Entropy **2013**, *15*, 5384-5418; doi:10.3390/e15125384

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Nonparametric Information Geometry: From Divergence Function to Referential-Representational Biduality on Statistical Manifolds

Jun Zhang

Department of Psychology and Department of Mathematics, University of Michigan, 530 Church Street, Ann Arbor, MI 48109, USA; E-Mail: junz@umich.edu; Tel: +1-734-763-6161

Received: 3 July 2013; in revised form: 11 October 2013 / Accepted: 22 October 2013 / Published: 4 December 2013

Abstract: Divergence functions are the non-symmetric "distance" on the manifold, $\mathcal{M}_\theta$, of parametric probability density functions over a measure space, (X, μ). Classical information geometry prescribes, on $\mathcal{M}_\theta$: (i) a Riemannian metric given by the Fisher information; (ii) a pair of dual connections (giving rise to the family of α-connections) that preserve the metric under parallel transport by their joint actions; and (iii) a family of divergence functions (α-divergence) defined on $\mathcal{M}_\theta \times \mathcal{M}_\theta$, which induce the metric and the dual connections. Here, we construct an extension of this differential geometric structure from $\mathcal{M}_\theta$ (that of parametric probability density functions) to the manifold, $\mathcal{M}$, of non-parametric functions on X, removing the positivity and normalization constraints. The generalized Fisher information and α-connections on $\mathcal{M}$ are induced by an α-parameterized family of divergence functions, reflecting the fundamental convex inequality associated with any smooth and strictly convex function. The infinite-dimensional manifold, $\mathcal{M}$, has zero curvature for all these α-connections; hence, the generally non-zero curvature of $\mathcal{M}_\theta$ can be interpreted as arising from an embedding of $\mathcal{M}_\theta$ into $\mathcal{M}$. Furthermore, when a parametric model (after a monotonic scaling) forms an affine submanifold, its natural and expectation parameters form biorthogonal coordinates, and such a submanifold is dually flat for $\alpha = \pm 1$, generalizing the results of Amari's α-embedding. The present analysis illuminates two different types of duality in information geometry, one concerning the referential status of a point (measurable function) expressed in the divergence function ("referential duality") and the other concerning its representation under an arbitrary monotone scaling ("representational duality").

Keywords: Fisher information; alpha-connection; infinite-dimensional manifold; convex function

1. Introduction

Information geometry is a differential geometric study of the manifold of probability measures or probability density functions [1]. Its role in understanding asymptotic inference was summarized in [2–7]. Information geometric methods have been applied to many areas of interest to statisticians, such as the study of estimating functions (e.g., [8,9]) and invariant priors for Bayesian inference (e.g., [10,11]), and to the machine learning community, such as the natural gradient descent method [12,13], support vector machine [14], boosting [15], turbo decoding [16], *etc*.

The differential geometric structure of statistical models with finite parameters is now well understood. Consider a family of probability functions (*i.e.*, probability measures on discrete support or probability density functions on continuous support) as parameterized by $\theta = [\theta^1, \cdots, \theta^n]$. The collection of such probability functions, where each function is indexed by a point, $\theta \in \mathbb{R}^n$, forms a manifold, $\mathcal{M}_\theta$, under suitable conditions. Rao [17] identified Fisher information to be the Riemannian metric for $\mathcal{M}_\theta$. Efron [18], through investigating a one-parameter family of statistical models, elucidated the meaning of curvature for asymptotic statistical inference and pointed out its flatness for the exponential model. In his reaction to Efron's work, Dawid [19] invoked the differential geometric notion of linear connections on a manifold as preserving parallelism during vector transportation and pointed out other possible constructions of linear connections on $\mathcal{M}_\theta$, in addition to the non-flat Levi-Civita connection associated with the Fisher metric. Amari [2,20], in his path-breaking work, systematically advanced the theory of information geometry by constructing a parametric family of α-connections, $\Gamma^{(\alpha)}$, $\alpha \in \mathbb{R}$, along with a dualistic interpretation of $\alpha \leftrightarrow -\alpha$ as conjugate connections on the manifold, $\mathcal{M}_\theta$. The e-connection ($\alpha = 1$) vanishes (*i.e.*, becomes identically zero) on the manifold of the exponential family of probability functions under natural parameters, whereas the m-connection ($\alpha = -1$) vanishes on the manifold of the mixture family of probability functions under mixture parameters. Therefore, not only have $\Gamma^{(\pm 1)}$ zero curvatures for both exponential and mixture families, but affine coordinates were found to yield $\Gamma^{(1)}$ and $\Gamma^{(-1)}$, themselves zero for the exponential and mixture families, respectively.

This classic information geometry dealing with parametric statistical models has been investigated in the non-parametric setting using the tools of infinite-dimensional analysis [21–23], with non-parametric Fisher information given by [23]. This is made possible, because topological issues were resolved by the pioneering work of [24] using the theory of Orlicz space for charting the exponential statistical manifold. Zhang and Hasto [25] characterized the probability manifold modeled on an ambient affine space via functional equations and generalized exponential charts. The goal of the present paper is to extend these non-parametric results by showing links among three inter-connected mathematical topics that underlie information geometry, namely: (i) divergence functions measuring the non-symmetric distance of any two points (density or measurable functions) on the manifold (the referential duality); (ii) convex analysis and the associated Legendre–Fenchel transformation linking the natural and expectation parameters of parametric models (the representational duality); and (iii) the resulting dual Riemannian

structure involving the Fisher metric and the family of α-connections. Results in the parametric setting were summarized in [26].

The Riemannian manifold of parametric statistical models is a special kind, one that involves dual (as known as conjugate) connections; historically, such a mathematical theory was independently developed to investigate hypersurface immersion (see [27,28]). Lauritzen [29] characterized the general differential geometric context under which a one-parameter family of α-connections arise, as well as the meaning of conjugacy for a pair of connections on statistical manifolds [30]. Kurose [31,32] and then Matsuzoe [33,34] elucidated information geometry from an affine differential geometric perspective. See, also [35] for a generalized notion of conjugate connections. It was Eguchi [36–38] who provided a generic way for inducing a metric and a pair of conjugate connections from an arbitrary divergence (what he called "contrast") function. The current exposition will build on this "Eguchi relation" between the metric and conjugate connections of the Riemannian manifold, $\mathcal{M}_\theta$, and the divergence function defined on $\mathcal{M}_\theta \times \mathcal{M}_\theta$.

The main results of this paper include the introduction of an α-parametric family of divergence functionals on measurable functions (including probability functions) using any smooth and strictly convex function and the induction by such divergence a metric and a family of conjugate connections that resemble, but generalize, the Fisher information proper and α-connections proper. In particular, we derive explicit expressions of the metric and conjugate connections on the infinite-dimensional manifold of all functions defined on the same support of the sample space. When finite-dimensional affine embedding is allowed, our formulae reduce to the familiar ones associated with the exponential family established in [2]. We carefully delineate two senses of duality associated with such manifolds, one related to the reference/comparison status of any pair of points (functions) and the other related to properly scaled representations of them.

1.1. Parametric Information Geometry Revisited

Here, we briefly summarize the well-known results of parametric information geometry in the classical (as opposed to quantum) sense. The motivation is two-fold. First, by reviewing the basic parametric results, we want to make sure that any generalization of the framework of information geometry will reduce to those formulae under appropriate conditions. Secondly, understanding how a divergence function is related to the dual Riemannian structure will enable us to approach the infinite-dimensional case by analogy, that is, through constructing more general classes of divergence functionals defined on function spaces.

1.1.1. Riemannian Manifold, Fisher Metric and α-Connections

Let $(\mathcal{X}, \mu)$ be a measure space with σ-algebra built upon the atoms, $d\zeta$, of $\mathcal{X}$. Let $\mathcal{M}_\mu$ denote the space of probability density functions, $p : \mathcal{X} \to \mathbb{R}_+ (\equiv \mathbb{R}^+ \cup \{0\})$, defined on the sample space, $\mathcal{X}$, with background measure $d\mu = \mu(d\zeta)$:

$$\mathcal{M}_\mu = \{p(\zeta) : \mathrm{E}_\mu\{p(\zeta)\} = 1;\, p(\zeta) > 0\ ,\ \forall \zeta \in \mathcal{X}\} \tag{1}$$

Here, and throughout this paper, $\mathrm{E}_\mu\{\cdot\} = \int_{\mathcal{X}}\{\cdot\}\, d\mu$ denotes the expectation of a measurable function (in curly brackets) with respect to the background measure, μ. We also denote $\mathrm{E}_p\{\cdot\} = \int_{\mathcal{X}}\{\cdot\}\, p\, d\mu$.

A parametric family of density functions, $p(\cdot|\theta)$, called a parametric statistical model, is the association of a density function, $\theta \mapsto p(\cdot|\theta)$, for each n-dimensional vector $\theta = [\theta^1, \cdots, \theta^n]$. The space of parametric statistical models forms a Riemannian manifold (where θ is treated as the local chart):

$$\mathcal{M}_\theta = \{p(\zeta|\theta) \in \mathcal{M}_\mu : \theta \in \Theta \subset \mathbb{R}^n\} \subset \mathcal{M}_\mu \tag{2}$$

with the so-called Fisher metric [17]:

$$g_{ij}(\theta) = \mathrm{E}_\mu \left\{ p(\zeta|\theta)\, \frac{\partial \log p(\zeta|\theta)}{\partial \theta^i}\, \frac{\partial \log p(\zeta|\theta)}{\partial \theta^j} \right\} \tag{3}$$

and α-connections [20,39]:

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ p(\zeta|\theta)\, \left(\frac{1-\alpha}{2}\, \frac{\partial \log p(\zeta|\theta)}{\partial \theta^i}\, \frac{\partial \log p(\zeta|\theta)}{\partial \theta^j} + \frac{\partial^2 \log p(\zeta|\theta)}{\partial \theta^i \partial \theta^j} \right) \frac{\partial \log p(\zeta|\theta)}{\partial \theta^k} \right\} \tag{4}$$

with the α-connections satisfying the dualistic relation:

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = \Gamma^{(-\alpha)}_{ij,k}(\theta) \tag{5}$$

Here, $*$, denotes conjugate (dual) connection. Recall that, in general, a metric is a bilinear map on the tangent space, and an affine connection is used to define parallel transport of vectors. The conjugacy in a pair of connections, $\Gamma \longleftrightarrow \Gamma^*$, is defined by their jointly preserving the metric when each acts on one of the two tangent vectors; that is, when the tangent vectors undergo parallel transport according to Γ or Γ^* respectively. Equivalently, and perhaps more fundamentally, the pair of conjugate connections preserve the dual pairing of vectors in the tangent space with co-vectors in the cotangent space [30]. Any Riemannian manifold with its metric, g, and conjugate connections, Γ, Γ^*, given in the form of Equations (3)–(5) is called a *statistical manifold* (in the narrower sense) and is denoted as $\{\mathcal{M}_\theta, g, \Gamma^{(\pm\alpha)}\}$. In the broader sense, a statistical manifold $\{\mathcal{M}, g, \Gamma, \Gamma^*\}$ is a differentiable manifold equipped with a Riemannian metric g and a pair of torsion-free conjugate connections $\Gamma \equiv \Gamma^{(1)}$, $\Gamma^* \equiv \Gamma^{(-1)}$, without necessarily requiring g and Γ, Γ^* to take the forms of Equations (3)–(5).

1.1.2. Exponential Family, Mixture Family and Their Generalization

An exponential family of probability density functions is defined as:

$$p^{(e)}(\zeta|\theta) = \exp\left(F_0(\zeta) + \sum_i \theta^i\, F_i(\zeta) - \Phi(\theta) \right) \tag{6}$$

where θ is its natural parameter and $F_i(\zeta)\,(i = 1, \cdots, n)$ is a set of linearly independent functions with the same support in $\mathcal{X}$, and the cumulant generating function ("potential function") $\Phi(\theta)$ is:

$$\Phi(\theta) = \log \mathrm{E}_\mu \left\{ \exp\left(F_0(\zeta) + \sum_i \theta^i\, F_i(\zeta) \right) \right\} \tag{7}$$

Substituting Equation (6) into Equations (3) and (4), the Fisher metric and the α-connections are simply:

$$g_{ij}(\theta) = \frac{\partial^2 \Phi(\theta)}{\partial\theta^i \partial\theta^j} \tag{8}$$

and:

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \frac{1-\alpha}{2} \frac{\partial^3 \Phi(\theta)}{\partial\theta^i \partial\theta^j \partial\theta^k} \tag{9}$$

whereas the Riemannian curvature tensor (of an α-connection) is given by [2], p.106:

$$R^{(\alpha)}_{ij\mu\nu}(\theta) = \frac{1-\alpha^2}{4} \sum_{l,k} (\Phi_{il\nu}\Phi_{jk\mu} - \Phi_{il\mu}\Phi_{jk\nu})\Phi^{lk} \tag{10}$$

where $\Phi^{ij} = g^{ij}$ is the matrix inverse of g_{ij} and subscripts of Φ indicate partial derivatives. Therefore, the α-connection for the exponential family is dually flat when $\alpha = \pm 1$. In particular, all components of $\Gamma^{(1)}_{ij,k}$ vanish, due to Equation (9), on the manifold formed by $p^{(e)}(\cdot|\theta)$ in which the natural parameter, θ, serves as the local coordinates.

