entropy

MDPI

Article

α-Geodesical Skew Divergence

Masanari Kimura [1,*] **and Hideitsu Hino** [2]

[1] Department of Statistical Science, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies (SOKENDAI), Kanagawa 240-0193, Japan
[2] The Institute of Statistical Mathematics, Tokyo 190-0014, Japan; hino@ism.ac.jp
* Correspondence: mkimura@ism.ac.jp

Abstract: The asymmetric skew divergence smooths one of the distributions by mixing it, to a degree determined by the parameter λ, with the other distribution. Such divergence is an approximation of the KL divergence that does not require the target distribution to be absolutely continuous with respect to the source distribution. In this paper, an information geometric generalization of the skew divergence called the α-geodesical skew divergence is proposed, and its properties are studied.

Keywords: KL-divergence; JS-divergence; skew divergence; information geometry

check for updates

Citation: Kimura, M.; Hino, H. α-Geodesical Skew Divergence. *Entropy* **2021**, *23*, 528. https://doi.org/10.3390/e23050528

Academic Editor: Frank Nielsen

Received: 1 April 2021
Accepted: 24 April 2021
Published: 25 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space where $\mathcal{X}$ denotes the sample space, $\mathcal{F}$ the σ-algebra of measurable events, and μ a positive measure. The set of the strictly positive probability measure $\mathcal{P}$ is defined as

$$\mathcal{P} := \left\{ f(x) > 0 \ (\forall x \in \mathcal{X}), \text{ and } \int_{\mathcal{X}} f(x) d\mu(x) = 1 \right\}, \tag{1}$$

and the set of nonnegative probability measure $\mathcal{P}_+$ is defined as

$$\mathcal{P}_+ := \left\{ f(x) \geq 0 \ (\forall x \in \mathcal{X}), \text{ and } \int_{\mathcal{X}} f(x) d\mu(x) = 1 \right\}. \tag{2}$$

Then a number of divergences that appear in statistics and information theory [1,2] are introduced.

Definition 1 (Kullback–Leibler divergence [3]). *The Kullback–Leibler divergence or KL-divergence $D_{KL} : \mathcal{P}_+ \times \mathcal{P} \to [0, \infty]$ is defined between two Radon–Nikodym densities p and q of μ-absolutely continuous probability measures by*

$$D_{KL}[p \| q] := \int_{\mathcal{X}} p \ln \frac{p}{q} d\mu. \tag{3}$$

KL-divergence is a measure of the difference between two probability distributions in statistics and information theory [4–7]. This is also called the relative entropy and is known not to satisfy the axiom of distance. Because the KL-divergence is asymmetric, several symmetrizations have been proposed in the literature [8–10].

Definition 2 (Jensen–Shannon divergence [8]). *The Jensen–Shannon divergence or JS-divergence $D_{JS} : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ is defined between two Radon–Nikodym densities p and q of μ-absolutely continuous probability measures by*

$$\begin{aligned} D_{JS}[p\|q] &:= \frac{1}{2}\left(D_{KL}\left[p\middle\|\frac{p+q}{2}\right]+D_{KL}\left[q\middle\|\frac{p+q}{2}\right]\right) \\ &= \frac{1}{2}\int_{\mathcal{X}}\left(p\ln\frac{2p}{p+q}+q\ln\frac{2q}{p+q}\right)d\mu \\ &= D_{JS}[q\|p]. \end{aligned} \tag{4}$$

The JS-divergence is a symmetrized and smoothed version of the KL-divergence, and it is bounded as

$$0 \le D_{JS}[p\|q] \le \ln 2. \tag{5}$$

This property contrasts with the fact that KL-divergence is unbounded.

Definition 3 (Jeffreys divergence [11]). *The Jeffreys divergence $D_J[p\|q] : \mathcal{P}\times\mathcal{P} \to [0,\infty]$ is defined between two Radon–Nikodym densities p and q of μ-absolutely continuous probability measures by*

$$D_J[p\|q] := D_{KL}[p\|q] + D_{KL}[q\|p]. \tag{6}$$

Such symmetrized KL-divergences have appeared in various pieces of literature [12–18].

For continuous distributions, the KL-divergence is known to have computational difficulty. To be more specific, if q takes a small value relative to p, the value of $D_{KL}[p\|q]$ may diverge to infinity. The simplest idea to avoid this is to use very small $\epsilon > 0$ and modify $D_{KL}[p\|q]$ as follows:

$$D^{+}_{KL}[p\|q] := \int_{\mathcal{X}} p\ln\frac{p}{q+\epsilon}d\mu.$$

However, such an extension is unnatural in the sense that $q+\epsilon$ no longer satisfies the condition for a probability measure: $\int_{\mathcal{X}} \epsilon + q(x)d\mu(x) \neq 1$. As a more natural way to stabilize KL-divergence, the following skew divergences have been proposed:

Definition 4 (Skew divergence [8,19]). *The skew divergence $D^{(\lambda)}_S[p\|q] : \mathcal{P}\times\mathcal{P} \to [0,\infty]$ is defined between two Radon–Nikodym densities p and q of μ-absolutely continuous probability measures by*

$$\begin{aligned} D^{(\lambda)}_S[p\|q] &:= D_{KL}[p\|(1-\lambda)p+\lambda q] \\ &= \int_{\mathcal{X}} p\ln\frac{p}{(1-\lambda)p+\lambda q}d\mu, \end{aligned} \tag{7}$$

where $\lambda \in [0,1]$.

Skew divergences have been experimentally shown to perform better in applications such as natural language processing [20,21], image recognition [22,23] and graph analysis [24,25]. In addition, there is research on quantum generalization of skew divergence [26].

The main contributions of this paper are summarized as follows:

- Several symmetrized divergences or skew divergences are generalized from an information geometry perspective.
- It is proved that the natural skew divergence for the exponential family is equivalent to the scaled KL-divergence.
- Several properties of geometrically generalized skew divergence are proved. Specifically, the functional space associated with the proposed divergence is shown to be a Banach space.

Implementation of the proposed divergence is available on GitHub (https://github.com/nocotan/geodesical_skew_divergence (accessed on 3 April 2021)).

2. α-Geodesical Skew Divergence

The skew divergence is generalized based on the following function.

Definition 5 (f-interpolation). *For any $a, b, \in \mathbb{R}$, $\lambda \in [0,1]$ and $\alpha \in \mathbb{R}$, f-interpolation is defined as*

$$m_f^{(\lambda,\alpha)}(a,b) = f_\alpha^{-1}\Big((1-\lambda)f_\alpha(a) + \lambda f_\alpha(b)\Big), \tag{8}$$

where

$$f_\alpha(x) = \begin{cases} x^{\frac{1-\alpha}{2}} & (\alpha \neq 1) \\ \ln x & (\alpha = 1) \end{cases} \tag{9}$$

is the function that defines the f-mean [27].

