) + (1-\alpha)\, B_F(\theta_2, (\theta_1\theta_2)_\alpha), \\ &= J_{F,\alpha}(\theta_1, \theta_2) =: B_\alpha(p_{(\mu_1,\Sigma_1)}, p_{(\mu_2,\Sigma_2)}), \\ &= \frac{1}{2}\left(\alpha\mu_1^\top\Sigma_1^{-1}\mu_1 + (1-\alpha)\mu_2^\top\Sigma_2^{-1}\mu_2 - \mu_\alpha^\top\Sigma_\alpha^{-1}\mu_\alpha + \log\frac{|\Sigma_1|^\alpha|\Sigma_2|^{1-\alpha}}{|\Sigma_\alpha|}\right), \\ F(\mu, \Sigma) &= \frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu + \log|\Sigma| + d\log 2\pi\right), \\ F(\theta_v, \theta_M) &= \frac{1}{2}\left(d\log\pi - \log|\theta_M| + \frac{1}{2}\theta_v^\top\theta_M^{-1}\theta_v\right), \\ \Delta(\theta_1, \theta_2) &= \exp(-J_{F,\alpha}(\theta_1, \theta_2)) + J_{F,\alpha}(\theta_1, \theta_2) - 1, \end{aligned}$$

where Σ_α *is the matrix harmonic barycenter:*

$$\Sigma_\alpha = \left(\alpha\Sigma_1^{-1} + (1-\alpha)\Sigma_2^{-1}\right)^{-1}, \tag{41}$$

and

$$\mu_\alpha = \Sigma_\alpha\left(\alpha\Sigma_1^{-1}\mu_1 + (1-\alpha)\Sigma_2^{-1}\mu_2\right). \tag{42}$$

4. The Extended and Normalized G-JSDs as Regularizations of the Ordinary JSD

The M-Jensen–Shannon divergence $\mathrm{JS}_M(p, q)$ can be interpreted as a regularization of the ordinary JSD:

Proposition 9 (JSD regularization). *For any arbitrary mean M, the following identity holds:*

$$\mathrm{JS}_M(p_1, p_2) = \mathrm{JS}(p_1, p_2) + \mathrm{KL}\left(\frac{p_1 + p_2}{2}, (p_1 p_2)_M\right). \tag{43}$$

Notice that $(p_1 p_2)_A = \frac{p_1+p_2}{2}$.

Proof. We have

$$\begin{aligned}
\mathrm{JS}_M(p_1, p_2) &:= \frac{1}{2}(\mathrm{KL}(p_1, (p_1 p_2)_M) + \mathrm{KL}(p_2, (p_1 p_2)_M)),\\
&= \frac{1}{2}\int \left(p_1 \log \frac{p_1\,(p_1 p_2)_A}{(p_1 p_2)_M\,(p_1 p_2)_A} + p_2 \log \frac{p_2\,(p_1 p_2)_A}{(p_1 p_2)_M\,(p_1 p_2)_A}\right) \mathrm{d}\mu,\\
&= \frac{1}{2}\int \left(p_1 \log \frac{p_1}{(p_1 p_2)_A} + p_1 \log \frac{(p_1 p_2)_A}{(p_1 p_2)_M} + p_2 \log \frac{p_2}{(p_1 p_2)_A} + p_2 \log \frac{(p_1 p_2)_A}{(p_1 p_2)_M}\right) \mathrm{d}\mu,\\
&= \frac{1}{2}\int \left(p_1 \log \frac{p_1}{(p_1 p_2)_A} + p_2 \log \frac{p_2}{(p_1 p_2)_A}\right) \mathrm{d}\mu + \int \frac{1}{2}(p_1 + p_2) \log \frac{(p_1 p_2)_A}{(p_1 p_2)_M}\, \mathrm{d}\mu,\\
&= \mathrm{JS}(p_1, p_2) + \int (p_1 p_2)_A \log \frac{(p_1 p_2)_A}{(p_1 p_2)_M}\, \mathrm{d}\mu,\\
&= \mathrm{JS}(p_1, p_2) + \mathrm{KL}((p_1 p_2)_A, (p_1 p_2)_M).
\end{aligned}$$

□

Remark 12. *One way to symmetrize the KLD is to consider two distinct symmetric means $M_1(a, b) = M_1(b, a)$ and $M_2(a, b) = M_2(b, a)$ and define*

$$\mathrm{KL}_{M_1,M_2}(p_1, p_2) = \mathrm{KL}((p_1 p_2)_{M_1}, (p_1 p_2)_{M_2}) = \mathrm{KL}_{M_1,M_2}(p_2, p_1).$$