On the other hand, the mixture family:

$$p^{(m)}(\zeta|\theta) = \sum_i \theta^i F_i(\zeta) \tag{11}$$

when viewed as a manifold charted by its mixture parameter, θ, with the constraints, $\sum_i \theta^i = 1$ and $\int_X F_i(\zeta) d\mu = 1$, turns out to have identically zero $\Gamma^{(-1)}_{ij,k}$. The connections, $\Gamma^{(1)}$ and $\Gamma^{(-1)}$, are also called the exponential and mixture connections, or e- and m-connection, respectively. The exponential family and the mixture family are special cases of the α-family [1,2] of density functions, $p(\zeta|\theta)$, whose denormalization satisfies (with constant κ):

$$l^{(\alpha)}(\kappa p) = F_0(\zeta) + \sum_i \theta^i F_i(\zeta) \tag{12}$$

under the α-embedding function, $l^{(\alpha)} : \mathbb{R}^+ \to \mathbb{R}$, defined as:

$$l^{(\alpha)}(t) = \begin{cases} \log t & \alpha = 1 \\ \frac{2}{1-\alpha} t^{(1-\alpha)/2} & \alpha \neq 1 \end{cases} \tag{13}$$

The α-embedding of a probability density function plays an important role in Tsallis statistics; see, e.g., [40]. Under α-embedding, the denormalized density functions form the so-called α-*affine manifold* [1], p.46. The Fisher metric and α-connections, under such α-representation, have the following expressions:

$$g_{ij}(\theta) = \mathrm{E}_\mu \left\{ \frac{\partial l^{(\alpha)}(p(\cdot|\theta))}{\partial\theta^i} \frac{\partial l^{(-\alpha)}(p(\cdot|\theta))}{\partial\theta^j} \right\} \tag{14}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \frac{\partial^2 l^{(\alpha)}(p(\cdot|\theta))}{\partial\theta^i \partial\theta^j} \frac{\partial l^{(-\alpha)}(p(\cdot|\theta))}{\partial\theta^k} \right\} \tag{15}$$

Clearly, on an α-affine manifold with any given α value, components of $\Gamma^{(\alpha)}$ are all identically zero by virtue of the definition of α-family Equation (12), and hence, $\pm\alpha$-connections are dually flat.

1.2. Divergence Function and Induced Statistical Manifold

It is well known that the statistical manifold, $\{\mathcal{M}_\theta, g, \Gamma^{(\pm\alpha)}\}$, with Fisher information as the metric, g, and the $(\pm\alpha)$-connections, $\Gamma^{(\pm\alpha)}$, as conjugate connections, can be induced from a parametric family of divergence functions called "α-divergence". Here, we briefly review the link of divergence functions to the dual Riemannian geometry of statistical manifolds.

1.2.1. Kullback-Leibler Divergence, Bregman Divergence and α-Divergence

Divergence functions are distance-like quantities; they measure the directed (non-symmetric) difference of two probability density functions in the infinite-dimensional function space or two points in a finite-dimensional vector space of the parameters of a statistical model. An example is the *Kullback-Leibler divergence* (also known as, KL cross-entropy) between two probability densities, $p, q \in \mathcal{M}_\mu$, here expressed in its extended form (i.e., without requiring p and q to be normalized):

$$K(p,q) = \int \left(q - p - p \log \frac{q}{p} \right) d\mu = K^*(q,p) \tag{16}$$

with a unique, global minimum of zero when $p = q$. For the exponential family Equation (6), the expression (16) takes the form of the so-called *Bregman divergence* [41] defined on $\Theta \times \Theta \subseteq \mathbb{R}^n \times \mathbb{R}^n$:

$$\mathrm{B}_\Phi(\theta_p, \theta_q) = \Phi(\theta_p) - \Phi(\theta_q) - \langle \theta_p - \theta_q, \partial\Phi(\theta_q) \rangle \tag{17}$$

where Φ is the potential function (7), ∂ is the gradient operator and $\langle \cdot, \cdot \rangle$ denotes the standard bilinear form (pairing) of a vector with a co-vector. The Bregman divergence (17) expresses the directed-distance of two members, p and q, of the exponential family as indexed, respectively, by the two parameters, θ_p and θ_q.

A generalization of the Kullback-Leibler divergence is the *α-divergence*, defined as:

$$\mathcal{A}^{(\alpha)}(p,q) = \frac{4}{1-\alpha^2} \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} p + \frac{1+\alpha}{2} q - p^{\frac{1-\alpha}{2}} q^{\frac{1+\alpha}{2}} \right\} \tag{18}$$

measuring the directed distance between any two density functions, p and q. It is easily seen that:

$$\lim_{\alpha \to -1} \mathcal{A}^{(\alpha)}(p,q) = K(p,q) = K^*(q,p) \tag{19}$$

$$\lim_{\alpha \to 1} \mathcal{A}^{(\alpha)}(p,q) = K^*(p,q) = K(q,p) \tag{20}$$

Note that traditionally (see [2,20]), the term $\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q$ is replaced by 1 in the integrand of Equation (18), and the term $q - p$ is absent in the integrand of Equation (16); this is trivially true when p, q are probability densities with a normalization of one. Zhu and Rohwer [42,43], in what they called the δ-divergence, $\delta = \frac{1-\alpha}{2}$, supplied these extra terms as the "extended" forms of α-divergence and of Kullback-Leibler divergence). The importance of these terms will be seen later (Section 2.2).

Note that, strictly speaking, when the underlying space is a finite-dimensional vector space, that is, the space, $\mathbb{R}^n$, for the parameters, θ, of a statistical model, $p(\cdot|\theta)$, then the term "divergence function" is appropriate. However, if the underlying sample space is infinite-dimensional that may be uncountable, that is, the manifold, $\mathcal{M}_\mu$, of non-parametric probability densities, p and q, then the term "divergence

functional" seems more appropriate. The latter implicitly defines a divergence function (through pullback) if the probability densities are embedded into a finite-dimensional submanifold, $\mathcal{M}_\theta$, in the case of a parametric statistical model, $p(\cdot|\theta)$. As an example, for the exponential family Equation (6), the Kullback-Leibler divergence Equation (16) in terms of p and q, implicitly defines a divergence in terms of θ_p, θ_q, *i.e.*, the Bregman divergence Equation (17). In the following, we use the term divergence *function* when we intend to blur the distinction between whether it is defined on the finite-dimensional vector space or on the infinite-dimensional function space and, in the latter case, whether it is pulled back into the finite dimensional submanifold. We will, however, use the term divergence *functional* when we emphasize the infinite-dimensional setting sans parametric embedding.

In general, a divergence function (also called "contrast function") is non-negative for all p, q and vanishes only when $p = q$; it is assumed to be sufficiently smooth. A divergence function will induce a Riemannian metric, g, in the form of Equation (3) by its second order properties and a pair of conjugate connections, Γ, Γ^*, in the forms of Equations (4) and (5) by its third order properties (relations were first formulated by Eguchi [36,37], which we are going to review next).

1.2.2. Induced Dual Riemannian Geometry

Let $\mathcal{M}$ be a Riemannian manifold endowed with a metric tensor field, g, whose restriction to any point, p, is a symmetric, positive bilinear form, $\langle\ ,\ \rangle$, on $T_p(\mathcal{M}) \times T_p(\mathcal{M})$. Here, $T_p(\mathcal{M})$ denotes the space of all tangent vectors at the point, $p \in \mathcal{M}$, and $\Sigma(\mathcal{M})$ denotes the collection of all vector fields on $\mathcal{M}$. Then:

$$g(u, v) = \langle u, v \rangle \tag{21}$$

with $u, v \in \Sigma(\mathcal{M})$. Let $w \in \Sigma(\mathcal{M})$ be another vector field, and d_w denotes the directional derivative (of a function, vector field, *etc.*) along the direction corresponding to w (taken at any given point, p, if explicitly written out). An affine connection, ∇, is a map, $\Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \Sigma(\mathcal{M}), (w, u) \mapsto \nabla_w u$, that is linear in u, w, while $\mathcal{F}$-linear in w, but not in u. A pair of connections, ∇, ∇^*, are said to be *conjugate* to each other if:

$$d_w\, g(u, v) = \langle \nabla_w u, v \rangle + \langle u, \nabla^*_w v \rangle \tag{22}$$

or in component form, denoted by Γ, Γ^*:

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i} \tag{23}$$

The "contravariant" form, Γ^l_{ij}, of the affine connection defined by:

$$\nabla_{\partial_i} \partial_j = \sum_l \Gamma^l_{ij} \partial_l \tag{24}$$

is related to the "covariant" form, $\Gamma_{ij,k}$ through:

$$\sum_l g_{lk}\, \Gamma^l_{ij} = \Gamma_{ij,k} \tag{25}$$

The Riemannian metric, g, and conjugate connections, ∇, ∇^*, on a statistical manifold can be induced by a divergence function, $\mathcal{D} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$, which, by definition, satisfies:

(i) $\mathcal{D}(p,q) \geq 0 \ \forall p,q \in \mathcal{M}$ with equality holding iff $p = q$;
(ii) $(d_u)_p \, \mathcal{D}(p,q)|_{p=q} = (d_v)_q \, \mathcal{D}(p,q)|_{p=q} = 0$;
(iii) $-(d_u)_p (d_v)_q \, \mathcal{D}(p,q)|_{p=q}$ is positive definite;

where the subscript, p, q, means that the directional derivative is taken with respect to the first and second arguments in $\mathcal{D}(p,q)$, respectively, along the direction, u or v. Eguchi [36,37] showed that any such divergence function, $\mathcal{D}$, satisfying (i)–(iii) will induce a Riemannian metric, g, and a pair of connections, ∇, ∇^* via:

$$g(u,v) = -(d_u)_p \, (d_v)_q \, \mathcal{D}(p,q)|_{p=q} \tag{26}$$

$$\langle \nabla_w u, v \rangle = -(d_w)_p \, (d_u)_p \, (d_v)_q \, \mathcal{D}(p,q)|_{p=q} \tag{27}$$

$$\langle u, \nabla^*_w v \rangle = -(d_w)_q \, (d_v)_q \, (d_u)_p \, \mathcal{D}(p,q)|_{p=q} \tag{28}$$

In index-laden component forms, they are:

$$g_{ij} = -(\partial_i)_p \, (\partial_j)_q \, \mathcal{D}(p,q)|_{p=q} \tag{29}$$

$$\Gamma_{ij,k} \equiv \langle \nabla_{\partial_i} \partial_j, \partial_k \rangle = -(\partial_i)_p \, (\partial_j)_p \, (\partial_k)_q \mathcal{D}(p,q)|_{p=q} \tag{30}$$

$$\Gamma^*_{ij,k} \equiv \langle \partial_k, \nabla^*_{\partial_i} \partial_j \rangle = -(\partial_i)_q \, (\partial_j)_q \, (\partial_k)_p \mathcal{D}(p,q)|_{p=q} \tag{31}$$

Equations (26)–(28) in coordinate-free form, or Equations (29)–(31) in index-laden form, link a divergence function, $\mathcal{D}$, to the Riemannian metric, g, and conjugate connections, ∇, ∇^*; henceforth, they will be called the *Eguchi relation*. It is easily verifiable that they satisfy Equation (22) or Equation (23), respectively. These relations are the stepping stones going from a divergence function defining (generally) non-symmetric distances between a pair of points on a manifold at large to the dual Riemannian geometric structure on the same manifold in the small. To apply to the infinite-dimensional context, we provide a proof (in Section 4) for the coordinate-free version Equations (26)–(28). This will allow us to first construct divergence functional on the infinite-dimensional function space (the Kullback-Leibler divergence being a special example) and then derive explicit expressions for the non-parametric Riemannian metric and conjugate connections by explicating d_u, d_v, d_w.

1.3. Goals and Approach

Our goals in this paper are several-fold. First, we want to provide a unified perspective for the divergence functions encountered in the literature. There are two broad classes, those defined on the infinite-dimensional function space and those defined on the finite-dimensional vector space. The former class include the one-parameter family of α-divergence (equivalently, the δ-divergence in [42,43]), the family of Jensen difference related to the Shannon entropy function [44], both specializing to Kullback-Leibler divergence as a limiting case. The latter class includes the Bregman divergence [41], also called "geometric divergence" [32], which turns out to be identical to the "canonical divergence" [1] on a dually flat manifold expressed in a pair of biorthogonal coordinates; those coordinates are induced by a pair of conjugate convex functions via the Legendre–Fenchel transform [2,20]. [15] recently investigated an infinite-dimensional version of the Bregman divergence, called the U-divergence. It will be shown in this paper that all of the above-mentioned divergence functions can be understood

as convex inequalities associated with some real-valued, strictly convex function defined on $\mathbb{R}$ (for the infinite-dimensional case) or $\mathbb{R}^n$ (for the finite-dimensional case), with the convex mixture parameter assuming the role of α in the induced α-connection. Note that $\alpha \longleftrightarrow -\alpha$ in such divergence functions corresponds to an exchange of the two points the divergence functions measure (generally in a non-symmetric fashion), while $\alpha \longleftrightarrow -\alpha$ in the induced connections corresponds to the conjugacy operation for the pairing of two metric-compatible connections. Hence, our approach to divergence functions from convex analysis will address both of these aspects coherently, and an intimate relation between these two senses of duality is expected to emerge from our formulation (see below).

The second goal of our paper is to provide a more general form for the Fisher metric Equation (3) and the α-connections Equation (4) (or equivalently, Equations (14) and (15) under α-embedding), while still staying within the framework of [29] in characterizing statistical manifolds. One specific aim is to derive explicit expressions for the Fisher metric and α-connections for the infinite-dimensional case. In the past, infinite-dimensional expression for the α-connection $\nabla^{(\alpha)}$, as a mixture of $\nabla^{(1)}$ and $\nabla^{(-1)}$, has emerged, but was given only implicitly with their interpretations debated [22,23]. Our approach exploits the coordinate-free version of the Eguchi relations Equations (26)–(28) directly, and derives Fisher metric and α-connections from the general form of divergence functions mentioned in the last paragraph. The affine connection, $\nabla^{(\alpha)}$, is formulated as the covariant derivative, which is characterized by a bilinear form. Since our divergence functional will be defined on the infinite-dimensional manifold, $\mathcal{M}$, without restricting the underlying functions (individual points on $\mathcal{M}$) to be normalized and positively-valued, the affine connections we derive are expected to have zero Riemann curvature as those in the ambient space. From this perspective, statistical curvature (the curvature of a statistical manifold) can be viewed as an embedding curvature, that is, curvature arising out of restricting to the submanifold, $\mathcal{M}_\mu$, of normalized and positive-valued functions (*i.e.*, non-parametric statistical manifold), and further to the finite-dimensional submanifold $\mathcal{M}_\theta$ (*i.e.*, parametric statistical models).