The f-mean function satisfies

$$\lim_{\alpha\to\infty} f_\alpha(x) = \begin{cases} \infty & (|x| < 1), \\ 1 & (|x| = 1), \\ 0 & (|x| > 1), \end{cases}$$

$$\lim_{\alpha\to-\infty} f_\alpha(x) = \begin{cases} 0 & (|x| < 1), \\ 1 & (|x| = 1), \\ \infty & (|x| > 1). \end{cases}$$

It is easy to see that this family includes various known weighted means including the e-mixture and m-mixture for $\alpha = \pm 1$ in the literature of information geometry [28]:

$$\begin{aligned}
(\alpha = 1) \quad & m_f^{(\lambda,1)}(a,b) = \exp\{(1-\lambda)\ln a + \lambda \ln b\} \\
(\alpha = -1) \quad & m_f^{(\lambda,-1)}(a,b) = (1-\lambda)a + \lambda b \\
(\alpha = 0) \quad & m_f^{(\lambda,0)}(a,b) = \Big((1-\lambda)\sqrt{a} + \lambda\sqrt{b}\Big)^2 \\
(\alpha = 3) \quad & m_f^{(\lambda,3)}(a,b) = \frac{1}{(1-\lambda)\frac{1}{a} + \lambda\frac{1}{b}} \\
(\alpha = \infty) \quad & m_f^{(\lambda,\infty)}(a,b) = \min\{a,b\} \\
(\alpha = -\infty) \quad & m_f^{(\lambda,-\infty)}(a,b) = \max\{a,b\}
\end{aligned}$$

The inverse function f_α^{-1} is convex when $\alpha \in [-1,1]$, and concave when $\alpha \in (-\infty,-1] \cup (1,\infty)$. It is worth noting that the f-interpolation is a special case of the Kolmogorov–Nagumo average [29–31] when α is restricted in the interval $[-1,1]$.

In order to consider the geometric meaning of this function, the notion of the statistical manifold is introduced.

2.1. Statistical Manifold

Let

$$\mathcal{S} = \{p_\xi = p(x;\xi) \in \mathcal{P} | \xi = (\xi^1,\ldots,\xi^n) \in \Xi\} \tag{10}$$

be a family of probability distribution on $\mathcal{X}$, where each element p_ξ is parameterized by n real-valued variables $\xi = (\xi^1,\ldots,\xi^n) \in \Xi \subset \mathbb{R}^n$. The set $\mathcal{S}$ is called a statistical model and is a subset of $\mathcal{P}$. We also denote $(\mathcal{S}, g_{ij})$ as a statistical model equipped with the Riemannian metric g_{ij}. In particular, let g_{ij} be the Fisher–Rao metric, which is the Riemannian metric induced from the Fisher information matrix [32].

In the rest of this paper, the abbreviations

$$\partial_i = \partial_{\xi^i} = \frac{\partial}{\partial \xi^i},$$
$$\ell = \ell_{\boldsymbol{x}}(\boldsymbol{\xi}) = \ln p_{\boldsymbol{\xi}}(\boldsymbol{x})$$

are used.

Definition 6 (Christoffel symbols)**.** *Let g_{ij} be a Riemannian metric, particularly the Fisher information matrix, then the Christoffel symbols are given by*

$$\Gamma_{ij,k} = \frac{1}{2}\Big(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\Big), \quad i,j,k = 1,\ldots,n. \tag{11}$$

Definition 7 (Levi-Civita connection)**.** *Let g be a Fisher–Riemannian metric on $\mathcal{S}$ which is a 2-covariant tensor defined locally by*

$$g(X_{\boldsymbol{\xi}}, Y_{\boldsymbol{\xi}}) = \sum_{i,j=1}^{n} g_{ij}(\boldsymbol{\xi}) a^i(\boldsymbol{\xi}) b^j(\boldsymbol{\xi}),$$

where $X_{\boldsymbol{\xi}} = \sum_{i=1}^n a^i(\boldsymbol{\xi})\partial_i p_{\boldsymbol{\xi}}$ and $Y_{\boldsymbol{\xi}} = \sum_{i=1}^n b^i(\boldsymbol{\xi})\partial_i p_{\boldsymbol{\xi}}$ are vector fields in the 0-representation on $\mathcal{S}$. Then, its associated Levi-Civita connection $\nabla^{(0)}$ is defined by

$$g(\nabla^{(0)}_{\partial_i}\partial_j, \partial_k) = \Gamma_{ij,k}. \tag{12}$$

The fact that $\nabla^{(0)}$ is a metrical connection can be written locally as

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}. \tag{13}$$

It is worth noting that the superscript α of $\nabla^{(\alpha)}$ corresponds to a parameter of the connection. Based on the above definitions, several connections parameterized by the parameter α are introduced. The case $\alpha = 0$ corresponds to the Levi-Civita connection induced by the Fisher metric.

Definition 8 ($\nabla^{(1)}$-connection)**.** *Let g be the Fisher-Riemannian metric on $\mathcal{S}$, which is a 2-covariant tensor. Then, the $\nabla^{(1)}$-connection is defined by*

$$g(\nabla^{(1)}_{\partial_i}\partial_j, \partial_k) = \mathbb{E}_{\boldsymbol{\xi}}[\partial_i\partial_j\ell\partial_k\ell]. \tag{14}$$

It can also be expressed equivalently by explicitly writing as the Christoffel coefficients

$$\Gamma^{(1)}_{ij,k}(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\xi}}[\partial_i\partial_j\ell\partial_k\ell]. \tag{15}$$

Definition 9 ($\nabla^{(-1)}$-connection)**.** *Let g be the Fisher–Riemannian metric on $\mathcal{S}$, which is a 2-covariant tensor. Then, the $\nabla^{(-1)}$-connection is defined by*

$$g(\nabla^{(-1)}_{\partial_i}\partial_j, \partial_k) = \Gamma^{(-1)}_{ij,k}(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\xi}}[(\partial_i\partial_j\ell + \partial_i\ell\partial_j\ell)\partial_k\ell]. \tag{16}$$

In the following, the ∇-flatness is considered with respect to the corresponding coordinates system. More details can be found in [28].

Proposition 1. *The exponential family is $\nabla^{(1)}$-flat.*

Proposition 2. *The exponential family is $\nabla^{(-1)}$-flat if and only if it is $\nabla^{(0)}$-flat.*

Proposition 3. *The mixture family is $\nabla^{(-1)}$-flat.*

Proposition 4. *The mixture family is $\nabla^{(1)}$-flat if and only if it is $\nabla^{(0)}$-flat.*

Proposition 5. *The relation between the foregoing three connections is given by*

$$\nabla^{(0)} = \frac{1}{2}\Big(\nabla^{(-1)} + \nabla^{(1)}\Big). \tag{17}$$

Proof. It suffices to show

$$\Gamma^{(0)}_{ij,k} = \frac{1}{2}\Big(\Gamma^{(-1)}_{ij,k} + \Gamma^{(1)}_{ij,k}\Big).$$

From the definitions of $\Gamma^{(-1)}$ and $\Gamma^{(1)}$,

$$\begin{aligned}\Gamma^{(-1)}_{ij,k} + \Gamma^{(1)}_{ij,k} &= \mathbb{E}_{\xi}[(\partial_i\partial_j\ell + \partial_i\ell\partial_j\ell)\partial_k\ell] + \mathbb{E}_{\xi}[\partial_i\partial_j\ell\partial_k\ell] \\ &= \mathbb{E}_{\xi}[(2\partial_i\partial_j\ell + \partial_i\ell\partial_j\ell)\partial_k\ell] \\ &= 2\mathbb{E}_{\xi}\Big[(\partial_i\partial_j\ell + \frac{1}{2}\partial_i\ell\partial_j\ell)\partial_k\ell\Big] \\ &= 2\Gamma^{(0)}_{ij,k},\end{aligned}$$

which proves the proposition. □

The connections $\nabla^{(-1)}$ and $\nabla^{(1)}$ are two special connections on $\mathcal{S}$ with respect to the mixture family and the exponential family, respectively. Moreover, they are related by the duality condition, and the following 1-parameter family of connections is defined.