We notice that $\sqrt{\mathrm{KL}^{A,G}}$ is not a metric distance by reporting a triple of points (p_1, p_2, p_3) that fails the triangle inequality. Consider $p_1 = (0.55, 0.45)$, $p_2 = (0.002, 0.998)$, and $p_3 = (0.045, 0.955)$. We have $\sqrt{\mathrm{KL}_{M_1,M_2}(p_1, p_2)} = 0.5374165\ldots$, $\sqrt{\mathrm{KL}_{M_1,M_2}(p_1, p_3)} = 0.1759400\ldots$, and $\sqrt{\mathrm{KL}_{M_1,M_2}(p_3, p_2)} = 0.08485931\ldots$. The triangle inequality defect is

$$\sqrt{\mathrm{KL}_{M_1,M_2}(p_1, p_2)} - (\sqrt{\mathrm{KL}_{M_1,M_2}(p_1, p_3)} + \sqrt{\mathrm{KL}_{M_1,M_2}(p_3, p_2)}) = 0.2766171\ldots.$$

We can also similarly symmetrize the extended KLD as follows:

$$\mathrm{KL}^{+}_{\tilde{M}_1,\tilde{M}_2}(q_1, q_2) = \mathrm{KL}^{+}((q_1 q_2)_{\tilde{M}_1}, (q_1 q_2)_{\tilde{M}_2}) = \mathrm{KL}_{\tilde{M}_1,\tilde{M}_2}(q_2, q_1).$$

In particular, when $M_1 = A$ and $M_2 = G$, we get the $\mathrm{KL}_{A,M}$ divergence:

$$\mathrm{KL}_{A,M}(p_1, p_2) = \frac{p_1 + p_2}{2} \log \frac{p_1 + p_2}{2\sqrt{p_1 p_2}} + \log Z_G(p_1, p_2),$$

which is related to Taneja T-divergence [48]:

$$T(p_1, p_2) = \int \frac{p_1 + p_2}{2} \log \frac{p_1 + p_2}{2\sqrt{p_1 p_2}}. \tag{44}$$

The T-divergence is an f-divergence [11,12] obtained for the generator $f_T(u) = \frac{1+u}{2} \log \frac{1+u}{2\sqrt{u}}$.

Corollary 4 (JSD lower bound on M-JSD). *We have $\mathrm{JS}_M(p, q) \geq \mathrm{JS}(p, q)$.*

Proof. Since $\mathrm{JS}_M(p, q) = \mathrm{JS}(p, q) + \mathrm{KL}\left(\frac{p+q}{2}, (pq)_M\right)$ and $\mathrm{KL} \geq 0$ by Gibbs' inequality, we have $\mathrm{JS}_M(p, q) \geq \mathrm{JS}(p, q)$. □

Since the extended M-JSD is $\mathrm{JS}^{+}_{\tilde{M}_\alpha,\beta}(p_1,p_2) = \mathrm{JS}_{M_\alpha,\beta}(p_1,p_2) + Z_\alpha - \log(Z_\alpha) - 1$, the extended M-JSD $\mathrm{JS}^{+}_{\tilde{M}_\alpha,\beta}$ can also be interpreted as another regularization of the Jensen–Shannon divergence when dealing with probability densities:

$$\mathrm{JS}^{+}_{\tilde{M}_\alpha,\beta}(p_1,p_2) = \mathrm{JS}(p_1,p_2) + \mathrm{KL}\left(\frac{p_1+p_2}{2}, (p_1p_2)_M\right) + Z_{M_\alpha}(p_1,p_2) - \log(Z_{M_\alpha}(p_1,p_2)) - 1. \tag{45}$$

It is well known that the JSD can be rewritten as a diversity index [4] using concave entropy:

$$\mathrm{JS}(p_1,p_2) = H\left(\frac{p_1+p_2}{2}\right) - \frac{H(p_1)+H(p_2)}{2}. \tag{46}$$

We generalize this decomposition as the difference of a cross-entropy term minus entropies, as follows:

Proposition 10 (M-JSD cross-entropy decomposition)**.** *We have*

$$\mathrm{JS}_M(p_1,p_2) = H^{\times}((p_1p_2)_A, (p_1p_2)_M) - \frac{H(p_1)+H(p_2)}{2}.$$

Proof. From Proposition 9, we have

$$\mathrm{JS}_M(p_1,p_2) = \mathrm{JS}(p_1,p_2) + \mathrm{KL}\left(\frac{p_1+p_2}{2}, (p_1p_2)_M\right).$$