Our third goal here is to clarify some fundamental issues in information geometry, including the meaning of duality and its relation to submanifold embedding. In its original development starting from [19], the flatness of the e-connection (or m-connection) is with respect to a particular family of density functions, namely, the exponential family (or mixture family). Later, Amari [2,20] generalized this observation to any α-family (*i.e.*, a density function that is, after denormalization, affine under α-embedding): the α-connection is flat (indeed, $\Gamma^{(\alpha)}_{ij,k}$ vanishes) for the α-affine manifold (which is reduced to the exponential model for $\alpha = 1$ and the mixture model for $\alpha = -1$). One may be led to infer that the α parameter in the α-connection and the α parameter in α-embedding are one and the same and, thereby, conclude that $\nabla^{(1)}$-flatness (or $\nabla^{(-1)}$-flatness) is exclusively associated with the exponential family expressed in its natural parameter (or the mixture family expressed in its mixture parameter). Here, we point out that these conclusions are unwarranted: the flatness of an α-connection and the embedding of a probability function into an affine submanifold under α-representation are two related, but separate, issues. We will show that the α-connections for the infinite-dimensional ambient manifold, $\mathcal{M}$, which contains the manifold of probability density functions, $\mathcal{M}_\mu$, as a submanifold, has zero (ambient) curvature for all α values. For finite-dimensional parametric statistical models, it is known that the α-connection will not in general have zero curvature even when $\alpha = \pm 1$. Here, we will give precise conditions under which $\nabla^{(\pm 1)}$ will be dually flat—*i.e.*, when the denormalized

statistical model can be affine embedded under any ρ-representation, where a strictly increasing function $\rho : \mathbb{R} \to \mathbb{R}$ generalizes the α-embedding function (13). In such cases, there exists a strictly convex potential function, akin to Equation (7), for the exponential statistical model, that will reduce the Fisher metric and α-connections to the forms of Equations (8) and (9). One may define the natural parameter and expectation parameter that are dual to each other and that form biorthogonal coordinates for the underlying manifold, just as for the exponential family.

Our analysis will clarify two different kinds of duality in information geometry, one related to the different status of a reference probability function and a comparison probability function (referential duality), the other related to the representation of each probability function via a pair of conjugate scaling (representational duality). Roughly speaking, the (± 1)-duality reflects the former, whereas the e/m-duality reflects the latter. Previously, they were non-distinguished; in our analysis, we are able to disambiguate these two senses of duality. For instance, we are able to devise a two-parameter family of divergence functions, where the two parameters play distinct roles in the induced geometry, one capturing referential duality and the other capturing representational duality. Interestingly, this two-parameter family of connections still takes the same form of the α-connection proper (with a single parameter), indicating that this extension is still within [29]'s conceptualization of dual connections in information geometry.

The technical challenge that we have to overcome in our derivations is doing calculus in the infinite-dimensional setting. Consider the set of measurable functions from $\mathcal{X}$ to $\mathbb{R}$, which, in the presence of charts modeled on (open) subsets, $\{E_i\}_{i \in I}$, of a Banach space, form a manifold, $\mathcal{M}$, of infinite dimension. Each point on $\mathcal{M}$ is a function, $p : \mathcal{X} \to \mathbb{R}$, over the sample space $\mathcal{X}$; and each chart, $\mathcal{U} \subset \mathcal{M}$, is afforded with a bijective map to the Banach space with a suitable norm (e.g., Orlicz space, as adopted by [21–24,45]). For non-parametric statistical models, [24] provided exponential charts modeled on Orlicz spaces, which was followed by the rest of the above-referenced works. We do not restrict ourselves to probability density functions and work, in general, with measurable functions (without positivity and normalization requirements); we treat probability functions as forming a submanifold in $\mathcal{M}$ defined by the positivity and normalization conditions. This approach gives us certain advantages in deriving, from divergence functions directly, the Riemannian geometry on $\mathcal{M}$, whereby $\mathcal{M}$ serves as an ambient space to embed a statistical manifold, $\mathcal{M}_\mu$, as a submanifold in a standard way (by restricting the tangent vector field of $\mathcal{M}$). The usual interpretation of the affine connection on $\mathcal{M}_\mu$ as the projection of a natural connection on $\mathcal{M}$ is then "borrowed" over from the finite-dimensional setting to this infinite-dimensional setting. Our approach followed that of [46], who treats the infinite dimensional manifold as a generic C^∞-Banach manifold and used the theory of canonical spray (and the Morse-Palais Lemma) to construct Riemannian metric and affine connections on such manifolds. However, we fell short of providing a topology on $\mathcal{M}$ as induced from the divergence functions and compare it with the one endowed by [24]. In particular, the conditions under which $\mathcal{M}_\mu$ forms a proper submanifold of $\mathcal{M}$ remain to be identified. Neither have we addressed topological issues for the well-definedness of conjugate connections on such infinite-dimensional manifolds. We refer the readers to [23], who investigated whether the entire family of α-connections is well-defined for $\mathcal{M}$ endowed with the same topology.

The structure of the rest of the paper is as follows. Section 2 will deal with information geometry under the infinite-dimensional setting and Section 3 under the finite-dimensional setting. For ease of presentation, results will be provided in the main text, while their proofs will be deferred to Section 4. Section 5 closes with a discussion of the implications of the current framework. A preliminary report of this work was presented to IGAIA2 (Tokyo) and appeared in [47].

2. Information Geometry on Infinite-Dimensional Function Space

In this section, we first review the basic apparatus of the differentiable manifold with particular emphasis paid to the infinite-dimensional (non-parametric) setting (Section 2.1). We then define a family of divergence functionals based on convex analysis (Section 2.2) and use them to induce the dual Riemannian geometry on the infinite-dimensional manifold (Section 2.3). The section is concluded with an investigation of a special case of homogeneous divergence, called (α, β)-divergence, in which the two parameters play distinct, but interrelated, roles for referential duality and representational duality, thereby generalizing the familiar α-divergence in a sensible way (Section 2.4).

2.1. Differentiable Manifold in the Infinite-Dimensional Setting

Let $\mathcal{U}$ be an open set on the base manifold, $\mathcal{M}$, containing a representative point, x_0, and $\mathrm{F} : \mathcal{U} \to \mathbb{R}$, a smooth function defined on this local patch, $\mathcal{U} \subset \mathcal{M}$. The set of smooth functions on $\mathcal{M}$ is denoted $\mathcal{F}(\mathcal{M})$. A curve, $t \mapsto x(t)$, on the manifold is a collection of points, $\{x(t) \in \mathcal{U} : t \in [0, 1]\}$, whereas a tangent vector (or simply "vector"), v at $x_0 \in \mathcal{U}$, represents an equivalent class of curves passing through $x_0 = x(0)$, all with the same direction and speed as specified by the vector, $v = \frac{dx}{dt}\big|_{t=0}$. We use $T_{x_0}(\mathcal{M})$ to denote the space of all tangent vectors ("tangent space") at a given x_0; it is obviously a vector space. The tangent manifold, $\mathcal{TM}$, is then the collection of tangent spaces for all points on $\mathcal{M}$: $\mathcal{TM} = \{\cup T_x(\mathcal{M}),\ x \in \mathcal{M}\}$. A vector field, $v(x)$, is the association of a vector, v, at each point, x, of the manifold, $\mathcal{M}$; it is a cross-section of $\mathcal{TM}$. The set of all smooth vector fields on $\mathcal{M}$ is denoted $\Sigma(\mathcal{M})$. The tangent vector, v, acting on a function, F, will yield a scalar, denoted $d_v\mathrm{F}$, called the *direction derivative* of F:

$$d_v\mathrm{F} = \lim_{t \to 0} \frac{1}{t} \left(\mathrm{F}(x(t)) - \mathrm{F}(x_0)\right) \tag{32}$$

The tangent vector, v, acting on a vector field, $u(x)$, is defined analogously:

$$d_v u = \lim_{t \to 0} \frac{1}{t} \left(u(x(t)) - u(x_0)\right) \tag{33}$$

In our setting, given a measure space, $(\mathcal{X}, \mu)$, where samples are drawn from the set $\mathcal{X}$ and μ is the background measure, we call any function that maps $\mathcal{X} \to \mathbb{R}$ a *ζ-function*. The set of all ζ-functions forms a vector space, where vector addition is point-wise: $(f_1 + f_2)(\zeta) = f_1(\zeta) + f_2(\zeta)$, and scalar multiplication is simple multiplication: $(cf)(\zeta) = cf(\zeta)$. We now consider the set of all ζ-functions with common support μ, which is assumed to form a manifold denoted as $\mathcal{M}_\mu$. A typical point of this manifold denotes a specific ζ-function, $p(\zeta) : \zeta \to p(\zeta)$, defined over $\mathcal{X}$, the sample space, which is infinite dimensional or even uncountable in general. Under suitable topology (e.g., [24]), all points, $\mathcal{M}_\mu$, form a manifold. On this manifold, any function, $\mathrm{F} : p \to \mathrm{F}(p)$, is referred to (in this paper)

as a ζ-*functional*, because it takes in a ζ-function $p(\cdot)$ and outputs a scalar. The set of ζ-functionals on $\mathcal{M}$ is denoted $\mathcal{F}(\mathcal{M})$. (Note that "ζ-function" and "ζ-functional" are both functions (also called "maps" or "mappings") in the mathematical sense, with pre-specified domains and ranges. We make the distinction that the ζ-function refers to a real-valued function (e.g., density functions, random variables) defined on the sample space, $\mathcal{X}$, and ζ-functional refers to a mapping from one or more ζ-functions to a real number.) A curve on $\mathcal{M}$ passing through a typical point, p, is nothing but a one-parameter *family* of ζ-functions, denoted as $p(\zeta|t)$, with $p(\zeta|0) = p$. Here, $\cdot|t$ is read as "given t", "indexed by t", "parameterized" by t—a one-parameter family of ζ-functions, $p(\zeta|t)$, is formed as t varies. For each fixed t, $p(\zeta|t)$ is a function, $\mathcal{X} \times I \to \mathbb{R}$. More generally, $p(\zeta|\theta)$, where $\theta = [\theta^1, \cdots, \theta^n] \subseteq \mathbb{R}^n$, is a ζ-function indexed by n parameters, $\theta^1, \cdots, \theta^n$. As θ varies, $p(\zeta|\theta)$ represents a finite dimensional submanifold, $\widetilde{\mathcal{M}}_\theta \subset \mathcal{M}$ where:

$$\widetilde{\mathcal{M}}_\theta = \{p(\zeta|\theta) \in \mathcal{M} : \theta \subseteq \mathbb{R}^n\} \subset \mathcal{M} \tag{34}$$

In this paper, they are referred to as *parametric models* (and *parametric statistical model* if $p(\zeta|\theta)$ is normalized and positive-valued).

In the infinite-dimensional setting, the following tangent vector, v:

$$v(\zeta) = \left.\frac{\partial p(\zeta|t)}{\partial t}\right|_{t=0} \tag{35}$$

is also a ζ-function. When the tangent vector, v, operates on the ζ-functional $\mathrm{F}(p)$:

$$d_v(\mathrm{F}(p)) = \lim_{t\to 0} \frac{\mathrm{F}(p(\cdot|t)) - \mathrm{F}(p(\cdot|0))}{t} \tag{36}$$

the outcome is another ζ-functional of both $p(\zeta)$ and $v(\zeta)$ and linear in the latter. A particular ζ-functional of interest in this paper is of the following form:

$$\mathrm{F}(p) = \int_{\mathcal{X}} f(p(\zeta))d\mu = \mathrm{E}_\mu\{f(p(\cdot))\} \tag{37}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a strictly convex function defined on the real line. In this case, $p(\zeta|t) = p(\zeta) + v(\zeta)t + o(t^2)$, so:

$$d_v(\mathrm{F}(p)) = \int_{\mathcal{X}} f'(p(\zeta))\, v(\zeta)\, d\mu \tag{38}$$

which is linear in $v(\cdot)$.

A vector field, as a cross-section of $\mathcal{TM}$, takes $p(\zeta)$ and associates a ζ-function. We denote a vector field as $u(\zeta|p) \in \Sigma(\mathcal{M})$, where the variable following the "|" sign indicates that u depends on the point, $p(\zeta)$, an element of the base manifold, $\mathcal{M}$ (we could also write it as $u(p(\zeta))(\zeta)$ or $u_p(\zeta)$). Though the vector fields defined above are not necessarily smooth, we will concentrate on smooth ones below. Of particular interest to us is the vector field, $\rho(p(\zeta))$, for some strictly increasing function, $\rho : \mathbb{R} \to \mathbb{R}$.

Differentiation of smooth vector fields can be defined analogously. The directional derivative, $d_v u$, of a vector field, $u(\zeta|p)$, which is a ζ-function also dependent on $p(\zeta)$, in the direction of $v = v(\zeta)$, which is another ζ-function, is:

$$d_v u(\zeta|p) = \lim_{t\to 0} \frac{u(\zeta|p(\zeta|t)) - u(\zeta|p(\zeta))}{t} \tag{39}$$

Note that $d_v u$ is another ζ-function; that is why we can write $d_v u(\zeta|p)$ also as $(d_v u)(\zeta)$. As an example, the derivative of the vector field, $\rho(p(\zeta))$, where $\rho : \mathbb{R} \to \mathbb{R}$, in the direction of $v(\zeta)$ is:

$$d_v \rho(p(\zeta|t)) = \lim_{t \to 0} \frac{\rho(p(\zeta|t)) - \rho(p)}{t} \; = \; \rho'(p(\zeta))\, v(\zeta) \tag{40}$$

With differentiation of vector fields defined, one can define the covariant derivative operation, ∇_w. When operating on a ζ-functional, the covariant derivative is simply the directional derivative (along direction w):

$$\nabla_w \mathrm{F}(p) = d_w \mathrm{F}(p) \tag{41}$$

when operating on a vector field, say $u(\zeta|p)$, ∇_w is defined as (see [46]):

$$\nabla_w u = d_w u + \mathrm{B}(\cdot|w(\cdot|p), u(\cdot|p)) \tag{42}$$

where $\mathrm{B} : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \Sigma(\mathcal{M})$ is a ζ-function, which is bilinear in the two tangent vectors (ζ-functions), w and u; it is the infinite-dimensional counterpart of the Christoffel symbol, Γ (for finite dimensions). We denote the conjugate covariant derivative, ∇^*_w (as defined by Equation (22)) in terms of B^* (with an asterisk denoting conjugacy):

$$(\nabla^*_w u)(\zeta) = (d_w u)(\zeta) + \mathrm{B}^*(\zeta|w(\zeta|p(\zeta)), u(\zeta|p(\zeta))) \tag{43}$$

(here, we write out the explicit dependency on ζ).