Definition 10 ($\nabla^{(\alpha)}$-connection)**.** *For $\alpha \in \mathbb{R}$, the $\nabla^{(\alpha)}$-connection on the statistical model $\mathcal{S}$ is defined as*

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}\nabla^{(-1)}. \tag{18}$$

Proposition 6. *The components $\Gamma^{(\alpha)}_{ij,k}$ can be written as*

$$\Gamma^{(\alpha)}_{ij,k} = \mathbb{E}_{\xi}\left[\Big(\partial_i\partial_j\ell + \frac{1-\alpha}{2}\partial_i\ell\partial_j\ell\Big)\partial_k\ell\right]. \tag{19}$$

The α-coordinate system associated with the $\nabla^{(\alpha)}$-connection is endowed with the α-geodesic, which is a straight line on the corresponding coordinates system. Then, we introduce some relevant notions.

Definition 11 (α-divergence [33])**.** *Let α be a real parameter. The α-divergence between two probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ is defined as*

$$D_\alpha[\boldsymbol{p}\|\boldsymbol{q}] = \frac{4}{1-\alpha^2}\Big(1 - \sum_i p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}}\Big). \tag{20}$$

The KL-divergence, which is a special case with $\alpha = 1$, induces the linear connection $\nabla^{(1)}$ as follows.

Proposition 7. *The diagonal part of the third mixed derivatives of the KL-divergence is the negative of the Christoffel symbol:*

$$-\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi_0^k} D_{KL}[p_{\xi_0}\|p_\xi]\Big|_{\xi=\xi_0} = \Gamma^{(1)}_{ij,k}(\xi_0). \tag{21}$$

Proof. The second derivative in the argument $\boldsymbol{\xi}$ is given by

$$\partial_{\xi^i}\partial_{\xi^j} D_{KL}[p_{\xi_0}\|p_\xi] = -\int_{\mathcal{X}} p_{\xi_0}(\boldsymbol{x})\partial_{\xi^i}\partial_{\xi^j}\ell_{\boldsymbol{x}}(\boldsymbol{\xi})d\boldsymbol{x},$$

and differentiating it with respect to $\boldsymbol{\xi}_0^k$ yields

$$\begin{aligned}-\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi_0^k} D_{KL}[p_{\xi_0}\|p_\xi] &= \partial_{\xi_0^k}\int_{\mathcal{X}} p_{\xi_0}(\boldsymbol{x})\partial_{\xi^i}\partial_{\xi^j}\ell_{\boldsymbol{x}}(\boldsymbol{\xi})d\boldsymbol{x}\\ &= \int_{\mathcal{X}} p_{\xi_0}(\boldsymbol{x})\partial_{\xi^i}\partial_{\xi^j}\ell_{\boldsymbol{x}}(\boldsymbol{\xi})\partial_{\xi_0^k}\ell_{\boldsymbol{x}}(\boldsymbol{\xi})d\boldsymbol{x}.\end{aligned}$$

Then, considering the diagonal part, one yields

$$\begin{aligned}-\partial_{\xi^i}\partial_{\xi^j}\partial_{\xi_0^k} D_{KL}[p_{\xi_0}\|p_\xi]\Big|_{\xi=\xi_0} &= \mathbb{E}_{\xi_0}[\partial_i\partial_j\ell(\boldsymbol{\xi})\partial_k\ell(\boldsymbol{\xi})]\\ &= \Gamma^{(1)}_{ij,k}(\boldsymbol{\xi}_0).\end{aligned}$$

□

More generally, the α-divergence with $\alpha \in \mathbb{R}$ induces the $\nabla^{(\alpha)}$-connection.

Definition 12 (α-representation [34]). *For some positive measure $m_i^{\frac{1-\alpha}{2}}$, the coordinate system $\boldsymbol{\theta} = (\theta^i)$ derived from the α-divergence is*

$$\theta^i = m_i^{\frac{1-\alpha}{2}} = f_\alpha(m_i) \tag{22}$$

and θ^i is called the α-representation of a positive measure $m_i^{\frac{1-\alpha}{2}}$.

Definition 13 (α-geodesic [28]). *The α-geodesic connecting two probability vectors $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is defined as*

$$r_i(t) = c(t) f_\alpha^{-1}\Big\{(1-t)f_\alpha(p(x_i)) + t f_\alpha(q(x_i))\Big\}, \quad t \in [0,1] \tag{23}$$

where $c(t)$ is determined as

$$c(t) = \frac{1}{\sum_{i=1}^n r_i(t)}. \tag{24}$$

It is known that the appropriate reparameterizations for the parameter t are necessary for a rigorous discussion in the space of probability measures [35,36]. However, as mentioned in the literature [35], an explicit expression for the reparametrizations $\tau_{p,a}$ and $\tau_{p,q}$ is unknown. A similar discussion has been made in the derivation of the ϕ_β-path [37], where it is mentioned that the normalizing factor is unknown in general. Furthermore, the f-mean is not convex depending on the α. For these reasons, it is generally difficult to discuss α-geodesics in probability measures by normalization or reparameterization, and to avoid unnecessary complexity, the parameter t is assumed to be appropriately reparameterized.

Let $\psi_\alpha(\boldsymbol{\theta}) = \frac{1-\alpha}{2}\sum_{i=1}^n m_i$. Then, the dual coordinate system $\boldsymbol{\eta}$ is given by $\boldsymbol{\eta} = \nabla\psi_\alpha(\boldsymbol{\theta})$ as

$$\eta_i = (\theta^i)^{\frac{1+\alpha}{1-\alpha}} = f_{-\alpha}(m_i). \tag{25}$$

Hence, it is the $(-\alpha)$-representation of m_i.

2.2. Generalization of Skew Divergences

From Definition 13, the f-interpoloation is considered as an unnormalized version of the α-geodesic. Using the notion of geodesics, skew divergence is generalized in terms of information geometry as follows.

Definition 14 (α-Geodesical Skew Divergence). *The α-geodesical skew divergence $D_{GS}^{(\alpha,\lambda)} : \mathcal{P} \times \mathcal{P} \to [0,\infty]$ is defined between two Radon–Nikodym densities p and q of μ-absolutely continuous probability measures by:*

$$\begin{aligned} D_{GS}^{(\alpha,\lambda)}\Big[p\|q\Big] &:= D_{KL}\Big[p\|m_f^{(\lambda,\alpha)}(p,q)\Big] \\ &= \int_{\mathcal{X}} p \ln \frac{p}{m_f^{(\lambda,\alpha)}(p,q)} d\mu, \end{aligned} \tag{26}$$

where $\alpha \in \mathbb{R}$ and $\lambda \in [0,1]$.