Since $\mathrm{KL}(p_1,p_2) = H^{\times}(p_1,p_2) - H(p_1)$, where $H^{\times}(p_1,p_2) = -\int p_1 \log p_2 \,\mathrm{d}\mu$ is the cross-entropy and $H(p) = -\int p \log p \,\mathrm{d}\mu = H^{\times}(p,p)$ is the entropy. Plugging Equation (46) into Equation (43), we get

$$\begin{aligned}
\mathrm{JS}_M(p_1,p_2) &= H\left(\frac{p_1+p_2}{2}\right) - \frac{H(p_1)+H(p_2)}{2} + H^{\times}\left(\frac{p_1+p_2}{2}, (p_1p_2)_M\right) - H\left(\frac{p_1+p_2}{2}\right), \\
&= H^{\times}\left(\frac{p_1+p_2}{2}, (p_1p_2)_M\right) - \frac{H(p_1)+H(p_2)}{2}.
\end{aligned}$$

Note that when $M = A$, the arithmetic mean, we have $H^{\times}\left(\frac{p_1+p_2}{2}, (p_1p_2)_M\right) = H\left(\frac{p_1+p_2}{2}\right)$ and we recover the fact that $\mathrm{JS}_M(p_1,p_2) = \mathrm{JS}(p_1,p_2)$. □

5. Estimation and Approximation of the Extended and Normalized M-JSDs

Let us recall the two definitions of the extended M-JSD and the normalized M-JSD (for the case of $\alpha = \beta = \frac{1}{2}$) between two normalized densities p_1 and p_2:

$$\begin{aligned}
\mathrm{JS}_M(p_1,p_2) &= \frac{1}{2}(\mathrm{KL}(p_1,(p_1p_2)_M) + \mathrm{KL}(p_2,(p_1p_2)_M)), \\
\mathrm{JS}^{+}_M(p_1,p_2) &= \frac{1}{2}(\mathrm{KL}^{+}(p_1,(p_1p_2)_{\tilde{M}}) + \mathrm{KL}^{+}(p_2,(p_1p_2)_{\tilde{M}})),
\end{aligned}$$

where $(p_1p_2)_M(x) = \frac{M(p_1(x),p_2(x))}{Z_M(p_1,p_2)}$ (with $Z_M(p_1,p_2) = \int M(p_1(x),p_2(x))\,\mathrm{d}\mu(x)$) and $(p_1p_2)_{\tilde{M}}(x) = M(p_1(x),p_2(x))$.

In practice, one needs to estimate the extended and normalized G-JSDs when they do not admit a closed-form formula.

5.1. Monte Carlo Estimators

To estimate $\mathrm{JS}_M(p_1, p_2)$, we can use Monte Carlo samplings to estimate both KLD integrals and mixture normalizers Z_M; for example, the normalizer $Z_M(p_1, p_2)$ is estimated by

$$\hat{Z}_M(p_1, p_2) = \frac{1}{s}\sum_{i=1}^{s}\frac{1}{r(x_i)}M(p_1(x_i), p_2(x_i)),$$

where $r(x)$ is the proposal distribution which can be chosen according to the mean M and the types of probability distributions p_1 and p_2, and $x_1, \ldots, x_s$ are s identically and independently sampled (iid.) from $r(x)$. However, since $(p_1p_2)_M(x)$ is now estimated as $(p_1p_2)_{\hat{M}(x)}$, it is no longer a normalized M-mixture, and thus we consider estimating

$$\mathrm{JS}^{+}_{\hat{M}}(p_1, p_2) = \frac{1}{2}\left(\mathrm{KL}^{+}\left(p_1, (p_1p_2)_{\hat{M}}\right) + \mathrm{KL}^{+}\left(p_2, (p_1p_2)_{\hat{M}}\right)\right)$$

to ensure the non-negativity of the divergence $\mathrm{JSD}_{\hat{M}}$.

Let us consider the estimation of the term

$$\mathrm{KL}^{+}(p_1, (p_1p_2)_{\tilde{M}}) = \int\left(p_1\log\frac{p_1}{M(p_1, p_2)} + M(p_1, p_2) - p_1\right)\mathrm{d}\mu.$$

By choosing the proposal distribution $r(x) = p_1(x)$, we have $\mathrm{KL}^{+}\left(p_1, (p_1p_2)_{\hat{M}}\right) \approx \widehat{\mathrm{KL}^{+}}(p_1, (p_1p_2)_{\tilde{M}})$ (for large enough s), where

$$\widehat{\mathrm{KL}}^{+}(p_1, (p_1p_2)_{\tilde{M}}) = \frac{1}{s}\sum_{i=1}^{s}\left(\log\frac{p_1(x_i)}{M(p_1(x_i), p_2(x_i))} + \frac{1}{p_1(x_i)}M(p_1(x_i), p_2(x_i)) - 1\right).$$

Monte Carlo (MC) stochastic integration [49] is a well-studied topic in statistics, with many results available regarding the consistency and variance of MC estimators.