The Riemann curvature tensor, R, which measures the curvature of a connection, ∇ (as specified by B), is defined by the map, $\Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \Sigma(\mathcal{M})$:

$$R(u, v, w) = R(u, v)w = \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w \tag{44}$$

where:

$$[u, v] = d_u v - d_v u \tag{45}$$

The torsion tensor, $T : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \Sigma(\mathcal{M})$, is given by:

$$T(u, v) = \nabla_u v - \nabla_v u - [u, v] \tag{46}$$

2.2. $\mathcal{D}^{(\alpha)}$-Divergence, a Family of Generalized Divergence Functionals

Divergence functionals are defined with respect to a pair of ζ-functions in an infinite-dimensional function space. A divergence functional, $\mathcal{D} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$, maps two ζ-functions to a non-negative real number. To the extent that ζ-functions can be parameterized by finite-dimensional vectors, $\theta \subseteq \mathbb{R}^n$, a divergence *functional* on $\mathcal{M} \times \mathcal{M}$ will implicitly induce a divergence *function* on the parameter space, which is a subset of $\mathbb{R}^n \times \mathbb{R}^n$. In this section, we will discuss the general form of the divergence functional and the associated infinite-dimensional manifold. Finite-dimensional embedding of ζ-functions (*i.e.*, parametric models) will be discussed in Section 3.

2.2.1. Fundamental Convex Inequality and Divergence

We start our exposition by reviewing the notion of a convex function on the real line, $f : \mathbb{R} \to \mathbb{R}$. We recall the *fundamental convex inequality* that defines a strictly convex function, f:

$$f\left(\frac{1-\alpha}{2}\,\gamma + \frac{1+\alpha}{2}\,\delta\right) \leq \frac{1-\alpha}{2}\,f(\gamma) + \frac{1+\alpha}{2}\,f(\delta) \tag{47}$$

for all $\gamma, \delta \in \mathbb{R}$, with equality holding, if and only if $\gamma = \delta$, for all $\alpha \in (-1, 1)$. Geometrically, the value of the function, f, at any point, ϵ, in between two end points, γ and δ, lies on or below the chord connecting its values at these two points. This property of a strictly convex function can also be stated in elementary algebra as the *Chord Theorem*, namely:

$$\frac{f(\epsilon) - f(\gamma)}{\epsilon - \gamma} \leq \frac{f(\delta) - f(\gamma)}{\delta - \gamma} \leq \frac{f(\delta) - f(\epsilon)}{\delta - \epsilon} \tag{48}$$

where:

$$\epsilon = \frac{1-\alpha}{2}\gamma + \frac{1+\alpha}{2}\delta \tag{49}$$

(here, we assumed $\gamma \leq \epsilon \leq \delta$ without loss of generality). In fact, the slope, $\frac{f(\delta)-f(\gamma)}{\delta-\gamma}$, is an increasing function in both δ and γ. The slopes for the chords connecting from the midpoint to either end point are, respectively:

$$L^{(\alpha)}(\gamma, \delta) = \frac{f(\delta) - f(\epsilon)}{\delta - \epsilon} = \frac{1}{\delta - \gamma}\frac{2}{1-\alpha}\left(f(\delta) - f\left(\frac{1-\alpha}{2}\gamma + \frac{1+\alpha}{2}\delta\right)\right) \tag{50}$$

$$\hat{L}^{(\alpha)}(\gamma, \delta) = \frac{f(\gamma) - f(\epsilon)}{\gamma - \epsilon} = \frac{1}{\delta - \gamma}\frac{2}{1+\alpha}\left(f\left(\frac{1-\alpha}{2}\gamma + \frac{1+\alpha}{2}\delta\right) - f(\gamma)\right) \tag{51}$$

with skew symmetry:

$$L^{(-\alpha)}(\gamma, \delta) = -\hat{L}^{(\alpha)}(\delta, \gamma)\ , \qquad \hat{L}^{(-\alpha)}(\gamma, \delta) = -L^{(\alpha)}(\delta, \gamma) \tag{52}$$

As $\alpha : -1 \to 1$ (*i.e.*, as point ϵ moves from γ to δ, the two fixed ends), both $L^{(\alpha)}(\gamma, \delta)$ and $\hat{L}^{(\alpha)}(\gamma, \delta)$, are increasing functions of α, but the chord theorem dictates that the latter is always no greater than the former. In fact, their difference has a non-negative value:

$$\begin{aligned} 0 \leq L^{(\alpha)}(\gamma, \delta) - \hat{L}^{(\alpha)}(\gamma, \delta) &= L^{(-\alpha)}(\delta, \gamma) - \hat{L}^{(-\alpha)}(\delta, \gamma) \\ &= \frac{1}{\delta - \gamma}\frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}f(\gamma) + \frac{1+\alpha}{2}f(\delta) - f\left(\frac{1-\alpha}{2}\gamma + \frac{1+\alpha}{2}\delta\right)\right) \end{aligned} \tag{53}$$

Though the above is obviously valid for $\alpha \in [-1, 1]$, it can be shown that it is also valid for any $\alpha \in \mathbb{R}$.

The fundamental convex inequality applies to any two real numbers, γ, δ. We can treat γ, δ as the values of two functions, $p, q : \mathcal{X} \to \mathbb{R}$, evaluated at any particular sample point, ζ, that is, $\gamma = p(\zeta)$, $\delta = q(\zeta)$. This allows us to define the following family of divergence functionals (see [48]).

PROPOSITION 1 Let $f : \mathbb{R} \to \mathbb{R}$ be smooth and strictly convex, and $\rho : \mathbb{R} \to \mathbb{R}$ be strictly increasing. For any two ζ-functions, p, q, and any $\alpha \in \mathbb{R}$:

$$\mathcal{D}_{f,\rho}^{(\alpha)}(p, q) = \frac{4}{1-\alpha^2}\,\mathrm{E}_\mu\left\{\frac{1-\alpha}{2}f(\rho(p)) + \frac{1+\alpha}{2}f(\rho(q)) - f\left(\frac{1-\alpha}{2}\rho(p) + \frac{1+\alpha}{2}\rho(q)\right)\right\} \tag{54}$$

is non-negative and equals zero, if and only

$$p(\zeta) = q(\zeta) \quad \text{almost} \quad \text{everywhere} \tag{55}$$

Proof. See Section 4.

Proposition 1 constructed a family (parameterized by α) of divergence functionals, $\mathcal{D}^{(\alpha)}$, for two ζ-functions, in which representational duality is embodied as:

$$\mathcal{D}^{(\alpha)}_{f,\rho}(p,q) = \mathcal{D}^{(-\alpha)}_{f,\rho}(q,p) \tag{56}$$

Its definition involves a strictly increasing function ρ, which can be taken to be the identity function if necessary. The reason ρ is introduced will be clear in the next subsection, where we introduce the notion of conjugate-scaled representations. Furthermore, in order to ensure that the integrals in Equation (54) are well defined, we require p, q to be elements of the set:

$$\{p(\zeta) : \mathrm{E}_\mu\{f(\rho(p))\} < \infty\} \tag{57}$$

$\mathcal{D}^{(\alpha)}$-divergence was first introduced in [48]. It generalized the familiar α-divergence Equation (18): take $f(p) = e^p$ and $\rho(p) = \log p$; then $\mathcal{D}^{(\alpha)}_{f,\rho}(p,q) = \mathcal{A}^{(\alpha)}(p,q)$. $\mathcal{D}^{(\alpha)}$-divergence became the U-divergence [15] when $f(p) = U(p), \rho(p) = (U')^{-1}(p), \alpha \to 1$ for any strictly convex and strictly increasing $U : \mathbb{R}_+ \to \mathbb{R}$. It was well known that U-divergence, when taking

$$U(t) = \frac{1}{\beta}(1 + (\beta - 1)t)^{\frac{\beta}{\beta - 1}} \quad (\beta \neq 0, 1) \tag{58}$$

specializes to β-divergence [49], defined as:

$$\mathcal{B}^{(\beta)}(p,q) = \mathrm{E}_\mu \left\{ p \frac{p^{\beta-1} - q^{\beta-1}}{\beta - 1} - \frac{p^\beta - q^\beta}{\beta} \right\} \tag{59}$$

and that both α- and β-divergence specialize to the Kullback-Leibler divergence as $\alpha \to \pm 1$ and $\beta \to 1$, respectively.

2.2.2. Conjugate-Scaled Representations of Measurable Functions

In one-dimension, any strictly convex function, $f : \mathbb{R} \to \mathbb{R}$, can be written as an integral of a strictly increasing function, g, and v*ice versa*: $f(\delta) = \int_\gamma^\delta g(t)dt + f(\gamma)$, with $g'(t) > 0$. The convex (Legendre–Fenchel) conjugate, $f^* : \mathbb{R} \to \mathbb{R}$, defined by:

$$f^*(t) = t\,(f')^{-1}(t) - f((f')^{-1}(t)) \tag{60}$$

has the integral expression $f^*(\lambda) = \int_{g(\gamma)}^\lambda g^{-1}(t)dt + f^*(g(\gamma))$, with g^{-1} also strictly monotonic and $\gamma, \delta, \lambda \in \mathbb{R}$. (Here, the monotonicity condition replaces the requirement of a positive semi-definite Hessian in the case of a convex function of several variables.) The Legendre–Fenchel inequality:

$$f(\delta) + f^*(\lambda) - \gamma\,\lambda \geq 0 \tag{61}$$

can be cast as the Young's inequality:

$$\int_\gamma^\delta g(t)\,dt + \int_{g(\gamma)}^\lambda g^{-1}(t)\,dt + \gamma\,g(\gamma) \geq \delta\,\lambda \tag{62}$$

with equality holding, if and only if $\lambda = g(\delta)$. The conjugate function, f^*, which is also strictly convex, satisfies $(f^*)^* = f$ and $(f^*)' = (f')^{-1}$.

We introduce the notion of ρ-representation of a ζ-function $p(\cdot)$ by defining a mapping, $p \mapsto \rho(p)$, for a strictly increasing function, $\rho : \mathbb{R} \to \mathbb{R}$. We say that a τ-representation of a ζ-function, $p \mapsto \tau(p)$, is conjugate to the ρ-representation *with respect to* a smooth and strictly convex function, $f : \mathbb{R} \to \mathbb{R}$, if:

$$\tau(p) = f'(\rho(p)) = ((f^*)')^{-1}(\rho(p)) \longleftrightarrow \rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p)) \tag{63}$$

As an example, we may let $\rho(p) = l^{(\alpha)}(p)$ be the *α-representation*, where $l^{(\alpha)}$ is given by Equation (13), and the conjugate representation is the $(-\alpha)$-representation $\tau(p) = l^{(-\alpha)}(p)$:

$$\rho(t) = l^{(\alpha)}(t) \longleftrightarrow \tau(p) = l^{(-\alpha)}(p) \tag{64}$$

In this case:

$$f(t) = \frac{2}{1+\alpha}\left(\left(\frac{1-\alpha}{2}\right)t\right)^{\frac{2}{1-\alpha}} , \quad f^*(t) = \frac{2}{1-\alpha}\left(\left(\frac{1+\alpha}{2}\right)t\right)^{\frac{2}{1+\alpha}} \tag{65}$$

so that:

$$f(\rho(p)) = \frac{2}{1+\alpha}p , \quad f^*(\tau(p)) = \frac{2}{1-\alpha}p \tag{66}$$

both linear in p. More generally, strictly increasing functions from $\mathbb{R} \to \mathbb{R}$ form a group, with functional composition as group composition operation and the functional inverse as the group inverse operation. That is, (i) for any two strictly increasing functions, ρ_1, ρ_2, their functional composition $\rho_2 \circ \rho_1$ is strictly increasing; (ii) the functional inverse, ρ^{-1}, of any strictly increasing function, ρ, is also strictly increasing; (iii) there exists a strictly increasing function, ι, the identity function, such that $\rho \circ \rho^{-1} = \rho^{-1} \circ \rho = \iota$. From this perspective, $f' = \tau \circ \rho^{-1}$, $(f^*)' = \rho \circ \tau^{-1}$, encountered above, are themselves two mutually inverse strictly increasing functions.

If, in the above discussions, $f' = \tau \circ \rho^{-1}$ is further assumed to be strictly convex, that is:

$$\frac{1-\alpha}{2}\tau(\rho^{-1}(\gamma)) + \frac{1+\alpha}{2}\tau(\rho^{-1}(\delta)) \geq \tau\left(\rho^{-1}\left(\frac{1-\alpha}{2}\gamma + \frac{1+\alpha}{2}\delta\right)\right) \tag{67}$$

for any $\gamma, \delta \in \mathbb{R}$ and $\alpha \in (-1, 1)$, then by taking τ^{-1} on both sides of the inequality and renaming $\rho^{-1}(\gamma)$ as γ and $\rho^{-1}(\delta)$ as δ, we obtain:

$$\tau^{-1}\left(\frac{1-\alpha}{2}\tau(\gamma) + \frac{1+\alpha}{2}\tau(\delta)\right) \geq \rho^{-1}\left(\frac{1-\alpha}{2}\rho(\gamma) + \frac{1+\alpha}{2}\rho(\delta)\right) \tag{68}$$

This is to say:

$$M_\tau^{(\alpha)}(\gamma, \delta) \geq M_\rho^{(\alpha)}(\gamma, \delta) \tag{69}$$

with equality holding, if and only if $\gamma = \delta$, where:

$$M_\rho^{(\alpha)}(\gamma, \delta) = \rho^{-1}\left(\frac{1-\alpha}{2}\rho(\gamma) + \frac{1+\alpha}{2}\rho(\delta)\right) \tag{70}$$

is the quasi-linear mean of two numbers γ, δ. Therefore, the following is also a divergence functional (see more discussions in Section 2.4)

$$\frac{4}{1-\alpha^2}\int_{\mathcal{X}}\left\{\tau^{-1}\left(\frac{1-\alpha}{2}\tau(p(\zeta)) + \frac{1+\alpha}{2}\tau(q(\zeta))\right) - \rho^{-1}\left(\frac{1-\alpha}{2}\rho(p(\zeta)) + \frac{1+\alpha}{2}\rho(q(\zeta))\right)\right\} d\mu \tag{71}$$

2.2.3. Canonical Divergence

The use of a pair of strictly increasing functions, f, f^*, allow us to define, in parallel with $\mathcal{D}^{(\alpha)}_{f,\rho}(p,q)$ given in Equation (54), the conjugate family, $\mathcal{D}^{(\alpha)}_{f^*,\tau}(p,q)$. The two families turn out to have the same form when $\alpha = \pm 1$; this is the so-called *canonical divergence*.