Some special cases of α-geodesical skew divergence are listed below:

$$\begin{aligned} (\forall \alpha \in \mathbb{R}, \lambda = 1)\ & D_{GS}^{(\alpha,1)}[p\|q] = D_{KL}[p\|q] \\ (\forall \alpha \in \mathbb{R}, \lambda = 0)\ & D_{GS}^{(\alpha,0)}[p\|q] = D_{KL}[p\|p] = 0 \\ (\alpha = 1, \forall \lambda \in [0,1])\ & D_{GS}^{(1,\lambda)}[p\|q] = \lambda D_{KL}[p\|q] \ \text{ (scaled KL-divergence)} \\ (\alpha = -1, \forall \lambda \in [0,1])\ & D_{GS}^{(-1,\lambda)}[p\|q] = D_S^{(\lambda)}[p\|q] \ \text{ (skew divergence)} \\ (\alpha = 0, \forall \lambda \in [0,1])\ & D_{GS}^{(0,\lambda)}[p\|q] = \int_{\mathcal{X}} p \ln \frac{p}{\{(1-\lambda)\sqrt{p} + \lambda\sqrt{q}\}^2} d\mu \\ (\alpha = 3, \forall \lambda \in [0,1])\ & D_{GS}^{(3,\lambda)}[p\|q] = D_S^{(\lambda)}[p\|q] + H(p) + H(q) \\ (\alpha = \infty, \forall \lambda \in [0,1])\ & D_{GS}^{(\infty,\lambda)}[p\|q] = \int_{\mathcal{X}} p \ln \frac{p}{\min\{p,q\}} d\mu \\ (\alpha = -\infty, \forall \lambda \in [0,1])\ & D_{GS}^{(-\infty,\lambda)}[p\|q] = \int_{\mathcal{X}} p \ln \frac{p}{\max\{p,q\}} d\mu \end{aligned}$$

Furthermore, α-geodesical skew divergence is a special form of the generalized skew K-divergence [10,38], which is a family of abstract means-based divergences. In this paper, the skew K-divergence touched upon in [10] is characterized in terms of α-geodesic on positive measures, and its geometric and functional analytic properties are investigated. When the Kolmogorov–Nagumo average (i.e., when the function f^{-1} in Equation (8) is a strictly monotone convex function) the geodesic has been shown to be well-defined [37].

2.3. Symmetrization of α-Geodesical Skew Divergence

It is easy to symmetrize the α-geodesical skew divergence as follows.

Definition 15 (Symmetrized α-Geodesical Skew Divergence). *The symmetrized α-geodesical skew divergence $\bar{D}_{GS}^{(\alpha,\lambda)} : \mathcal{P} \times \mathcal{P} \to [0,\infty]$ is defined between two Radon–Nikodym densities p and q of μ-absolutely continuous probability measures by:*

$$\bar{D}_{GS}^{(\alpha,\lambda)}[p\|q] := \frac{1}{2}\left(D_{GS}^{(\alpha,\lambda)}[p\|q] + D_{GS}^{(\alpha,\lambda)}[q\|p] \right), \tag{27}$$

where $\alpha \in \mathbb{R}$ and $\lambda \in [0,1]$.

It is seen that $\bar{D}_{GS}^{(\alpha,\lambda)}[p\|q]$ includes several symmetrized divergences.

$$\bar{D}_{GS}^{(\alpha,1)}[p\|q] = \frac{1}{2}\left(D_{KL}[p\|q] + D_{KL}[q\|p]\right), \quad \text{(half of Jeffreys divergence)}$$

$$\bar{D}_{GS}^{(-1,\frac{1}{2})}[p\|q] = \frac{1}{2}\left(D_{KL}\left[p\|\frac{p+q}{2}\right] + D_{KL}\left[q\|\frac{p+q}{2}\right]\right), \quad \text{(JS-divergence)}$$

$$\bar{D}_{GS}^{(-1,\lambda)}[p\|q] = \frac{1}{2}\left(D_{KL}\left[p\|(1-\lambda)p+\lambda q\right] + D_{KL}\left[q\|(1-\lambda)q+\lambda p\right]\right).$$

The last one is the λ-JS-divergence [39], which is a generalization of the JS-divergence.

3. Properties of α-Geodesical Skew Divergence

In this section, the properties of the α-geodesical skew divergence are studied.

Proposition 8 (Non-negativity of the α-geodesical skew divergence)**.** *For $\alpha \geq -1$ and $\lambda \in [0,1]$, the α-geodesical skew divergence $D_{GS}^{(\alpha,\lambda)}[p\|q]$ satisfies the following inequality:*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \geq 0. \tag{28}$$

Proof. When λ is fixed, the f-interpolation has the following inverse monotonicity with respect to α:

$$m_f^{(\lambda,\alpha)}(p,q) \geq m_f^{(\lambda,\alpha')}(p,q), \ (\alpha \leq \alpha'). \tag{29}$$

From Gibbs' inequality [40] and Equation (29), one obtains

$$\begin{aligned} D_{GS}^{(\alpha,\lambda)}[p\|q] &= \int_{\mathcal{X}} p \ln \frac{p}{m_f^{(\alpha,\lambda)}(p,q)} d\mu \\ &\geq \left(\int_{\mathcal{X}} p d\mu\right) \ln \frac{p}{m_f^{(\alpha,\lambda)}(p,q)} \\ &\geq 1 \cdot \ln 1 = 0. \end{aligned}$$

□

Proposition 9 (Asymmetry of the α-geodesical skew divergence)**.** *α-Geodesical skew divergence is not symmetric in general:*

$$D_{GS}^{(\alpha,\lambda)}[p|q] \neq D_{GS}^{(\alpha,\lambda)}[q\|p]. \tag{30}$$

Proof. For example, if $\lambda = 1$, then $\forall \alpha \in \mathbb{R}$, it holds that

$$D_{GS}^{(\alpha,1)}[p\|q] - D_{GS}^{(\alpha,1)}[q\|p] = D_{KL}[p\|q] - D_{KL}[q\|p],$$

and the asymmetry of the KL-divergence results in an asymmetry of the geodesic skew divergence. □

When a function $f(x)$ of $x \in [0,1]$ satisfies $f(x) = f(1-x)$, it is referred to as centrosymmetric.

Proposition 10 (Non-centrosymmetricity of the α-geodesical skew divergence with respect to λ)**.** *α-Geodesical skew divergence is not centrosymmetric in general with respect to the parameter $\lambda \in [0,1]$:*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \neq D_{GS}^{(\alpha,1-\lambda)}[p\|q]. \tag{31}$$

Proof. For example, if $\lambda = 1$, then $\forall \alpha \in \mathbb{R}$, we have

$$\begin{aligned} D_{GS}^{(\alpha,\lambda)}[p\|q] - D_{GS}^{(\alpha,1-\lambda)}[p\|q] &= D_{GS}^{(\alpha,1)}[p\|q] - D_{GS}^{(\alpha,0)}[p\|q] \\ &= \int_{\mathcal{X}} p \ln \frac{p}{q} - \int_{\mathcal{X}} p \ln \frac{p}{p} \\ &= \int_{\mathcal{X}} p \ln \frac{p}{q} \geq 0. \end{aligned} \tag{32}$$

□

Proposition 11 (Monotonicity of the α-geodesical skew divergence with respect to α)**.** *α-Geodesical skew divergence satisfies the following inequality for all $\alpha \in \mathbb{R}, \lambda \in [0,1]$.*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \geq D_{GS}^{(\alpha',\lambda)}[p\|q], \ (\alpha \geq \alpha').$$

Proof. Obvious from the inverse monotonicity of the f-interpolation (29) and the monotonicity of the logarithmic function. □

Figure 1 shows the monotonicity of the geodesic skew divergence with respect to α. In this figure, divergence is calculated between two binomial distributions.

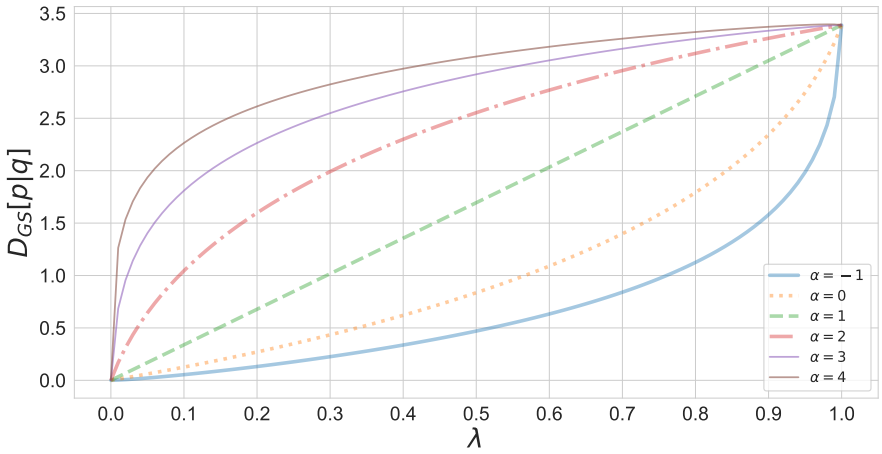

Figure 1. Monotonicity of the α-geodesical skew divergence with respect to α. The α-geodesical skew divergence between the binomial distributions $p = B(10, 0.3)$ and $q = B(10, 0.7)$ has been calculated.