Note that even if we have a generic formula for the G-JSD between two densities of an exponential family given by Corollary 2, the cumulant function $F(\theta)$ may not be in closed form [50,51]. This is the case when the sufficient statistic vector of the exponential family is $t(x) = (x, x^2, \ldots, x^m)$ (for $m \geq 5$), yielding the polynomial exponential family (also called exponential–polynomial family [51]).

5.2. Approximations via γ-Divergences

One way to circumvent the lack of computational tractable density normalizers is to consider the family of γ-divergences [39] instead of the KLD:

$$\tilde{D}_\gamma(q_1, q_2) = \frac{1}{\gamma(1+\gamma)}\log I_\gamma(q_1, q_2) - \frac{1}{\gamma}\log I_\gamma(q_1, q_2) + \frac{1}{1+\gamma}\log I_\gamma(q_1, q_2), \quad \gamma > 0,$$

where

$$I_\gamma(q_1, q_2) = \int q_1(x)\, q_2^\gamma(x)\,\mathrm{d}\mu(x).$$

The γ-divergences are projective divergences, i.e., they enjoy the property that

$$\tilde{D}_\gamma(\lambda_1 q_1, \lambda_2 q_2) = \tilde{D}_\gamma(q_1, q_2), \quad \forall \lambda_1 > 0, \lambda_2 > 0.$$

Furthermore, we have $\lim_{\gamma\to 0}\tilde{D}_\gamma(p_1, p_2) = \mathrm{KL}(p_1, p_2)$. (Note that the KLD is not projective.)

Let us define the projective M-JSD:

$$\mathrm{JS}_{\tilde{M},\gamma}(p_1, p_2) = \frac{1}{2}\left(\tilde{D}_\gamma(p_1, (p_1p_2)_{\tilde{M}}) + \tilde{D}_\gamma(p_2, (p_1p_2)_{\tilde{M}})\right). \tag{47}$$

We have, for $\gamma = \epsilon$, a small enough value (e.g., $\epsilon \leq 10^{-3}$), $\mathrm{JS}_M(p_1, p_2) \approx \mathrm{JS}_{\tilde{M},\gamma}(p_1, p_2)$, since

$$\mathrm{KL}(p_1, (p_1p_2)_M) \approx_{\gamma=\epsilon} \tilde{D}_\gamma(p_1, (p_1p_2)_{\tilde{M}}).$$

In particular, for exponential family densities $p_{\theta_1}(x) = \frac{q_{\theta_1}(x)}{Z(\theta_1)}$ and $p_{\theta_2}(x) = \frac{q_{\theta_2}(x)}{Z(\theta_2)}$, we have

$$I_\gamma(p_{\theta_1}, p_{\theta_2}) = \exp(F(\theta_1 + \gamma\theta_2) - F(\theta_1) - \gamma F(\theta_2)),$$

provided that $\theta_1 + \gamma\theta_2$ belongs to the natural parameter space (otherwise, the integral I_γ diverges).

Even when $F(\theta)$ is not known in closed form, we may estimate the γ-divergence by estimating the I_γ integrals as follows:

$$\hat{I}_\gamma(q_{\theta_1}, q_{\theta_2}) \approx \frac{1}{s}\sum_{i=1}^{s} q_2(x_i),$$

where $x_1, \ldots, x_s$ are iid. sampled from $p_1(x)$. For example, we may use Monte Carlo importance sampling methods [52] or exponential family Langevin dynamics [53] to sample densities of exponential family densities with computationally intractable normalizers (e.g., polynomial exponential families).

6. Summary and Concluding Remarks

In this paper, we first recalled the Jensen–Shannon symmetrization (JS-symmetrization) scheme of [17] for an arbitrary statistical dissimilarity $D(\cdot,\cdot)$ using an arbitrary weighted scalar mean M_α as follows:

$$D^{\mathrm{JS}}_{M_\alpha,\beta}(p_1, p_2) := \beta\, D\Big(p_1, (p_1p_2)_{M_\alpha}\Big) + (1-\beta)\, D\Big(p_2, (p_1p_2)_{M_\alpha}\Big), \quad (\alpha,\beta) \in (0,1)^2,$$

In particular, we showed that the skewed Bhattacharyya distance and the Chernoff information can both be interpreted as JS-symmetrizations of the reverse Kullback–Leibler divergence.