Taking the limit, $\alpha \to -1$, the inequality Equation (53) becomes:

$$\frac{f(\delta) - f(\gamma)}{\delta - \gamma} - f'(\gamma) \geq 0 \tag{72}$$

where f is strictly convex. A similar inequality is obtained when $\alpha \to 1$. Hence, the divergence functionals, $\mathcal{D}^{(\pm 1)}_{f,\rho}(p,q)$, take the form:

$$\begin{aligned}
\mathcal{D}^{(-1)}_{f,\rho}(p,q) &= \mathrm{E}_\mu\{f(\rho(q)) - f(\rho(p)) - (\rho(q) - \rho(p))f'(\rho(p))\} && (73)\\
&= \mathrm{E}_\mu\{f^*(\tau(p)) - f^*(\tau(q)) - (\tau(p) - \tau(q))(f^*)'(\tau(q))\} = \mathcal{D}^{(-1)}_{f^*,\tau}(q,p) && (74)\\
\mathcal{D}^{(1)}_{f,\rho}(p,q) &= \mathrm{E}_\mu\{f(\rho(p)) - f(\rho(q)) - (\rho(p) - \rho(q))f'(\rho(q))\} && (75)\\
&= \mathrm{E}_\mu\{f^*(\tau(q)) - f^*(\tau(p)) - (\tau(q) - \tau(p))(f^*)'(\tau(p))\} = \mathcal{D}^{(1)}_{f^*,\tau}(q,p) && (76)
\end{aligned}$$

The canonical divergence functional, $\mathcal{A} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$, is defined (with the aid of a pair of conjugate representations) as:

$$\mathcal{A}_f(\rho(p), \tau(q)) = \mathrm{E}_\mu\{f(\rho(p)) + f^*(\tau(q)) - \rho(p)\,\tau(q)\} = \mathcal{A}_{f^*}(\tau(q), \rho(p)) \tag{77}$$

where $\int_{\mathcal{X}} f(\rho(p))d\mu$ can be called the (generalized) cumulant generating functional and $\int_{\mathcal{X}} f^*(\tau(p))d\mu$, the (generalized) entropy functional. Thus, a dualistic relation exists between $\alpha = 1 \longleftrightarrow \alpha = -1$ and between $(f,\rho) \longleftrightarrow (f^*,\tau)$:

$$\begin{aligned}
\mathcal{D}^{(1)}_{f,\rho}(p,q) &= \mathcal{D}^{(-1)}_{f,\rho}(q,p) = \mathcal{D}^{(1)}_{f^*,\tau}(q,p) = \mathcal{D}^{(-1)}_{f^*,\tau}(p,q) && (78)\\
&= \mathcal{A}_f(\rho(p), \tau(q)) = \mathcal{A}_{f^*}(\tau(q), \rho(p)) && (79)
\end{aligned}$$

We can see that under conjugate $(\pm\alpha)$-representations Equation (64), $\mathcal{A}_f$ is simply the α-divergence proper $\mathcal{A}^{(\alpha)}$:

$$\mathcal{A}_f(\rho(p), \tau(q)) = \mathcal{A}^{(\alpha)}(p,q) \tag{80}$$

In fact:

$$\frac{1-\alpha^2}{4}\mathcal{A}^{(\alpha)}(u,v) = \mathrm{E}_\mu\left\{\frac{1-\alpha}{2}\,u^{\frac{2}{1-\alpha}} + \frac{1+\alpha}{2}\,v^{\frac{2}{1+\alpha}} - u\,v\right\} \geq 0 \tag{81}$$

is an expression of Young's inequality between two functions $u = (l^{(\alpha)})^{-1}(p), v = (l^{(-\alpha)})^{-1}(q)$ under conjugate exponents, $\frac{2}{1-\alpha}$ and $\frac{2}{1+\alpha}$.

2.3. Geometry Induced by the $\mathcal{D}^{(\alpha)}$-Divergence

The last two sections showed that the divergence functional, $\mathcal{D}^{(\alpha)}$, we constructed on $\mathcal{M}$ according to Equation (54) generalizes the α-divergence in a sensible way. Now, we investigate the metric and conjugate connections that such divergence functionals induce; this is accomplished by invoking Eguchi relations Equations (26)–(28).

PROPOSITION 2. At any given $p \in \mathcal{M}$ and for any vector fields, $u, v \in \Sigma(\mathcal{M})$:

(i) the metric tensor field, $g : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \mathcal{F}(\mathcal{M})$, is given by:

$$g(u, v) = \mathrm{E}_\mu\{g(p(\zeta))\, u(\zeta|p(\zeta))\, v(\zeta|p(\zeta))\} \tag{82}$$

where:

$$g(t) = f''(\rho(t))(\rho'(t))^2 \tag{83}$$

(ii) the family of covariant derivatives (connections) $\nabla^{(\alpha)} : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \Sigma(\mathcal{M})$ is given as:

$$\nabla^{(\alpha)}_w u = (d_w u)(\zeta) + b^{(\alpha)}(p(\zeta))u(\zeta|p(\zeta))w(\zeta|p(\zeta)) \tag{84}$$

where:

$$b^{(\alpha)}(t) = \frac{1-\alpha}{2}\frac{f'''(\rho(t))\rho'(t)}{f''(\rho(t))} + \frac{\rho''(t)}{\rho'(t)} \tag{85}$$

(iii) the family of conjugate covariant derivatives is:

$$\nabla^{*(\alpha)}_w u = (d_w u)(\zeta) + b^{(-\alpha)}(p(\zeta))u(\zeta|p(\zeta))w(\zeta|p(\zeta)) \tag{86}$$

Proof. See Section 4.

Note that the $g(\cdot)$ term in Equation (82) and the $b^{(\alpha)}(\cdot)$ term in covariant derivatives Equation (84) depend on p, the point on the base manifold, where the metric and covariant derivatives are evaluated. They both depend on the auxiliary "scaling functions", f and ρ. We may cast them into an equivalent, dually symmetric form as follows.

COROLLARY 3. The $g(\cdot)$ function in expressing the metric Equation (82) and $b^{(\alpha)}(\cdot)$ in expressing the covariant derivatives Equation (84) can be expressed in dualistic forms:

$$g(t) = \rho'(t)\,\tau'(t) \tag{87}$$

and:

$$b^{(\alpha)}(t) = \frac{d}{dt}\left(\frac{1+\alpha}{2}\log\rho'(t) + \frac{1-\alpha}{2}\log\tau'(t)\right) \tag{88}$$

Proof. See Section 4.

Corollary 3 makes it immediately evident that the Riemannian metrics induced by $\mathcal{D}^{(\alpha)}_{f,\rho}(p, q)$ and by $\mathcal{D}^{(\alpha)}_{f^*,\tau}(p, q)$ are identical for all α values, while the connections (covariant derivatives) induced by the two families of divergence are conjugate to each other, expressed as $\alpha \longleftrightarrow -\alpha$. This implies that the conjugacy embodied by the definition of the pair of connections is related to both referential duality and representational duality.

It can be proven that the covariant derivative of the kind of Equation (84) are both curvature-free and torsion-free.

PROPOSITION 4. For the entire family of covariant derivatives indexed by α ($\alpha \in \mathbb{R}$):

(i) the Riemann curvature tensor $R^{(\alpha)}(u, v, w) \equiv 0$;
(ii) the torsion tensor $T^{(\alpha)}(u, v) \equiv 0$.

Proof. See Section 4.

In other words, the manifold, $\mathcal{M}$, has zero-curvature and zero-torsion for all α. As such, it can serve as an ambient manifold to embed the manifold, $\mathcal{M}_\mu$, of non-parametric probability density functions and the manifold, $\mathcal{M}_\theta$, of parametric density functions, and any curvature on $\mathcal{M}_\mu$ or $\mathcal{M}_\theta$ may be interpreted as arising from embedding or restriction to a lower dimensional space. See, also, [50] for a discussion of curvatures of statistical manifolds.

2.4. Homogeneous (α, β)-Divergence and the Induced Geometry

Suppose that f is, in addition to being strictly convex, strictly increasing. We may set $\rho(t) = f^{-1}(\varepsilon t) \longleftrightarrow f(t) = \varepsilon \rho^{-1}(t)$, so that the divergence functional becomes:

$$\mathcal{D}_\rho^{(\alpha)}(p,q) = \frac{4\,\varepsilon}{1-\alpha^2}\int_{\mathcal{X}} \left\{ \frac{1-\alpha}{2}p(\zeta) + \frac{1+\alpha}{2}q(\zeta) - \rho^{-1}\left(\frac{1-\alpha}{2}\rho(p(\zeta)) + \frac{1+\alpha}{2}\rho(q(\zeta)) \right) \right\} d\mu \tag{89}$$

Now, the second term in the integrand is just the quasi-linear mean, $M_\rho^{(\alpha)}$, introduced in Equation (70), where ρ is strictly increasing and concave here. As an example, take $\rho(p) = \log p$, $\epsilon = 1$; then $M_\rho^{(\alpha)}(p,q) = p^{\frac{1-\alpha}{2}} q^{\frac{1+\alpha}{2}}$, and $\mathcal{D}_\rho^{(\alpha)}(p,q)$ is the α-divergence Equation (18), while:

$$\mathcal{D}_\rho^{(1)}(p,q) = \int_{\mathcal{X}} (p - q - (\rho(p) - \rho(q)))\,(\rho^{-1})'(\rho(q))\,d\mu = \mathcal{D}_\rho^{(-1)}(q,p) \tag{90}$$

is an immediate generalization of the extended Kullback-Leibler divergence in Equation (16).

If we impose a homogeneous requirement ($\kappa \in \mathbb{R}^+$) on $\mathcal{D}_\rho^{(\alpha)}$:

$$\mathcal{D}_\rho^{(\alpha)}(\kappa p, \kappa q) = \kappa \mathcal{D}_\rho^{(\alpha)}(p,q) \tag{91}$$

then (see [48]) $\rho(p) = l^{(\beta)}(p)$; so Equation (89) becomes a two-parameter family:

$$\mathcal{D}^{(\alpha,\beta)}(p,q) \equiv \frac{4}{1-\alpha^2}\frac{2}{1+\beta}\mathrm{E}_\mu \left\{ \frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q - \left(\frac{1-\alpha}{2}p^{\frac{1-\beta}{2}} + \frac{1+\alpha}{2}q^{\frac{1-\beta}{2}} \right)^{\frac{2}{1-\beta}} \right\} \tag{92}$$

Here $(\alpha, \beta) \in [-1,1] \times [-1,1]$, and $\varepsilon = 2/(1+\beta)$ in Equation (89) is chosen to make $\mathcal{D}^{(\alpha,\beta)}(p,q)$ well defined for $\beta = -1$. We call this family (α, β)-*divergence*; it belongs to the general class of f-divergence studied by [51]. Note that the α parameter encodes referential duality, and the β parameter encodes representational duality. When *either* $\alpha = \pm 1$ *or* $\beta = \pm 1$, the one-parameter version of the generic alpha-connection results. The family, $\mathcal{D}^{(\alpha,\beta)}$, is then a generalization of Amari's α-divergence Equation (18) with:

$$\lim_{\alpha \to -1} \mathcal{D}^{(\alpha,\beta)}(p,q) = \mathcal{A}^{(-\beta)}(p,q) \tag{93}$$

$$\lim_{\alpha \to 1} \mathcal{D}^{(\alpha,\beta)}(p,q) = \mathcal{A}^{(\beta)}(p,q) \tag{94}$$

$$\lim_{\beta \to 1} \mathcal{D}^{(\alpha,\beta)}(p,q) = \mathcal{A}^{(\alpha)}(p,q) \tag{95}$$

$$\lim_{\beta \to -1} \mathcal{D}^{(\alpha,\beta)}(p,q) = J^{(\alpha)}(p,q) \tag{96}$$

where $J^{(\alpha)}$ denotes the Jensen difference discussed by [44]:

$$J^{(\alpha)}(p,q) \equiv \frac{4}{1-\alpha^2}\,\mathrm{E}_\mu\left(\frac{1-\alpha}{2}p\log p+\frac{1+\alpha}{2}q\log q - \left(\frac{1-\alpha}{2}p+\frac{1+\alpha}{2}q\right)\log\left(\frac{1-\alpha}{2}p+\frac{1+\alpha}{2}q\right)\right) \tag{97}$$

$J^{(\alpha)}$ reduces to Kullback-Leibler divergence Equation (16) when $\alpha \to \pm 1$. Lastly, we note that in $\mathcal{D}^{(\alpha,\beta)}$, when either α or β equals zero, the Levi-Civita connection results.

Note that the divergence given by Equation (92), which first appeared in [48], was called "(α,β)-divergence" in [47]. Cichocki *et al.* [52], following their review of α,β,γ divergence by [53], introduced the following two-parameter family:

$$D_{AB}^{\alpha,\beta} = -\frac{1}{\alpha\beta}\,\mathrm{E}_\mu\left\{p^\alpha q^\beta - \frac{\alpha}{\alpha+\beta}p^{\alpha+\beta} - \frac{\beta}{\alpha+\beta}q^{\alpha+\beta}\right\} \tag{98}$$

and called it (α,β)-divergence. Essentially, it is α-divergence under β- (power) embedding:

$$D_{AB}^{\alpha,\beta} = (\alpha+\beta)^2\,\mathcal{A}^{\frac{\beta-\alpha}{\alpha-\beta}}(p^{\alpha+\beta}, q^{\alpha+\beta}) \tag{99}$$

Clearly, by taking $f(t)=e^t, \rho(t)=(\alpha+\beta)\log t$ and renaming $\frac{\beta-\alpha}{\alpha+\beta}$ as α, $D_{AB}^{\alpha,\beta}$ is a special case of the $\mathcal{D}_{f,\rho}^{(\alpha)}(p,q)$, as is $\mathcal{D}^{(\alpha,\beta)}(p,q)$ by Zhang [47,48]. The two definitions of (α,β)-divergences, both special cases of $\mathcal{D}_{f,\rho}^{(\alpha)}(p,q)$, only differ by a $l^{(\beta)}$-embedding in the density functions, leading to a superficial difference in the homogeneity/scaling of the divergence function.

With respect to the geometry induced from the (α,β)-divergence of Equation (92), we have the following result.