Proposition 12 (Subadditivity of the α-geodesical skew divergence with respect to α)**.** *α-Geodesical skew divergence satisfies the following inequality for all $\alpha, \beta \in \mathbb{R}, \lambda \in [0,1]$*

$$D_{GS}^{(\alpha+\beta,\lambda)}[p\|q] \leq D_{GS}^{(\alpha,\lambda)}[p\|q] + D_{GS}^{(\beta,\lambda)}[p\|q].$$

Proof. For some α and λ, $m_f^{(\lambda,\alpha)}$ takes the form of the Kolmogorov mean [29], which is obvious from its continuity, monotonicity and self-distributivity. □

Proposition 13 (Continuity of the α-geodesical skew divergence with respect to α and λ)**.** *α-Geodesical skew divergence has the continuity property.*

Proof. We can prove from the continuity of the KL-divergence and the Kolmogorov mean. □

Figure 2 shows the continuity of the geodesic skew divergence with respect to α and λ. Both source and target distributions are binomial distributions. From this figure, it can be seen that the divergence changes smoothly as the parameters change.

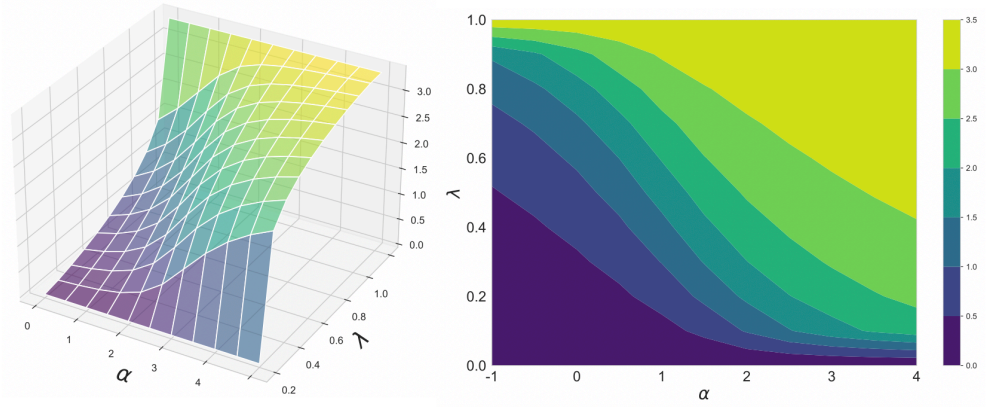

Figure 2. Continuity of the α-geodesmcal skew divergence with respect to α and λ. The α-geodesical skew divergence between the binomial distributions $p = B(10, 0.3)$ and $q = B(10, 0.7)$ has been calculated.

Lemma 1. *Suppose* $\alpha \to \infty$. *Then,*

$$\lim_{\alpha\to\infty} D_{GS}^{(\alpha,\lambda)}[p\|q] = \int_{\mathcal{X}} p \ln \frac{p}{\min\{p,q\}} d\mu \tag{33}$$

holds for all $\lambda \in [0,1]$.

Proof. Let $u = \frac{1-\alpha}{2}$. Then $\lim_{\alpha\to\infty} u = -\infty$. Assuming $p_0 \leq p_1$, it holds that

$$\begin{aligned}
\lim_{\alpha\to\infty} m_f^{(\lambda,\alpha)}(p_0,p_1) &= \lim_{u\to-\infty} \left((1-\lambda)p_0^u + \lambda p_1^u\right)^{\frac{1}{u}} \\
&= p_0 \lim_{u\to-\infty} \left((1-\lambda) + \lambda\left(\frac{p_1}{p_0}\right)^u\right)^{\frac{1}{u}} \\
&= p_0 = \min\{p_0,p_1\}.
\end{aligned}$$

Then, the following equality

$$\begin{aligned}
\lim_{\alpha\to\infty} D_{GS}^{(\alpha,\lambda)}[p\|q] &= \int_{\mathcal{X}} p \ln \frac{p}{\lim_{\alpha\to\infty} m_f^{(\lambda,\alpha)}(p_0,p_1)} d\mu \\
&= \int_{\mathcal{X}} p \ln \frac{p}{\min\{p,q\}} d\mu
\end{aligned}$$

holds. □

Lemma 2. *Suppose* $\alpha \to -\infty$. *Then,*

$$\lim_{\alpha\to\infty} D_{GS}^{(\alpha,\lambda)}[p\|q] = \int_{\mathcal{X}} p \ln \frac{p}{\max\{p,q\}} d\mu \tag{34}$$

holds for all $\lambda \in [0,1]$.

Proof. Let $u = \frac{1-\alpha}{2}$. Then $\lim_{\alpha\to-\infty} u = \infty$. Assuming $p_0 \leq p_1$, it holds that

$$\begin{aligned}\lim_{\alpha\to\infty} m_f^{(\lambda,\alpha)}(p_0,p_1) &= \lim_{u\to-\infty}\left((1-\lambda)p_0^u + \lambda p_1^u\right)^{\frac{1}{u}} \\ &= p_1 \lim_{u\to-\infty}\left((1-\lambda)\left(\frac{p_0}{p_1}\right)^u + \lambda\right)^{\frac{1}{u}} \\ &= p_1 = \max\{p_0,p_1\}.\end{aligned}$$

Then, the following equality

$$\begin{aligned}\lim_{\alpha\to-\infty} D_{GS}^{(\alpha,\lambda)}[p\|q] &= \int_{\mathcal{X}} p\ln\frac{p}{\lim_{\alpha\to-\infty} m_f^{(\lambda,\alpha)}(p_0,p_1)}d\mu \\ &= \int_{\mathcal{X}} p\ln\frac{p}{\max\{p,q\}}d\mu\end{aligned}$$

holds. □

Proposition 14 (Lower bound of the α-geodesical skew divergence)**.** *α-Geodesical skew divergence satisfies the following inequality for all $\alpha\in\mathbb{R}, \lambda\in[0,1]$.*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \geq \int_{\mathcal{X}} p\ln\frac{p}{\max\{p,q\}}d\mu. \tag{35}$$

Proof. It follows from the definition of the inverse monotonicity of f-interpolation (29) and Lemma 2. □

Proposition 15 (Upper bound of the α-geodesical skew divergence)**.** *α-Geodesical skew divergence satisfies the following inequality for all $\alpha\in\mathbb{R}, \lambda\in[0,1]$.*

$$D_{GS}^{(\alpha,\lambda)}[p\|q] \leq \int_{\mathcal{X}} p\ln\frac{p}{\min\{p,q\}}d\mu. \tag{36}$$

Proof. It follows from the definition of the f-interpolation (29) and Lemma 1. □

Theorem 1 (Strong convexity of the α-geodesical skew divergence)**.** *α-Geodesical skew divergence $D_{GS}^{(\alpha,\lambda)}[p\|q]$ is strongly convex in p with respect to the total variation norm.*