Then we defined two types of geometric Jensen–Shannon divergence between probability densities. The first type JS_M requires normalization of M-mixtures and relies on the Kullback–Leibler divergence: $\mathrm{JS}_M = \mathrm{KL}^{\mathrm{JS}}_{M_{\frac{1}{2}},\frac{1}{2}}$. The second type $\mathrm{JS}^+_{\tilde{M}}$ does not normalize M-mixtures and uses the extended Kullback–Leibler divergence KL^+ to take into account unnormalized mixtures: $\mathrm{JS}^+_{\tilde{M}} = \mathrm{KL}^{\mathrm{JS}^+}_{\tilde{M}_{\frac{1}{2}},\frac{1}{2}}$. When M is the arithmetic mean A, both M-JSD types coincide with the ordinary Jensen–Shannon divergence of Equation (2).

We have shown that both M-JSD types can be interpreted as regularized Jensen–Shannon divergences JS with additive terms. Namely, we have

$$\begin{aligned}
\mathrm{JS}_M(p_1, p_2) &= \mathrm{JS}(p_1, p_2) + \mathrm{KL}((p_1p_2)_A, (p_1p_2)_M), \\
\mathrm{JS}^+_{\tilde{M}}(p_1, p_2) &= \mathrm{JS}_M(p_1, p_2) + Z_M(p_1, p_2) - \log Z_M(p_1, p_2) - 1, \\
&= \mathrm{JS}(p_1, p_2) + \mathrm{KL}((p_1p_2)_A, (p_1p_2)_M) + Z_M(p_1, p_2) - \log Z_M(p_1, p_2) - 1,
\end{aligned}$$

where $Z_M(p_1, p_2) = \int M(p_1, p_2)\, \mathrm{d}\mu$ is the M-mixture normalizer. The gap between these two types of M-JSD is

$$\begin{aligned}
\Delta_M(p_1, p_2) &= \mathrm{JS}^+_{\tilde{M}}(p_1, p_2) - \mathrm{JS}_M(p_1, p_2), \\
&= Z_M(p_1, p_2) - \log Z_M(p_1, p_2) - 1.
\end{aligned}$$

When taking the geometric mean $M = G$, we showed that both G-JSD types can be expressed using the Jeffreys divergence and the Bhattacharyya divergence (or Bhattacharyya coefficient):

$$\begin{aligned} \mathrm{JS}_G(p_1, p_2) &= \frac{1}{4} J(p_1, p_2) - B(p_1, p_2), \\ \mathrm{JS}^+_{\tilde{G}}(p_1, p_2) &= \frac{1}{4} J(p_1, p_2) + \exp(-B(p_1, p_2)) - 1, \\ &= \frac{1}{4} J(p_1, p_2) + \mathrm{BC}(p_1, p_2) - 1. \end{aligned}$$

Thus, the gap between these two types of G-JSD is

$$\begin{aligned} \Delta_G(p_1, p_2) &:= \mathrm{JS}^+_{\tilde{G}}(p_1, p_2) - \mathrm{JS}_G(p_1, p_2), \\ &= \mathrm{BC}(p_1, p_2) + B(p_1, p_2) - 1, \\ &= Z_G(p_1, p_2) - \log Z_G(p_1, p_2) - 1, \end{aligned}$$

since $Z_G(p_1, p_2) = \int \sqrt{p_1\, p_2}\mathrm{d}\mu = \mathrm{BC}(p_1, p_2)$.

Although the square root of the Jensen–Shannon divergence yields a metric distance, this is no longer the case for the geometric-JSD and the extended geometric-JSD: we reported counterexamples in Remark 8. Moreover, we have shown that the KL symmetrization $\sqrt{\mathrm{KL}((p_1p_2)_A, (p_1p_2)_G)}$ is not a metric distance (Remark 12).

We discussed the merit of the extended G-JSD, which does not require normalization of the geometric mixture, in Section 5, and we showed how to approximate the G-JSD using the projective γ-divergences [39] for $\gamma = \epsilon$, a small enough value (i.e., $\gamma = \epsilon = 10^{-3}$). From the viewpoint of information geometry, the extended G-JSD has been shown to be an *f*-divergence [13] (separable divergence), while the G-JSD is not separable in general because of the normalization of mixtures (with the exception of the ordinary JSD, which is an *f*-divergence because the arithmetic mixtures do not require normalization).