PROPOSITION 5. The metric g and affine connections (covariant derivatives) $\nabla^{(\alpha,\beta)}$ corresponding to the (α,β)-divergence are given by:

$$g(u,v) = \int_{\mathcal{X}} \frac{1}{p}\,u\,v\,d\mu \tag{100}$$

$$\nabla_u^{(\alpha,\beta)} v = d_u v - \frac{1+\alpha\beta}{2p}\,u\,v \tag{101}$$

$$\nabla_u^{*(\alpha,\beta)} v = d_u v - \frac{1-\alpha\beta}{2p}\,u\,v \tag{102}$$

where $u,v \in \Sigma(\mathcal{M})$ and $p=p(\zeta)$ is the point at which g and ∇ are evaluated.

Proof. The proof is immediate upon substituting Equations (64) and (65) to Equations (83) and (85). ◇

This is to say, with respect to the (α,β)-divergence, the product of the two parameters, $\alpha\beta$, acts as the "alpha" parameter in the family of induced connections, so:

$$\nabla^{*(\alpha,\beta)} = \nabla^{(-\alpha,\beta)} = \nabla^{(\alpha,-\beta)} \tag{103}$$

Setting $\lim_{\beta\to 1}\nabla^{(\alpha,\beta)}$ yields Amari's one-parameter family of α-connections in the infinite-dimensional setting, taking the very simple form:

$$\nabla_u^{(\alpha)} v = d_u v - \frac{1+\alpha}{2p}u\,v \tag{104}$$

The same is true when $\lim_{\alpha\to 1}\nabla^{(\alpha,\beta)}$ (the connections are indexed by β, of course).

3. Parametric Statistical Manifold As Finite-Dimensional Embedding

3.1. Finite-Dimensional Parametric Models

Now, we restrict attention to a finite-dimensional submanifold of measurable functions whose ρ-representation are parameterized using $\theta = [\theta^1, \cdots, \theta^n] \subseteq \mathbb{R}^n$. In this case, the divergence functional of the two functions, p and q, assumed to be specified, respectively, by θ_p and θ_q in the parametric model, becomes an implicit function of θ_p, θ_q. In other words, through introducing parametric models (i.e., a finite-dimensional submanifold) of the infinite-dimensional manifold of measurable functions, we arrive at a divergence function defined ("pulled back") over the vector space. We denote the ρ-representation of a parameterized measurable function as $\rho(p(\zeta|\theta))$, and the corresponding divergence function by $D(\theta_p, \theta_q)$. It is important to realize that, while $f(\cdot)$ is strictly convex, $\mathcal{F}(p) = \int_{\mathcal{X}} f(p(\zeta|\theta))\, d\mu$ is not at all convex in θ in general.

3.1.1. Riemannian Geometry of Parametric Models

The parametric family of functions, $p(\zeta|\theta)$, forms a submanifold of $\mathcal{M}$ defined by:

$$\widetilde{\mathcal{M}}_\theta = \{p(\zeta|\theta) \in \mathcal{M} : \theta \subseteq \mathbb{R}^n\} \tag{105}$$

where $p(\zeta|\theta)$ is a ζ-function indexed by θ, *i.e.*, θ is treated as a parameter to specify a ζ-function. $\widetilde{\mathcal{M}}_\theta$ is a finite-dimensional submanifold of $\mathcal{M}$. We also denote the manifold of a parametric statistical model as:

$$\mathcal{M}_\theta = \{p(\zeta|\theta) \in \mathcal{M}_\mu : \theta \subseteq \Theta \subset \mathbb{R}^n\} \tag{106}$$

The θ values themselves, called the *natural parameter* of the parametric (statistical) model, $p(\cdot|\theta)$, are coordinates for $\widetilde{\mathcal{M}}_\theta$ (or $\mathcal{M}_\theta$). The tangent vector fields, u, v, w, of $\mathcal{M}$ in the directions that are also tangent for $\widetilde{\mathcal{M}}_\theta$ (or $\mathcal{M}_\theta$) take the form:

$$u = \frac{\partial p(\zeta|\theta)}{\partial \theta^i}\ , \qquad v = \frac{\partial p(\zeta|\theta)}{\partial \theta^k}\ , \qquad w = \frac{\partial p(\zeta|\theta)}{\partial \theta^j} \tag{107}$$

The following proposition gives the metric and the family of α-connections in the parametric case. For convenience, we denote $\rho(\zeta, \theta) \equiv \rho(p(\zeta|\theta))$, $\tau(\zeta, \theta) \equiv \tau(p(\zeta|\theta))$ in this subsection.

PROPOSITION 6. For parametric models $p(\zeta|\theta)$, the metric tensor takes the form:

$$g_{ij}(\theta) = \mathrm{E}_\mu \left\{ f''(\rho(\zeta, \theta))\, \frac{\partial \rho(\zeta, \theta)}{\partial \theta^i}\, \frac{\partial \rho(\zeta, \theta)}{\partial \theta^j} \right\} \tag{108}$$

and the α-connections take the form:

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2}\, f'''(\rho(\zeta, \theta))\, A_{ijk} + f''(\rho(\zeta, \theta))\, B_{ijk} \right\} \tag{109}$$

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \frac{1+\alpha}{2}\, f'''(\rho(\zeta, \theta))\, A_{ijk} + f''(\rho(\zeta, \theta))\, B_{ijk} \right\} \tag{110}$$

where:

$$A_{ijk}(\zeta, \theta) = \frac{\partial \rho(\zeta, \theta)}{\partial \theta^i}\, \frac{\partial \rho(\zeta, \theta)}{\partial \theta^j}\, \frac{\partial \rho(\zeta, \theta)}{\partial \theta^k}\ , \qquad B_{ijk}(\zeta, \theta) = \frac{\partial^2 \rho(\zeta, \theta)}{\partial \theta^i \partial \theta^j}\, \frac{\partial \rho(\zeta, \theta)}{\partial \theta^k} \tag{111}$$

Proof. See Section 4.

Note that strict convexity of f requires that $f'' > 0$; thereby the positive-definiteness of $g_{ij}(\theta)$ is guaranteed. Clearly, the α-connections form conjugate pairs $\Gamma^{*(\alpha)}_{ij,k}(\theta) = \Gamma^{(-\alpha)}_{ij,k}(\theta)$.

As an example, we take the embedding $f(t) = e^t$ and $\rho(p) = \log p$, with $\tau(p) = p$, the identity function; then, the expressions in Proposition 6 reduce to the Fisher information and α-connections of the exponential family in Equations (3) and (4).

COROLLARY 7. In dualistic form, the metric and α-connections are:

$$g_{ij}(\theta) = \mathrm{E}_\mu \left\{ \frac{\partial \rho(\zeta,\theta)}{\partial \theta^i} \frac{\partial \tau(\zeta,\theta)}{\partial \theta^j} \right\} \tag{112}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial^2 \tau(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(\zeta,\theta)}{\partial \theta^k} + \frac{1+\alpha}{2} \frac{\partial^2 \rho(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \tau(\zeta,\theta)}{\partial \theta^k} \right\} \tag{113}$$

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \frac{1+\alpha}{2} \frac{\partial^2 \tau(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(\zeta,\theta)}{\partial \theta^k} + \frac{1-\alpha}{2} \frac{\partial^2 \rho(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \tau(\zeta,\theta)}{\partial \theta^k} \right\} \tag{114}$$

Proof. See Section 4.

An immediate consequence of this corollary is as follows. If we construct the divergence function, $\mathcal{D}^{(\alpha)}_{f^*,\tau}(\theta_p, \theta_q)$, then the induced metric, $\tilde{g}_{ij}$, and the induced conjugate connections, $\tilde{\Gamma}^{(\alpha)}_{ij,k}, \tilde{\Gamma}^{*(\alpha)}_{ij,k}$, will be related to those induced from $\mathcal{D}^{(\alpha)}_{f,\rho}(\theta_p, \theta_q)$ (and denoted without the $\tilde{}$) via:

$$\tilde{g}_{ij}(\theta) = g_{ij}(\theta) \tag{115}$$

with:

$$\tilde{\Gamma}^{(\alpha)}_{ij,k}(\theta) = \Gamma^{(-\alpha)}_{ij,k}(\theta) \ , \qquad \tilde{\Gamma}^{*(\alpha)}_{ij,k}(\theta) = \Gamma^{(\alpha)}_{ij,k}(\theta) \tag{116}$$

So, the difference between using $\mathcal{D}^{(\alpha)}_{f,\rho}(\theta_p, \theta_q)$ and $\mathcal{D}^{(\alpha)}_{f^*,\tau}(\theta_p, \theta_q)$ reflects a conjugacy in the ρ- and τ-scalings of $p(\zeta|\theta)$. Corollary 7 says that the conjugacy in the connection pair $\Gamma \longleftrightarrow \Gamma^*$ reflects, in addition to the referential duality $\theta_p \longleftrightarrow \theta_q$, the representational duality between ρ-scaling and τ-scaling of a ζ-function:

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = \tilde{\Gamma}^{(\alpha)}_{ij,k}(\theta) \tag{117}$$

3.1.2. Example: The Parametric (α, β)-Manifold

We have introduced the two-parameter family of divergence *functionals* $\mathcal{D}^{(\alpha,\beta)}(p, q)$ in Section 2.4. Now, pulling back to $\widetilde{\mathcal{M}}_\theta$ (or to $\mathcal{M}_\theta$), we have the two-parameter family of divergence *functions* $D^{(\alpha,\beta)}(\theta_p, \theta_q)$ defined by:

$$D^{(\alpha,\beta)}(\theta_p, \theta_q) = \mathcal{D}^{(\alpha,\beta)}_{f,\rho}(p(\cdot|\theta_p), q(\cdot|\theta_q)) \tag{118}$$

There are two ways to reduce to Amari's alpha-divergence (indexed by β here to avoid confusion): (i) take $\alpha = 1$ and $\rho(p) = l^{(\beta)}(p) \longleftrightarrow \tau(p) = l^{(-\beta)}(p)$; or (ii) take $\alpha = -1$ and $\rho(p) = l^{(-\beta)}(p) \longleftrightarrow \tau(p) = l^{(\beta)}(p)$.

COROLLARY 8. The metric and affine connections for the parametric (α, β)-manifold are:

$$g_{ij}(\theta) = \mathrm{E}_p \left\{ \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right\} \tag{119}$$

$$\Gamma^{(\alpha,\beta)}_{ij,k}(\theta) = \mathrm{E}_p \left\{ \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} \frac{\partial \log p}{\partial \theta^k} + \frac{1-\alpha\beta}{2} \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \frac{\partial \log p}{\partial \theta^k} \right\} \tag{120}$$

$$\Gamma^{*(\alpha,\beta)}_{ij,k}(\theta) = \mathrm{E}_p \left\{ \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} \frac{\partial \log p}{\partial \theta^k} + \frac{1+\alpha\beta}{2} \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \frac{\partial \log p}{\partial \theta^k} \right\} \tag{121}$$

Proof. Direct substituion of expressions of $\rho(p)$ and $\tau(p)$. $\diamond$

This two-parameter family of affine connections, $\Gamma^{(\alpha,\beta)}_{ij,k}(\theta)$, indexed now by the numerical product, $\alpha\beta \in [-1, 1]$, is actually the alpha-connection proper (*i.e.*, the one-parameter family of its generic form; see [29])

$$\Gamma^{(\alpha,\beta)}_{ij,k}(\theta) = \Gamma^{(-\alpha,-\beta)}_{ij,k}(\theta) \tag{122}$$

with biduality compactly expressed as

$$\Gamma^{*(\alpha,\beta)}_{ij,k}(\theta) = \Gamma^{(-\alpha,\beta)}_{ij,k}(\theta) = \Gamma^{(\alpha,-\beta)}_{ij,k}(\theta) \tag{123}$$

3.2. Affine Embedded Submanifold

We now define the notion of ρ-affinity. A parametric model, $p(\zeta|\theta)$, is said to be *ρ-affine* if its ρ-representation can be embedded into a finite-dimensional affine space, *i.e.*, if there exists a set of linearly independent functions $\lambda_i(\zeta)$ over the same support, $\mathcal{X} \ni \zeta$, such that:

$$\rho(p(\zeta|\theta)) = \sum_i \theta^i \lambda_i(\zeta) \tag{124}$$

As noted in Section 3.1.1, the parameter $\theta = [\theta^1, \cdots, \theta^n] \in \Theta$ is its natural parameter.

For any measurable function, $p(\zeta)$, the projection of its τ-representation onto the functions $\lambda_i(\zeta)$

$$\eta_i = \int_{\mathcal{X}} \tau(p(\zeta))\, \lambda_i(\zeta)\, d\mu \tag{125}$$

forms a vector $\eta = [\eta_1, \cdots, \eta_n] \subseteq \mathbb{R}^n$. We call η the expectation parameter of $p(\zeta)$, and the functions $\lambda(\zeta) = [\lambda_1(\zeta), \cdots, \lambda_n(\zeta)]$ the affine basis functions.

The above notion of ρ-affinity is a generalization of α-affine manifolds [1,2], where ρ- and τ-representations are just α- and $(-\alpha)$-representations, respectively. Note that elements of the ρ-affine manifold may not be a probability model; rather, after denormalization, probability models can become ρ-affine. The issue of normalization will be discussed in Section 5.

3.2.1. Biorthogonality of Natural and Expectation Parameters

PROPOSITION 9. When a parametric model is ρ-affine,

(i) the function:

$$\Phi(\theta) = \int_{\mathcal{X}} f(\rho(p(\zeta|\theta)))\, d\mu \tag{126}$$

is strictly convex;

(ii) the divergence functional, $\mathcal{D}^{(\alpha)}_{f,\rho}(p, q)$, takes the form of the divergence function:

$$D^{(\alpha)}_{\Phi}(\theta_p, \theta_q) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}\,\Phi(\theta_p) + \frac{1+\alpha}{2}\,\Phi(\theta_q) - \Phi\left(\frac{1-\alpha}{2}\,\theta_p + \frac{1+\alpha}{2}\theta_q\right)\right) \tag{127}$$

(iii) the metric tensor, affine connections and the Riemann curvature tensor take the forms:

$$g_{ij}(\theta) \;=\; \Phi_{ij}\;; \quad \Gamma^{(\alpha)}_{ij,k}(\theta) \;=\; \frac{1-\alpha}{2}\,\Phi_{ijk} \;=\; \Gamma^{*(-\alpha)}_{ij,k}(\theta) \tag{128}$$

$$R^{(\alpha)}_{ij\mu\nu}(\theta) \;=\; \frac{1-\alpha^2}{4}\sum_{l,k}(\Phi_{il\nu}\Phi_{jk\mu} - \Phi_{il\mu}\Phi_{jk\nu})\Phi^{lk} \;=\; R^{*(\alpha)}_{ij\mu\nu}(\theta) \tag{129}$$

Here, Φ_{ij}, Φ_{ijk} denote, respectively, second and third partial derivatives of $\Phi(\theta)$:

$$\Phi_{ij} = \frac{\partial^2 \Phi(\theta)}{\partial \theta^i \partial \theta^j} \ , \qquad \Phi_{ijk} = \frac{\partial^3 \Phi(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k} \tag{130}$$

and Φ^{ij} is the matrix inverse of Φ_{ij}.