Proof. Let $r := m_f^{(\alpha,\lambda)}(p,q)$ and $f_j := \frac{p_j}{r}$ $(j=0,1)$, so that $f_t = \frac{p_t}{r}$ $(t\in(0,1))$. From Taylor's theorem, for $g(x) := x\ln x$ and $j=0,1$, it holds that

$$g(f_j) = g(f_t) + g'(f_t)(f_j-f_t) + (f_j-f_t)^2\int_0^1 g''((1-s)f_t+sf_j)(1-s)ds.$$

Let

$$\begin{aligned}\delta &:= (1-t)g(f_0) + tg(f_1) - g(f_t) \\ &= (1-t)t(f_1-f_0)^2\int_0^1\left(\frac{t}{(1-s)f_t+sf_0} + \frac{1-t}{(1-s)f_t+sf_1}\right)(1-s)ds \\ &= (1-t)t(f_1-f_0)^2\int_0^1\left(\frac{t}{f_{u_0}(t,s)} + \frac{1-t}{f_{u_1}(t,s)}\right)(1-s)ds,\end{aligned}$$

where

$$u_j(t,s) \coloneqq (1-s)t + jt,$$
$$f_{\mu_j}(t,s) \coloneqq (1-s)f_t + sf_j.$$

Then,

$$\begin{aligned}\Delta &\coloneqq (1-t)H(p_0) + tH(p_1) - H(p_t)\\ &= \int \delta dr\\ &= (1-t)t\int_0^1 (1-s)ds\Big[tI(u_0(t,s)) + (1-t)I(u_1(t,s))\Big],\end{aligned}$$

where

$$\|p_1 - p_0\| \coloneqq \int |dp_1 - dp_0|d\mu,$$
$$H(p) \coloneqq D_{GS}^{(\alpha,\lambda)}[p\|r] = \int p\ln\frac{p}{r}d\mu,$$
$$I(u) \coloneqq \int \frac{(f_1 - f_0)^2}{f_u}dr.$$

Now, it is suffice to prove that $\Delta \geq \frac{t(1-t)}{2}\|p_1 - p_0\|^2$. For all $u \in (0,1)$, it is seen that p_1 is absolutely continuous with respect to p_u. Let $g_u \coloneqq \frac{p_1}{p_u} = \frac{f_1}{f_u}$. One obtains

$$\begin{aligned}I(u) &= \frac{1}{(1-u)^2}\int \frac{(f_1 - f_u)^2}{f_u}dr\\ &= \frac{1}{(1-u)^2}\int (g_u - 1)^2 dp_u\\ &\geq \frac{1}{(1-u)^2}\left(\int |g_u - 1|dp_u\right)^2\\ &= \frac{1}{(1-u)^2}\|p_1 - p_u\|^2 = \|p_1 - p_0\|^2,\end{aligned}$$

and hence, for $j = 0, 1$,

$$\Delta \geq \frac{t(1-t)}{2}\|p_1 - p_0\|^2.$$

□

4. Natural α-Geodesical Skew Divergence for Exponential Family

In this section, the exponential family is considered in which the probability density function is given by

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\Big\{\boldsymbol{\theta}\cdot\boldsymbol{x} + k(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\Big\}, \tag{37}$$

where $\boldsymbol{x}$ is a random variable. In the above equation, $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^n)$ is an n-dimensional vector parameter to specify distribution, $k(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ and ψ corresponds to the normalization factor.

In skew divergence, the probability distribution of the target is a weighted average of the two distributions. This implicitly assumes that interpolation of the two probability distributions is properly given by linear interpolation. Here, in the exponential family, the interpolation between natural parameters rather than interpolation between probability

distributions themselves is considered. Namely, the geodesic connecting two distributions $p(\boldsymbol{x};\boldsymbol{\theta}_p)$ and $q(\boldsymbol{x};\boldsymbol{\theta}_q)$ on the $\boldsymbol{\theta}$-coordinate system is considered:

$$\boldsymbol{\theta}(\lambda) = (1-\lambda)\boldsymbol{\theta}_p + \lambda\boldsymbol{\theta}_q, \tag{38}$$

where $\lambda \in [0,1]$ is the parameter. The probability distributions on the geodesic $\boldsymbol{\theta}(\lambda)$ are

$$\begin{aligned} p(\boldsymbol{x};\lambda) &= p(\boldsymbol{x};\boldsymbol{\theta}(\lambda)) \\ &= \exp\Big\{\lambda(\boldsymbol{\theta}_q - \boldsymbol{\theta}_p)\cdot\boldsymbol{x} + \boldsymbol{\theta}_p\cdot\boldsymbol{x} - \psi(\lambda)\Big\}. \end{aligned} \tag{39}$$

Hence, a geodesic itself is a one-dimensional exponential family, where λ is the natural parameter. A geodesic consists of a linear interpolation of the two distributions in the logarithmic scale because

$$\ln p(\boldsymbol{x};\lambda) = (1-\lambda)\ln p(\boldsymbol{x};\boldsymbol{\theta}_p) + \lambda \ln p(\boldsymbol{x};\boldsymbol{\theta}_q) - \psi(\lambda). \tag{40}$$

This corresponds to the case $\alpha = 1$ on the f-interpolation with normalization factor $c(\lambda) = \exp\{-\psi(\lambda)\}$,

$$p(\boldsymbol{x};\boldsymbol{\theta}(\lambda)) = m_f^{(\lambda,1)}(p(\boldsymbol{x};\boldsymbol{\theta}_p), p(\boldsymbol{x};\boldsymbol{\theta}_q)). \tag{41}$$

This induces the natural geodesic skew divergence with $\alpha = 1$ as

$$\begin{aligned} D_{GS}^{(1,\lambda)}[p\|q] &= \int_{\mathcal{X}} p\ln\left(\frac{p}{m_f^{(\lambda,1)}(p,q)}\right)d\mu \\ &= \int_{\mathcal{X}} p\ln p - p\ln\Big(m_f^{(\lambda,1)}(p,q)\Big)d\mu \\ &= \int_{\mathcal{X}} p\ln p - p\ln\Big(\exp\{(1-\lambda)\ln p + \lambda\ln q\}\Big)d\mu \\ &= \int_{\mathcal{X}} \Big(p\ln p - (1-\lambda)p\ln p - \lambda p\ln q\Big)d\mu \\ &= \int_{\mathcal{X}} \Big(\lambda p\ln p - \lambda p\ln q\Big)d\mu \\ &= \lambda\int_{\mathcal{X}} p\ln\frac{p}{q}d\mu \\ &= \lambda D_{KL}[p\|q], \end{aligned}$$

and this is equal to the scaled KL divergence.

More generally, let $\theta_P^{(\alpha)}$ and $\theta_Q^{(\alpha)}$ be the parameter representations on the α-coordinate system of probability distributions P and Q. Then, the geodesics between them are represented as in Figure 3, and it induces the α-geodesical skew divergence.

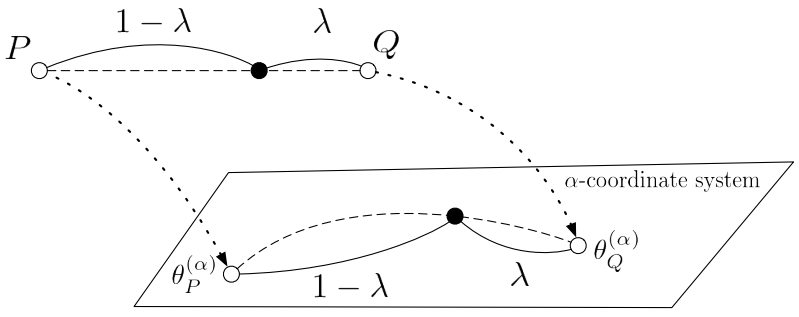

Figure 3. The geodesic between two probability distributions on the α-coordinate system.