We studied power JSDs by considering the power means and studied the $\pm\infty$ limits, the extended max-JSD, and the min-JSD: We proved that the extended max-JSD is upper-bounded by the total variation distance $\mathrm{TV}(p_1, p_2) = \frac{1}{2} \int |p_1 - p_2| \,\mathrm{d}\mu$:

$$0 \leq \mathrm{JS}^+_{\widetilde{\max}}(p_1, p_2) \leq \mathrm{TV}(p_1, p_2),$$

and that the extended min-JSD is lower-bounded as follows:

$$\mathrm{JS}^+_{\widetilde{\min}}(p_1, p_2) \geq \frac{1}{4} J(p_1, p_2) - \mathrm{TV}(p_1, p_2),$$

where J denotes the Jeffreys divergence: $J(p_1, p_2) = \mathrm{KL}(p_1, p_2) + \mathrm{KL}(p_2, p_1)$.

The advantage of using the extended G-JSD is that we do not need to normalize geometric mixtures, while this novel divergence is proven to be an *f*-divergence [13] and retains the property that it amounts to a regularization of the ordinary Jensen–Shannon divergence by an extra additive gap term.

Finally, we expressed JS_G (Equation (41)) and $\mathrm{JS}^+_{\tilde{G}}$ (Equation (41)) for exponential families, characterized the gap between these two types of divergences as a function of the cumulant and partition functions, and reported a corresponding explicit formula for the multivariate Gaussian (exponential) family. The G-JSD between Gaussian distributions has already been used successfully in many applications [30,32–38].

Funding: This research received no external funding.

Acknowledgments: The Author would like to thank the two Reviewers for their insightful, detailed, and constructive comments and feedback.

Conflicts of Interest: Author Frank Nielsen was employed by the company Sony Computer Science Laboratories. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Nomenclature

Means:

$M_\alpha(a,b)$	weighted scalar mean
$M_\alpha^\phi(a,b)$	weighted quasi-arithmetic scalar mean for generator $\phi(u)$
$A(a,b)$	arithmetic mean
$A_\alpha(a,b)$	weighted arithmetic mean
$G_\alpha(a,b)$	weighted geometric mean
$G(a,b)$	geometric mean
$P_\gamma(a,b)$	power mean with $P_0 = G$ and $P_1 = A$
$P_{\gamma,\alpha}(a,b)$	weighted power mean

Densities on measure space $(\mathcal{X}, \mathcal{E}, \mu)$:

$p, p_1, p_2, \ldots$	normalized density
$q, q_1, q_2, \ldots$	unnormalized density
$Z(q)$	density normalizer $p = \frac{q}{Z(q)}$
$Z_M(p_1,p_2)$	normalizer of M-mixture ($\alpha = \frac{1}{2}$)
$\hat{Z}_M(p_1,p_2)$	Monte Carlo estimator of $Z_M(p_1,p_2)$
$Z_{M,\alpha}(p_1,p_2)$	normalizer of weighted M-mixture
$(p_1p_2)_M$	M-mixture
$(p_1p_2)_{M,\alpha}$	weighted M-mixture

Dissimilarities, divergences, and distances:

$\mathrm{KL}(p_1,p_2)$	Kullback–Leibler divergence (KLD)
$\mathrm{KL}^+(q_1,q_2)$	extended Kullback–Leibler divergence
$\mathrm{KL}^*(p_1,p_2)$	reverse Kullback–Leibler divergence
$H^\times(p_1,p_2)$	cross-entropy
$H(p)$	Shannon discrete or differential entropy
$J(p_1,p_2)$	Jeffreys divergence
$\mathrm{TV}(p_1,p_2)$	total variation distance
$B(p_1,p_2)$	Bhattacharyya "distance" (not metric)
$B_\alpha(p_1,p_2)$	α-skewed Bhattacharyya "distance"
$C(p_1,p_2)$	Chernoff information or Chernoff distance
$T(p_1,p_2)$	Taneja T-divergence
$I_f(p_1,p_2)$	Ali–Silvey–Csiszár f-divergence
$D(p_1,p_2)$	arbitrary dissimilarity measure
$D^*(p_1,p_2)$	reverse dissimilarity measure
$D^+(q_1,q_2)$	extended dissimilarity measure
$\tilde{D}(q_1,q_2)$	projective dissimilarity measure
$\tilde{D}_\gamma(q_1,q_2)$	γ-divergence
$\hat{D}^+(q_1,q_2)$	Monte Carlo estimation of dissimilarity D^+

Jensen–Shannon divergences and generalizations:

$\mathrm{JS}(p_1,p_2)$	Jensen–Shannon divergence (JSD)
$\mathrm{JS}_{\alpha,\beta}(p_1,p_2)$	β-weighted α-skewed mixture JSD
$\mathrm{JS}_M(p_1,p_2)$	M-JSD for M-mixtures

$\mathrm{JS}_G(p_1, p_2)$	geometric JSD
$\mathrm{JS}_{\tilde{G}}(p_1, p_2)$	extended geometric JSD
$\mathrm{JS}^*_G(p_1, p_2)$	left-sided geometric JSD (right-sided for KL^*)
$\mathrm{JS}^+_{\widetilde{\min}}(p_1, p_2)$	min-JSD
$\mathrm{JS}^+_{\widetilde{\max}}(p_1, p_2)$	max-JSD
$\Delta_M(p_1, p_2)$	gap between extended and normalized M-JSDs

References

1. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
2. Jeffreys, H. *The Theory of Probability*; OuP Oxford: Oxford, UK, 1998.
3. Johnson, D.H.; Sinanovic, S. Symmetrizing the Kullback-Leibler distance. *IEEE Trans. Inf. Theory* **2001**, *1*, 1–10.
4. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
5. Fuglede, B.; Topsoe, F. Jensen-Shannon divergence and Hilbert space embedding. In Proceedings of the International Symposium on Information Theory (ISIT), Chicago, IL, USA, 27 June–2 July 2004; p. 31.
6. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [CrossRef]
7. Okamura, K. Metrization of powers of the Jensen-Shannon divergence. *arXiv* **2023**, arXiv:2302.10070. [CrossRef]
8. Sibson, R. Information radius. *Z. FÜR Wahrscheinlichkeitstheorie Und Verwandte Geb.* **1969**, *14*, 149–160. [CrossRef]
9. Briët, J.; Harremoës, P. Properties of classical and quantum Jensen-Shannon divergence. *Phys. Rev. A* **2009**, *79*, 052311. [CrossRef]
10. Virosztek, D. The metric property of the quantum Jensen-Shannon divergence. *Adv. Math.* **2021**, *380*, 107595. [CrossRef]
11. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. Methodol.* **1966**, *28*, 131–142. [CrossRef]
12. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
13. Amari, S.i. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Tokyo, Japan, 2016.
14. Csiszár, I.; Shields, P.C. *Information Theory and Statistics: A Tutorial. (Foundations and Trends® in Communications and Information Theory)*; Now Publishers Inc.: Hanover, MA, USA, 2004; Volume 1, pp. 417–528.
15. Osterreicher, F.; Vajda, I. A new class of metric divergences on probability spaces and its applicability in statistics. *Ann. Inst. Stat. Math.* **2003**, *55*, 639–653. [CrossRef]
16. Schoenberg, I.J. Metric spaces and completely monotone functions. *Ann. Math.* **1938**, *39*, 811–841. [CrossRef]
17. Nielsen, F. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485. [CrossRef]
18. Bullen, P.S. *Handbook of Means and Their Inequalities*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 560.
19. Yamano, T. Some bounds for skewed α-Jensen-Shannon divergence. *Results Appl. Math.* **2019**, *3*, 100064. [CrossRef]
20. Nielsen, F. Revisiting Chernoff information with likelihood ratio exponential families. *Entropy* **2022**, *24*, 1400. [CrossRef]
21. Jerfel, G.; Wang, S.; Wong-Fannjiang, C.; Heller, K.A.; Ma, Y.; Jordan, M.I. Variational refinement for importance sampling using the forward Kullback-Leibler divergence. In Proceedings of the Uncertainty in Artificial Intelligence, PMLR, Online, 27–30 July 2021; pp. 1819–1829.
22. Asadi, M.; Ebrahimi, N.; Kharazmi, O.; Soofi, E.S. Mixture models, Bayes Fisher information, and divergence measures. *IEEE Trans. Inf. Theory* **2018**, *65*, 2316–2321. [CrossRef]
23. Grosse, R.B.; Maddison, C.J.; Salakhutdinov, R.R. Annealing between distributions by averaging moments. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–12.
24. Amari, S.I. Integration of stochastic models by minimizing α-divergence. *Neural Comput.* **2007**, *19*, 2780–2796. [CrossRef]
25. Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhyā Indian J. Stat.* **1946**, *7*, 401–406.
26. Melville, P.; Yang, S.M.; Saar-Tsechansky, M.; Mooney, R. Active learning for probability estimation using Jensen-Shannon divergence. In Proceedings of the European Conference on Machine Learning, Porto, Portugal, 3–7 October 2005; pp. 268–279.
27. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
28. Sutter, T.; Daunhawer, I.; Vogt, J. Multimodal generative learning utilizing Jensen-Shannon-divergence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6100–6110.
29. Michalowicz, J.V.; Nichols, J.M.; Bucholtz, F. Calculation of differential entropy for a mixed Gaussian distribution. *Entropy* **2008**, *10*, 200–206. [CrossRef]
30. Deasy, J.; Simidjievski, N.; Liò, P. Constraining variational inference with geometric Jensen-Shannon divergence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 10647–10658.