Proof. See Section 4.

Recall that, for a convex function of several variables, $\Phi : \mathbb{R}^n \to \mathbb{R}$, its convex conjugate Φ^* is defined through the Legendre–Fenchel transform:

$$\Phi^*(\eta) = \langle \eta, (\partial \Phi)^{-1}(\eta) \rangle - \Phi((\partial \Phi)^{-1}(\eta)) \tag{131}$$

where $\partial \Phi$ stands for the gradient (sub-differential) of Φ, and $\langle \ , \ \rangle$ denotes the standard inner product. It can be shown that the function, Φ^*, is also convex and has Φ as its conjugate $(\Phi^*)^* = \Phi$. The Hessian (second derivatives) of a strictly convex function (Φ and Φ^*) is positive-definite. The Legendre–Fenchel inequality Equation (131) can be expressed using dual variables, θ, η, as:

$$\Phi(\theta) + \Phi^*(\eta) - \sum_i \eta_i \theta^i \geq 0 \tag{132}$$

where equality holds, if and only if:

$$\theta = (\partial \Phi^*)(\eta) = (\partial \Phi)^{-1}(\eta) \longleftrightarrow \eta = \partial \Phi(\theta) = (\partial \Phi^*)^{-1}(\theta) \tag{133}$$

COROLLARY 10. For a ρ-affine manifold:

(i) define

$$\tilde{\Phi}(\theta) = \int_{\mathcal{X}} f^*(\tau(p(\zeta|\theta))) \, d\mu \tag{134}$$

then $\Phi^*(\eta) \equiv \tilde{\Phi}((\partial \Phi)^{-1}(\eta))$ is the convex (Legendre–Fenchel) conjugate of $\Phi(\theta)$;

(ii) the pair of convex functions, Φ, Φ^*, form a pair of "potentials" to induce η, θ:

$$\frac{\partial \Phi(\theta)}{\partial \theta^i} = \eta_i \longleftrightarrow \frac{\partial \Phi^*(\eta)}{\partial \eta_i} = \theta^i \tag{135}$$

(iii) the expectation parameter, $\eta \in \Xi$, and the natural parameter, $\theta \in \Theta$, form biorthogonal coordinates:

$$\frac{\partial \eta_i}{\partial \theta^j} = g_{ij}(\theta) \longleftrightarrow \frac{\partial \theta^i}{\partial \eta_j} = \tilde{g}^{ij}(\eta) \tag{136}$$

where $\tilde{g}^{ij}(\eta)$ is the matrix inverse of $g_{ij}(\theta)$, the metric tensor of the parametric (statistical) manifold.

Proof. See Section 4.

Note that while the function, $\Phi(\theta)$, can be viewed as the generalized cumulant generating function (or partition function), the function, $\Phi^*(\eta)$, is the generalized entropy function. For an exponential family, the two are well known to form one-to-one correspondence; either can be used on that index as a density function of the exponential family.

3.2.2. Dually Flat Affine Manifolds

When $\alpha = \pm 1$, part (iii) of Proposition 9 dictates that all components of the curvature tensor vanish, *i.e.*, $R^{(\pm 1)}_{ij\mu\nu}(\theta) = 0$. In this case, there exists a coordinate system under which either $\Gamma^{*(-1)}_{ij,k}(\theta) = 0$ or $\Gamma^{(1)}_{ij,k}(\theta) = 0$. This is the well-studied "dually flat" parametric statistical manifold [1,2,20], under which divergence functions have a unique, canonical form.

COROLLARY 11. When $\alpha \to \pm 1$, $D^{(\alpha)}_\Phi$ reduces to the Bregman divergence Equation (17):

$$D^{(-1)}_\Phi(\theta_p, \theta_q) = D^{(1)}_\Phi(\theta_q, \theta_p) = \Phi(\theta_q) - \Phi(\theta_p) - \langle \theta_q - \theta_p, \partial\Phi(\theta_p) \rangle = B_\Phi(\theta_q, \theta_p) \quad (137)$$

$$D^{(1)}_\Phi(\theta_p, \theta_q) = D^{(-1)}_\Phi(\theta_q, \theta_p) = \Phi(\theta_p) - \Phi(\theta_q) - \langle \theta_p - \theta_q, \partial\Phi(\theta_q) \rangle = B_\Phi(\theta_p, \theta_q) \quad (138)$$

or equivalently, to the canonical divergence functions:

$$D^{(1)}_\Phi(\theta_p, (\partial\Phi)^{-1}(\eta_q)) = \Phi(\theta_p) + \Phi^*(\eta_q) - \langle \theta_p, \eta_q \rangle \equiv A_\Phi(\theta_p, \eta_q) \quad (139)$$

$$D^{(-1)}_\Phi((\partial\Phi)^{-1}(\theta_p), \theta_q) = \Phi(\theta_q) + \Phi^*(\eta_p) - \langle \eta_p, \theta_q \rangle \equiv A_{\Phi^*}(\eta_p, \theta_q) \quad (140)$$

Proof. Immediate by substitution using the definition Equation (131). ⋄

The canonical divergence, $A_\Phi(\theta_p, \eta_q)$, based on the Legendre–Fenchel inequality was introduced by [2,20], where the functions, Φ, Φ^*, the cumulant generating functions of an exponential family, were referred to as dual "potential" functions. This form Equation (139) is "canonical", because it is uniquely specified in a dually flat manifold using a pair of biorthogonal coordinates.

We point out that there are *two* kinds of duality associated with the divergence defined on dually flat statistical manifold, one between $D^{(-1)}_\Phi \longleftrightarrow D^{(1)}_\Phi$ and between $D^{(-1)}_{\Phi^*} \longleftrightarrow D^{(1)}_{\Phi^*}$; the other between $D^{(-1)}_\Phi \longleftrightarrow D^{(-1)}_{\Phi^*}$ and between $D^{(1)}_\Phi \longleftrightarrow D^{(1)}_{\Phi^*}$. The first kind is related to the duality in the choice of the reference and the comparison status for the two points (θ *versus* η) for computing the value of the divergence and, hence, called "referential duality". The second kind is related to the duality in the choice of the representation of the point as a vector in the parameter *versus* gradient space (θ *versus* η) in the expression of the divergence function and, hence, called "representational duality". More concretely:

$$D^{(-1)}_\Phi(\theta_p, \theta_q) = D^{(-1)}_{\Phi^*}(\partial\Phi(\theta_q), \partial\Phi(\theta_p)) = D^{(1)}_{\Phi^*}(\partial\Phi(\theta_p), \partial\Phi(\theta_q)) = D^{(1)}_\Phi(\theta_q, \theta_p) \quad (141)$$

The biduality is compactly reflected in the canonical divergence as:

$$A_\Phi(\theta_p, \eta_q) = A_{\Phi^*}(\eta_q, \theta_p) \quad (142)$$

4. Proofs

PROOF OF *Eguchi Relation* EQUATIONS (26)–(28). Assume the divergence function, $\mathcal{D}$, is Fréchet differentiable up to third order. For any two points, $p, q \in \mathcal{M}$, and two vector fields, $u, v \in \Sigma(\mathcal{M})$, let us denote $G(p, q)(u, v) = -(d_u)_p\,(d_v)_q\,\mathcal{D}(p, q)$, the mixed second derivative, which is bilinear in u, v. Then, $\langle u, v \rangle_p = G(p, p)(u, v)$. Suppressing the dependency on (u, v) in G, we take directional derivative with respect to $w \in \Sigma(\mathcal{M})$:

$$d_w\, G(p, q) = (d_w)_p\, G(p, q) + (d_w)_q\, G(p, q) \quad (143)$$

and then evaluating at $q = p$ to obtain:

$$d_w \langle u, v \rangle = \langle \nabla_w u, v \rangle + \langle u, \nabla^*_w v \rangle \tag{144}$$

which is the defining Equation (22) for conjugate connections. Therefore, what remains to be checked is whether ∇ as defined by:

$$\langle \nabla_w u, v \rangle = -(d_w)_p \, (d_u)_p \, \widetilde{G}(p, q)(v) \Big|_{p=q} \tag{145}$$

indeed transforms as an affine connection (and similarly for ∇^*), where $\widetilde{G}(p, q)(v) = -(d_v)_q \, \mathcal{D}(p, q)$ is linear in v. It is easy to verify that $\nabla_w(u_1 + u_2) = \nabla_w u_1 + \nabla_w u_2$, $\nabla_{w_1+w_2} u = \nabla_{w_1} u + \nabla_{w_2} u$, and $\nabla_{fw} u = f \nabla_w u$ for $f \in \mathcal{F}$. We need only to prove:

$$\langle \nabla_w (fu), v \rangle = f \langle \nabla_w u, v \rangle + \langle u, v \rangle \, d_w f \tag{146}$$

which is immediately obtained by:

$$(d_w)_p \, (f d_u)_p \, \widetilde{G}(p, q) = f \cdot \Big((d_w)_p \, (d_u)_p \, \widetilde{G}(p, q) \Big) + (d_u)_p \, (\widetilde{G}(p, q)) \, (d_w f)_p \tag{147}$$

$\diamond$

Proof of Proposition 1. We only need to prove that for a strictly convex function, $f : \mathbb{R} \to \mathbb{R}$ and $\alpha \in \mathbb{R}$, the following quantity:

$$d_f^{(\alpha)}(\gamma, \delta) = \frac{4}{1 - \alpha^2} \left(\frac{1 - \alpha}{2} f(\gamma) + \frac{1 + \alpha}{2} f(\delta) - f\left(\frac{1 - \alpha}{2} \, \gamma + \frac{1 + \alpha}{2} \, \delta \right) \right) \tag{148}$$

is non-negative for all real numbers, $\gamma, \delta \in \mathbb{R}$, with $d_f^{(\alpha)}(\gamma, \delta) = 0$, if and only if $\gamma = \delta$.

Clearly, for any $\alpha \in (-1, 1)$, $1 - \alpha^2 > 0$; so, from the fundamental convex inequality Equation (47), the functions $d_f^{(\alpha)}(\gamma, \delta) \geq 0$ for all $\gamma, \delta \in \mathbb{R}$, with equality holding, if and only if $\gamma = \delta$. When $\alpha > 1$, we rewrite $\delta = \frac{2}{\alpha+1} \, \lambda + \frac{\alpha-1}{\alpha+1} \, \gamma$ as a convex mixture of λ and γ (*i.e.*, $\frac{2}{\alpha+1} = \frac{1-\alpha'}{2}$, $\frac{\alpha-1}{\alpha+1} = \frac{1+\alpha'}{2}$ with $\alpha' \in (-1, 1)$). Strict convexity of f guarantees:

$$\frac{2}{\alpha + 1} \, f(\lambda) + \frac{\alpha - 1}{\alpha + 1} \, f(\gamma) \geq f(\delta) \tag{149}$$

or moving left-hand side to right-hand side:

$$\frac{2}{1 + \alpha} \left(\frac{1 - \alpha}{2} \, f(\gamma) + \frac{1 + \alpha}{2} \, f(\delta) - f\left(\frac{1 - \alpha}{2} \, \gamma + \frac{1 + \alpha}{2} \, \delta \right) \right) \leq 0 \tag{150}$$

This, along with $1 - \alpha^2 < 0$ proves the non-negativity of $d_f^{(\alpha)}(\gamma, \delta) \geq 0$ for $\alpha > 1$, with equality holding, if and only if $\lambda = \gamma$, *i.e.*, $\gamma = \delta$. The case of $\alpha < -1$ is similarly proven by applying Equation (47) to the three points, γ, λ, and their convex mixture $\delta = \frac{2}{1-\alpha} \, \lambda + \frac{-1-\alpha}{1-\alpha} \, \gamma$. Finally, the continuity of $d_f^{(\alpha)}(\gamma, \delta)$ with respect to α guarantees that the above claim is also valid in the case of $\alpha = \pm 1$. $\diamond$

Proof of Proposition 2. With respect to Equation (54), note that $(d_u)_p$ means that the functional derivative is with respect to p only (point q is treated as fixed):

$$(d_u)_p \mathcal{D}_{f,\rho}^{(\alpha)}(p, q) = \frac{2}{1 + \alpha} \int_{\mathcal{X}} \left\{ f'(\rho(p)) - f'\left(\frac{1 - \alpha}{2} \rho(p) + \frac{1 + \alpha}{2} \rho(q) \right) \right\} \rho'(p) \, u \, d\mu \tag{151}$$

Applying functional derivative $(d_v)_q$, now with respect to q only, to the above equation yields:

$$(d_v)_q \left((d_u)_p \mathcal{D}^{(\alpha)}_{f,\rho}(p,q) \right) = - \int_{\mathcal{X}} f'' \left(\frac{1-\alpha}{2}\rho(p) + \frac{1+\alpha}{2}\rho(q) \right) \rho'(p)\rho'(q)\, u\, v\, d\mu \tag{152}$$

Setting $p = q$ and invoking Equation (26) yields Equation (82) with Equation (83).