5. Function Space Associated with the α-Geodesical Skew Divergence

To discuss the functional nature of the α-geodesical skew divergence in more depth, the function space it constitutes is considered. For an α-geodesical skew divergence $f_q^{(\alpha,\lambda)}(p) = D_{GS}^{(\alpha,\lambda)}[p\|q]$ with one side of the distribution fixed, let the entire set be

$$\mathcal{F}_q = \left\{ f_q^{(\alpha,\lambda)} \mid \alpha \in \mathbb{R}, \lambda \in [0,1] \right\}. \tag{42}$$

For $f_q^{(\alpha,\lambda)} \in \mathcal{F}_q$, its semi-norm is defined by

$$\left\| f_q^{(\alpha,\lambda)} \right\|_p := \int_{\mathcal{X}} \left(\left| f_q^{(\alpha,\lambda)} \right|^p d\mu \right)^{\frac{1}{p}}. \tag{43}$$

By defining addition and scalar multiplication for $f_q^{(\alpha,\lambda)}, g_q^{(\alpha,\lambda)} \in \mathcal{F}_q$, $c \in \mathbb{R}$ as follows, $\mathcal{F}_q$ becomes a semi-norm vector space:

$$(f_q^{(\alpha,\lambda)} + g_q^{(\alpha,\lambda)})(u) := f_q^{(\alpha,\lambda)}(u) + g_q^{(\alpha,\lambda)}(u) = D_{GS}^{(\alpha,\lambda)}[u\|q] + D_{GS}^{(\alpha',\lambda')}[u\|q], \tag{44}$$

$$(cf)(u) := c f_q^{(\alpha,\lambda)}(u) = c \cdot D_{GS}^{(\alpha,\lambda)}[u\|q]. \tag{45}$$

Theorem 2. *Let $\mathcal{N}$ be the kernel of $\|\cdot\|_p$ as follows:*

$$\mathcal{N} := ker(\|\cdot\|_p) = \left\{ f_q^{(\alpha,\lambda)} \mid f_q^{(\alpha,\lambda)} = 0 \right\}. \tag{46}$$

Then the quotient space $\mathcal{V} := (\mathcal{F}_q, \|\cdot\|_p)/N$ is a Banach space.

Proof. It is sufficient to prove that $f_q^{(\alpha,\lambda)}$ is integrable to the power of p and that $\mathcal{V}$ is complete. From Proposition 15, the α-geodesical skew divergence is bounded from above for all $\alpha \in \mathbb{R}$ and $\lambda \in [0,1]$. Since $f_q^{(\alpha,\lambda)}$ is continuous, we know that it is p-power integrable.

Let $\{f_n\}$ be a Cauchy sequence of $\mathcal{V}$:

$$\lim_{n,m\to\infty} \|f_n - f_m\|_p = 0.$$

As $n(k)$, $k = 1, 2, \ldots$, can be taken to be monotonically increasing and

$$\|f_n - f_{n(k)}\|_p < 2^{-k}$$

with respect to $n > n(k)$, let

$$\|f_{n(k+1)} - f_{n(k)}\|_p < 2^{-k}.$$

If $g_n = |f_{n(1)}| + \sum_{j=1}^{n-1} |f_{n(j+1)} - f_{n(j)}| \in \mathcal{V}$, it is non-negatively monotonically increasing at each point, and from the subadditivity of the norm, $\|g_n\|_p \leq \|f_{n(1)}\|_p + \sum_{j=1}^{n-1} 2^{-j}$. From the monotonic convergence theorem, we have

$$\left\| \lim_{n\to\infty} g_n \right\|_p = \lim_{n\to\infty} \|g_n\|_p \leq \|f_{n(1)}\|_p + 1 < \infty.$$

That is, $\lim_{n\to\infty} g_n$ exists almost everywhere, and $\lim_{n\to\infty} g_n \in \mathcal{V}$. From $\lim_{n\to\infty} g_n < \infty$, we have

$$f_{n(1)} + \sum_{j=1}^{n-1} (f_{n(j+1)} - f_{n(j)}) = \lim_{n\to\infty} f_{n(1)}$$

converges absolutely almost everywhere to $|\lim_{n\to\infty} f_{n(n)}| \leq \lim_{n\to\infty} g_n, a.e.$. That is, $\lim_{n\to\infty} f_{n(n)} \in \mathcal{V}$. Then

$$\left| \lim_{n\to\infty} f_n - f_{n(n)} \right| \leq \lim_{n\to\infty} g_n$$

and from the superior convergence theorem, we can obtain

$$\lim_{n\to\infty} \left\| \lim_{n\to\infty} f_n - f_{n(n)} \right\|_p = 0$$

We have now confirmed the completeness of $\mathcal{V}$. □

Corollary 1. *Let*

$$\mathcal{F}_+ = \left\{ f_q^{(\alpha,\lambda)} \mid \alpha \in \mathbb{R}, \lambda \in (0,1], q \in \mathcal{P} \right\}. \tag{47}$$

Then the space $\mathcal{V}_+ := (\mathcal{F}_+, \|\cdot\|_p)$ *is a Banach space.*

Proof. If we restrict $\lambda \in (0,1]$, $D_{GS}^{(\alpha,\lambda)}[u\|q] = 0$ if and only if $u = q$. Then, $\mathcal{V}_+$ has the unique identity element, and then $\mathcal{V}_+$ is a complete norm space. □

Consider the second argument Q of $D_{GS}^{(\alpha,\lambda)}(P||Q)$ is fixed, which is referred to as the reference distribution. Figure 4 shows values of the α-geodesical skew divergence for a fixed reference Q, where both P and Q are restricted to be Gaussian. In this figure, the reference distribution is $\mathcal{N}(0, 0.5)$ and the parameters of input distributions are varied in $\mu \in [0, 4.5]$ and $\sigma^2 \in [0.5, 2.3]$. From this figure, one can see that a larger value of α emphasizes the discrepancy between distributions P and Q. Figure 5 illustrates a coordinate system associated with the α-geodesical skew divergence for different α. As seen from the figure, for the same pair of distributions P and Q, the value of divergence with $\alpha = 3$ is larger than that with $\alpha = -1$.

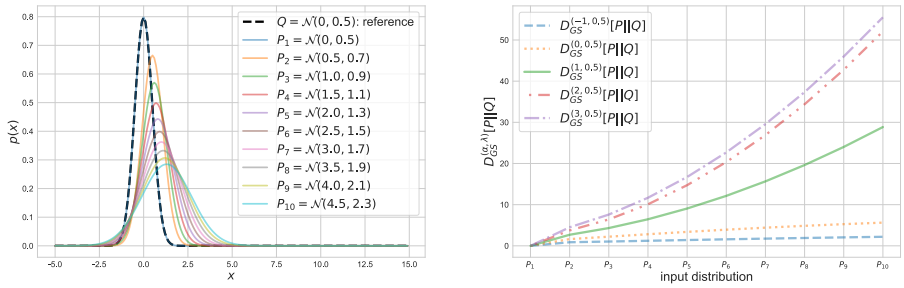

Figure 4. α-geodesical skew divergence between two normal distributions. The reference distribution is $Q = \mathcal{N}(0, 0.5)$. For $P_1, P_2, \ldots, P_j$, $(j = 1, 2, \ldots, 10)$, let their mean and variance be μ_j and σ_j^2, respectively, where $\mu_{j+1} - \mu_j = 0.5$ and $\sigma_{j+1}^2 - \sigma_j^2 = 0.2$.