31. Deasy, J.; McIver, T.A.; Simidjievski, N.; Lio, P. α-VAEs: Optimising variational inference by learning data-dependent divergence skew. In Proceedings of the ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models, Virtual, 23 July 2021.
32. Kumari, J.; Deepak, G.; Santhanavijayan, A. RDS: Related document search for economics data using ontologies and hybrid semantics. In Proceedings of the International Conference on Data Analytics and Insights, Kolkata, India, 11–13 May 2023; pp. 691–702.
33. Ni, S.; Lin, C.; Wang, H.; Li, Y.; Liao, Y.; Li, N. Learning geometric Jensen-Shannon divergence for tiny object detection in remote sensing images. *Front. Neurorobot.* **2023**, *17*, 1273251. [CrossRef]
34. Sachdeva, R.; Gakhar, R.; Awasthi, S.; Singh, K.; Pandey, A.; Parihar, A.S. Uncertainty and Noise Aware Decision Making for Autonomous Vehicles—A Bayesian Approach. *IEEE Trans. Veh. Technol.* **2024**, *74*, 378–389. [CrossRef]
35. Wang, J.; Massiceti, D.; Hu, X.; Pavlovic, V.; Lukasiewicz, T. NP-SemiSeg: When neural processes meet semi-supervised semantic segmentation. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 36138–36156.
36. Serra, G.; Stavrou, P.A.; Kountouris, M. On the computation of the Gaussian rate–distortion–perception function. *IEEE J. Sel. Areas Inf. Theory* **2024**, *5*, 314–330. [CrossRef]
37. Thiagarajan, P.; Ghosh, S. Jensen–Shannon divergence based novel loss functions for Bayesian neural networks. *Neurocomputing* **2025**, *618*, 129115. [CrossRef]
38. Hanselmann, N.; Doll, S.; Cordts, M.; Lensch, H.P.; Geiger, A. EMPERROR: A Flexible Generative Perception Error Model for Probing Self-Driving Planners. *IEEE Robot. Autom. Lett.* **2025**, *10*, 5807–5814. [CrossRef]
39. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081. [CrossRef]
40. Jones, L.K.; Byrne, C.L. General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Trans. Inf. Theory* **2002**, *36*, 23–30. [CrossRef]
41. Nishimura, T.; Komaki, F. The information geometric structure of generalized empirical likelihood estimators. *Commun. Stat. Methods* **2008**, *37*, 1867–1879. [CrossRef]
42. Nielsen, F. Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. *Pattern Recognit. Lett.* **2014**, *42*, 25–34. [CrossRef]
43. Barndorff-Nielsen, O. *Information and Exponential Families: In Statistical Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
44. Kailath, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [CrossRef]
45. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [CrossRef]
46. Grünwald, P.D. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA USA, 2007.
47. Cena, A.; Pistone, G. Exponential statistical manifold. *Ann. Inst. Stat. Math.* **2007**, *59*, 27–56. [CrossRef]
48. Taneja, I.J. New developments in generalized information measures. In *Advances in Imaging and Electron Physics*; Elsevier: Amsterdam, The Netherlands, 1995; Volume 91, pp. 37–135.
49. Rubinstein, R.Y.; Kroese, D.P. *Simulation and the Monte Carlo Method*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
50. Cobb, L.; Koppstein, P.; Chen, N.H. Estimation and moment recursion relations for multimodal distributions of the exponential family. *J. Am. Stat. Assoc.* **1983**, *78*, 124–130. [CrossRef]
51. Hayakawa, J.; Takemura, A. Estimation of exponential-polynomial distribution by holonomic gradient descent. *Commun. Stat.-Theory Methods* **2016**, *45*, 6860–6882. [CrossRef]
52. Kloek, T.; Van Dijk, H.K. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econom. J. Econom. Soc.* **1978**, *46*, 1–19. [CrossRef]
53. Banerjee, A.; Chen, T.; Li, X.; Zhou, Y. Stability based generalization bounds for exponential family Langevin dynamics. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MA, USA, 17–23 July 2022; pp. 1412–1449.