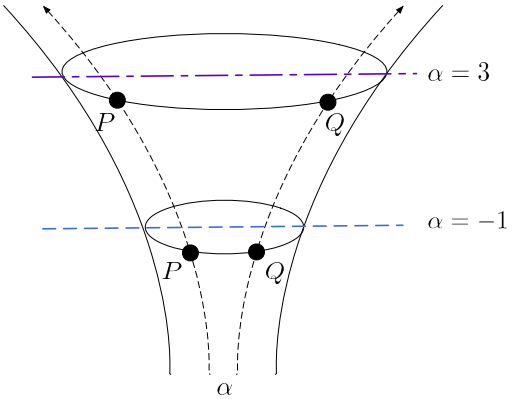

Figure 5. Coordinate system of $\mathcal{F}_q$ or $\mathcal{F}_+$. Such a coordinate system is not Euclidean.

6. Conclusions and Discussion

In this paper, a new family of divergence is proposed to address the computational difficulty of KL-divergence. The proposed α-geodesical skew divergence is a natural derivation from the concept of α-geodesics in information geometry and generalizes many existing divergences.

Furthermore, α-geodesical skew divergence leads to several applications. For example, the new divergence can be applied to the annealed importance sampling by the same analogy as in previous studies using q-paths [41]. It could also be applied to linguistics, a field in which skew divergence was originally used [19].

Author Contributions: Formal analysis, M.K. and H.H.; Investigation, M.K.; Methodology, M.K. and H.H.; Software, M.K.; Supervision, H.H.; Validation, H.H.; Visualization, M.K.; Writing—original draft, M.K.; Writing—review & editing, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JPSJ (KAKENHI) grant number JP17H01793, JST CREST Grant No. JPMJCR2015 and NEDO JPNP18002.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors express special thanks to the editor and reviewers, whose comments led to valuable improvements to the manuscript.

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

References

1. Deza, M.M.; Deza, E. Encyclopedia of distances. In *Encyclopedia of Distances*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–583.
2. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633. [CrossRef]
3. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
4. Sakamoto, Y.; Ishiguro, M.; Kitagawa, G. *Akaike Information Criterion Statistics*; D. Reidel: Dordrecht, The Netherlands, 1986; Volume 81, p. 26853.
5. Goldberger, J.; Gordon, S.; Greenspan, H. An Efficient Image Similarity Measure Based on Approximations of KL-Divergence Between Two Gaussian Mixtures. *ICCV* **2003**, *3*, 487–493.
6. Yu, D.; Yao, K.; Su, H.; Li, G.; Seide, F. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 16–31 May 2013.
7. Solanki, K.; Sullivan, K.; Madhow, U.; Manjunath, B.; Chandrasekaran, S. Provably secure steganography: Achieving zero KL divergence using statistical restoration. In Proceedings of the 2006 International Conference on Image Processing, Atlanta, GA, USA, 8–11 October 2006.
8. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
9. Menéndez, M.; Pardo, J.; Pardo, L.; Pardo, M. The jensen-shannon divergence. *J. Frankl. Inst.* **1997**, *334*, 307–318. [CrossRef]
10. Nielsen, F. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485. [CrossRef]
11. Jeffreys, H. An Invariant Form for the Prior Probability in Estimation Problems. Available online: https://royalsocietypublishing.org/doi/10.1098/rspa.1946.0056 (accessed on 24 April 2021).
12. Chatzisavvas, K.C.; Moustakidis, C.C.; Panos, C. Information entropy, information distances, and complexity in atoms. *J. Chem. Phys.* **2005**, *123*, 174111. [CrossRef]
13. Bigi, B. Using Kullback-Leibler distance for text categorization. In *European Conference on Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 305–319.
14. Wang, F.; Vemuri, B.C.; Rangarajan, A. Groupwise point pattern registration using a novel CDF-based Jensen-Shannon divergence. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006.
15. Nishii, R.; Eguchi, S. Image classification based on Markov random field models with Jeffreys divergence. *J. Multivar. Anal.* **2006**, *97*, 1997–2008. [CrossRef]

16. Bayarri, M.; García-Donato, G. Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2008**, *70*, 981–1003. [CrossRef]
17. 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]
18. Nielsen, F. On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid. *Entropy* **2020**, *22*, 221. [CrossRef]
19. Lee, L. Measures of distributional similarity. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, MD, USA, 20–26 June 1999.
20. Lee, L. On the Effectiveness of the Skew Divergence for Statistical Language Analysis. In Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, Key West, FL, USA, 4–7 January 2001; pp. 176–783.
21. Xiao, F.; Wu, Y.; Zhao, H.; Wang, R.; Jiang, S. Dual skew divergence loss for neural machine translation. *arXiv* **2019**, arXiv:1908.08399.
22. Carvalho, B.M.; Garduño, E.; Santos, I.O. Skew divergence-based fuzzy segmentation of rock samples. *J. Phys. Conf. Ser.* **2014**, *490*, 012010. [CrossRef]
23. Revathi, P.; Hemalatha, M. Cotton leaf spot diseases detection utilizing feature selection with skew divergence method. *Int. J. Sci. Eng. Technol.* **2014**, *3*, 22–30.
24. Ahmed, N.; Neville, J.; Kompella, R.R. Network Sampling via Edge-Based Node Selection with Graph Induction. Available online: https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2743&context=cstech (accessed on 24 April 2021).
25. Hughes, T.; Ramage, D. Lexical semantic relatedness with random graph walks. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007.
26. Audenaert, K.M. Quantum skew divergence. *J. Math. Phys.* **2014**, *55*, 112202. [CrossRef]
27. Hardy, G.H.; Littlewood, J.E.; Pólya, G. *Inequalities*; Cambridge University Press: Cambridge, UK, 1952.
28. Amari, S.I. *Information Geometry and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2016.
29. Kolmogorov, A.N.; Castelnuovo, G. *Sur la Notion de la Moyenne*; Atti Accad. Naz: Lincei, French, 1930.
30. Nagumo, M. Über eine klasse der mittelwerte. *Jpn. J. Math.* **1930**, *7*, 71–79. [CrossRef]
31. Nielsen, F. Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. *Pattern Recognit. Lett.* **2014**, *42*, 25–34. [CrossRef]
32. Amari, S.I. *Differential-Geometrical Methods in Statistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 28.
33. Amari, S. Differential-geometrical methods in statistics. *Lect. Notes Stat.* **1985**, *28*, 1.
34. Amari, S. α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931. [CrossRef]
35. Ay, N.; Jost, J.; Lê, H.V.; Schwachhöfer, L. *Information Geometry*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017. [CrossRef]
36. Morozova, E.A.; Chentsov, N.N. Markov invariant geometry on manifolds of states. *J. Sov. Math.* **1991**, *56*, 2648–2669. [CrossRef]
37. Eguchi, S.; Komori, O. Path Connectedness on a Space of Probability Density Functions. In *Lecture Notes in Computer Science*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 615–624. [CrossRef]
38. Nielsen, F. On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius. *Entropy* **2021**, *23*, 464. [CrossRef]
39. Nielsen, F. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv* **2010**, arXiv:1009.4004.
40. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
41. Brekelmans, R.; Masrani, V.; Bui, T.D.; Wood, F.D.; Galstyan, A.; Steeg, G.V.; Nielsen, F. Annealed Importance Sampling with q-Paths. *arXiv* **2020**, arXiv:2012.07823.