Next, applying $(d_w)_p$ to Equation (152), and realizing that u, v are both vector fields:

$$(d_w)_p((d_v)_q\,(d_u)_p \mathcal{D}^{(\alpha)}_{f,\rho}(p,q)) = \tag{153}$$

$$- \int_{\mathcal{X}} \frac{1-\alpha}{2} \left(f''' \left(\frac{1-\alpha}{2}\rho(p) + \frac{1+\alpha}{2}\rho(q) \right) (\rho'(p))^2 \rho'(q)\, u\, v\, w\, d\mu \tag{154}$$

$$- \int_{\mathcal{X}} f'' \left(\frac{1-\alpha}{2}\rho(p) + \frac{1+\alpha}{2}\rho(q) \right) \rho'(q)\, v\, \left(\rho''(p)\, u\, w + \rho'(p)\, (d_w u) \right) d\mu \tag{155}$$

Setting $p = q$, invoking Equation (27) and:

$$g(\nabla_w u, v) = \int f''(\rho(p))\, (\rho'(p))^2\, \nabla_w u\, v(\zeta|p) d\mu \tag{156}$$

and realizing that $v(\zeta|p)$ can be arbitrary, we have:

$$f''(\rho)\, (\rho')^2\, \nabla^{(\alpha)}_w u = \frac{1-\alpha}{2}\, f'''(\rho)\, (\rho')^3\, u\, w + f''(\rho)\, \rho'\, (\rho'' u\, w + \rho'\, (d_w u)) \tag{157}$$

where we have short-handed ρ for $\rho(p(\zeta))$. Remember that $\nabla^{(\alpha)}_w u$ is a ζ-function; the above equation yields:

$$\nabla^{(\alpha)}_w u = d_w u + \frac{1-\alpha}{2}\, \frac{f'''(\rho)}{f''(\rho)}\, \rho' u\, w + \frac{\rho''}{\rho'} u\, w = d_w u + \left(\frac{1-\alpha}{2}\, \frac{f'''(\rho)}{f''(\rho)} \rho' + \frac{\rho''}{\rho'} \right) uw \tag{158}$$

Thus, we obtain Equation (84) with Equation (85). The expression for $\nabla^{*(\alpha)}$ is obtained analogously. $\diamond$

PROOF OF COROLLARY 3. From the identities:

$$f''(\rho) = \frac{\tau'}{\rho'}\ , \quad f'''(\rho) = \frac{\rho'\tau'' - \rho''\tau'}{(\rho')^3} \tag{159}$$

we obtain Equations (87) and (88) after substitution. $\diamond$

PROOF OF PROPOSITION 4. We first derive a general formula for the Riemann curvature tensor for the infinite-dimensional manifold, since that given by a popular text book ([46], p.226) appears to miss some terms. From Equation (42):

$$d_u(\nabla_v w) = d_u(d_v w) + \mathrm{B}(d_u v, w) + \mathrm{B}(v, d_u w) + (d_u \mathrm{B})(v, w) \tag{160}$$

so that:

$$\begin{aligned} \nabla_u(\nabla_v w) &= d_u(\nabla_v w) + \mathrm{B}(u, \nabla_v w) \\ &= (d_u(d_v w) + \mathrm{B}(d_u v, w) + \mathrm{B}(v, d_u w) + (d_u \mathrm{B})(v, w)) + (\mathrm{B}(u, d_v w) + \mathrm{B}(u, \mathrm{B}(v, w))) \end{aligned}$$

here $d_u \mathrm{B} = \mathrm{B}' u$ refers to the derivative on the B-form itself and not on its v, w arguments. The expression for $\nabla_v(\nabla_u w)$ simply exchanges $u \to v$ in the above. Now:

$$\nabla_{[u,v]} w = d_{[u,v]} w + \mathrm{B}([u, v], w) \tag{161}$$

where $[u, v] = d_u v - d_v u$ is a vector field, such that:

$$d_{[u,v]} w = d_u(d_v w) - d_v(d_u w) \tag{162}$$

Substituting them into Equation (44), we get a general expression of the Riemann curvature tensor in infinite-dimensional setting:

$$R(u, v, w) = \mathrm{B}(u, \mathrm{B}(v, w)) - \mathrm{B}(v, \mathrm{B}(u, w)) + (d_u \mathrm{B})(v, w) - (d_v \mathrm{B})(u, w) \tag{163}$$

The expression for $T(u, v)$ in Equation (46) becomes:

$$T(u, v) = \mathrm{B}(u, v) - \mathrm{B}(v, u) \tag{164}$$

In the current case, B evaluated at $p(\zeta)$ is the bilinear form:

$$\mathrm{B}(u, v) = b^{(\alpha)}(p(\zeta)) u(\zeta|p) v(\zeta|p) \tag{165}$$

Substituting this into the above, and realizing that $(d_u \mathrm{B})(v, w)$ is simply $(b^{(\alpha)})\prime\, u\, v\, w$, we immediately have $R^{(\alpha)}(u, v, w) = 0$, as well as $T^{(\alpha)}(u, v) = 0$. ⋄

Proof of Proposition 6. Given Equation (107) as the tangent vector fields for parametric models with holonomic coordinates θ, we note that:

$$d_u \rho \;=\; \rho'(p)\, u \;=\; \rho'(p)\, \frac{\partial p}{\partial \theta^i} \;=\; \frac{\partial \rho(p)}{\partial \theta^i} \tag{166}$$

$$d_w \rho \;=\; \rho'(p)\, w \;=\; \rho'(p)\, \frac{\partial p}{\partial \theta^j} \;=\; \frac{\partial \rho(p)}{\partial \theta^j} \tag{167}$$

so Equation (108) follows. Next, from:

$$d_w u = \frac{\partial^2 p}{\partial \theta^i \partial \theta^j} \tag{168}$$

we have:

$$\begin{aligned} \rho''(p)\, u\, w + \rho'(p)\, (d_w u) \;&=\; \rho''(p) \frac{\partial p}{\partial \theta^i} \frac{\partial p}{\partial \theta^j} + \rho'(p) \frac{\partial^2 p}{\partial \theta^i \partial \theta^j} \;=\; \frac{\partial}{\partial \theta^i} \left(\rho'(p) \frac{\partial p}{\partial \theta^j} \right) \\ &=\; \frac{\partial}{\partial \theta^i} \left(\frac{\partial \rho}{\partial \theta^j} \right) \;=\; \frac{\partial^2 \rho}{\partial \theta^i \partial \theta^j} \end{aligned} \tag{169}$$

Observing $\Gamma_{ij,k} = \langle \nabla_w u, v \rangle$, expression (109) results after substituting the above derived expressions into Equation (84) with Equation (85). ⋄

Proof of Corollary 7. Applying Equations (166) and (167) to Equation (82) with Equation (87) immediately yields Equation (112). Next, from Corollary 3:

$$b^{(\alpha)}(t) = \frac{1 - \alpha}{2} \frac{\tau''(t)}{\tau'(t)} + \frac{1 + \alpha}{2} \frac{\rho''(t)}{\rho'(t)} \tag{170}$$

It follows that:

$$\Gamma^{(\alpha)}_{ij,k} = \langle \nabla^{(\alpha)}_w u, v \rangle = \left(\frac{1 - \alpha}{2} \left(\rho'\, \tau'' u\, w + \rho'\, \tau'\, d_w u \right) + \frac{1 + \alpha}{2} \left(\rho''\, \tau' u\, w + \rho'\, \tau'\, d_w u \right) \right) v \tag{171}$$

$$= \rho'\, v\, \frac{1 - \alpha}{2} d_w (d_u \tau) + \tau'\, v\, \frac{1 + \alpha}{2} d_w (d_u \rho) = \frac{1 - \alpha}{2} (d_v \rho)\, d_w (d_u \tau) + \frac{1 + \alpha}{2} (d_v \tau)\, d_w (d_u \rho) \tag{172}$$

Note that given holonomic coordinates Equation (107):

$$d_w(d_u\rho) = \frac{\partial}{\partial\theta^j}\left(\frac{\partial\rho(p)}{\partial\theta^i}\right) = \frac{\partial^2\rho(p)}{\partial\theta^i\partial\theta^j} \tag{173}$$

Substituting into Equation (84) with Equation (88) yields Equations (113) and (114). ⋄

PROOF OF PROPOSITION 9. The assumption Equation (124) implies that $\frac{\partial\rho}{\partial\theta^i} = \lambda_i(\zeta)$, so from Equation (108):

$$\frac{\partial^2\Phi(\theta)}{\partial\theta^i\partial\theta^j} = \int_{\mathcal{X}} f''(\rho(p(\zeta|\theta))\,\lambda_i(\zeta)\,\lambda_j(\zeta)\,d\mu \tag{174}$$

That the above expression is positive definite is seen by observing:

$$\sum_{ij}\frac{\partial^2\Phi(\theta)}{\partial\theta^i\partial\theta^j}\xi^i\xi^j = \int_{\mathcal{X}} f''(\rho(p(\zeta|\theta))\left(\sum_i \lambda_i(\zeta)\xi^i\right)^2 d\mu > 0 \tag{175}$$

for any $\xi = [\xi^1, \cdots, \xi^n] \in \mathbb{R}^n$, due to the linear independence of the λ_i components and the strict convexity of f. Hence, $\Phi(\theta)$ is strictly convex in θ, proving (i). An immediate consequence is that expression (127) is non-negative and vanishes, if an only if $\theta_p = \theta_q$. This establishes (ii), *i.e.*, $D_\Phi^{(\alpha)}(\theta_p, \theta_q)$ is a divergence functions. Part (iii) follows from a straight-forward application of Eguchi relations Equations (29)–(31). ⋄

PROOF OF COROLLARY 10. First, since $f'(\rho(t)) = \tau(t)$, we have the identity:

$$f^*(\tau(p(\zeta|\theta)) + f(\rho(p(\zeta|\theta))) = f'(\rho(p(\zeta|\theta)))\,\rho(p(\zeta|\theta)) \tag{176}$$

From (126), taking a derivative with respect to θ^i, while noting that $p(\zeta|\theta)$ satisfies (124), gives:

$$\frac{\partial\Phi(\theta)}{\partial\theta^i} = \int_{\mathcal{X}} f'\left(\sum_j \theta^j\lambda_j(\zeta)\right)\lambda_i(\zeta)\,d\mu = \int_{\mathcal{X}} \tau(p(\zeta|\theta))\lambda_i(\zeta)\,d\mu = \eta_i \tag{177}$$

and that:

$$\sum_i \theta^i\frac{\partial\Phi(\theta)}{\partial\theta^i} - \Phi(\theta) = \int_{\mathcal{X}}\left\{f'\left(\sum_j\theta^j\lambda_j(\zeta)\right)\left(\sum_i\theta^i\lambda_i(\zeta)\right) - f\left(\sum_j\theta^j\lambda_j(\zeta)\right)\right\}d\mu \tag{178}$$

$$= \int_{\mathcal{X}} f^*(\tau(p(\zeta|\theta)))d\mu = \tilde{\Phi}(\theta) \tag{179}$$

It follows from Equation (131) that Φ^*, as defined in (i), is the conjugate of Φ, and that the relation in (ii) is the basic Legendre–Fenchel duality. Finally, the biorthogonality of η and θ as expressed by (iii) also becomes evident on account of (ii). ⋄

5. Discussions

This paper constructs a family of divergence functionals, induced by any smooth and strictly convex function, to measure the non-symmetric "distance" between two measurable functions defined on a sample space. Subject to an arbitrary monotone scaling, the divergence functional induces a Riemannian manifold with a metric tensor generalizing the conventional Fisher information and a pair

of conjugate connections generalizing the conventional $(\pm\alpha)$-connections. Such manifolds manifest biduality: referential duality (in choosing a reference point) and representational duality (in choosing a monotone scale). The (α, β)-divergence we gave as an example of this bidualistic structure extends the α-divergence, with α and β representing referential duality and representational duality, respectively. It induces the conventional Fisher metric and the conventional α-connection (with $\alpha\beta$ as a single parameter). Finally, for the ρ-affine submanifold, a pair of conjugated potentials exist to induce the natural and expectation parameters as biorthogonal coordinates on the manifold.

Our approach demonstrated an intimate connection between convex analysis and information geometry. The divergence functionals (and the divergence functions in the finite-dimensional case) are associated with the fundamental convex inequality of a convex function, $f : \mathbb{R} \to \mathbb{R}$ (or $\Phi : \mathbb{R}^n \to \mathbb{R}$), with the convex mixture coefficient as the α-parameter in the induced geometry. Referential duality is associated with $\alpha \longleftrightarrow -\alpha$, and representational duality is associated with the convex conjugacy $f \longleftrightarrow f^*$ (or $\Phi \longleftrightarrow \Phi^*$). Thus, our analysis reveals that the e/m-duality and (± 1)-duality that were used almost interchangeably in the current literature are not the same thing!

The kind of referential duality (originating from non-symmetric status for a referent and for a comparison object), while common in psychological and behavioral contexts [54,55], has always been implicitly acknowledged in statistics. Formal investigation of such non-symmetry between a reference probability distribution and comparison probability distribution in constructing divergence functions leads to the framework of preferred point geometry [56–61]. Preferred point geometry reformulates Amari's [20] expected geometry and Barndorff-Nielsen's [3] observed geometry by studying the product manifold, $\mathcal{M}_\theta \times \mathcal{M}_\theta$, formed by an ordered pair of probability densities, (p, q), and defining a family of Riemannian metric defined on the product manifold. The precise relation of the preferred point approach with our approach to referential duality needs future exploration.

With respect to representational duality, it is worth mentioning the field of affine differential geometry which studies hypersurface realization of the dual Riemannian manifold involving a pair of conjugate connections (see [27,28]). [31–34] investigated affine immersion of statistical manifolds. [62–65] further illuminated a conformal structure when the (normalized) probability density functions undergo the $l^{(\alpha)}$ embedding. Such an embedding appears in the context of Tsallis statistics, where Shannon entropy and Kullback-Leibler cross-entropy (divergence) is generalized to a one-parameter family of entropy and cross-entropy (see, e.g., [40]). We demonstrated ([48], and here) that the ρ-affine manifold (Section 3.2) has the structure of an α-Hessian structure [26], a generalization of Hessian manifold [66,67]. It remains to be illuminated whether a conformal structure arises for ρ-affine probability density functions after normalization.

It should be noted that, while any divergence function determines uniquely a statistical manifold (in the broad sense of [29]), the converse is not true. Though a statistical manifold equipped with an arbitrary metric tensor and a pair of conjugate, torsion-free connections always admits a divergence function [68], it is not unique in general, except when the connections are dually flat, in which case the divergence is uniquely determined as the canonical divergence. In this sense, there is nothing special about our use of $\mathcal{D}^{(\alpha)}$-divergence apart from it generalizing familiar divergences (including α-divergence in particular). Rather, $\mathcal{D}^{(\alpha)}$-divergence is merely a vehicle for us to derive the underlying dual Riemannian geometry. It remains to be elucidated *why* the convex mixture parameter turns out to be the α-parameter

in the family of connections of the induced geometry. It seems that our generalizations of the Fisher metric and of conjugate α-connections hinge on this miraculous identification. Generalization from α-affinity/embedding to ρ-affinity/embedding, and the resulting generalized biorthogonality between natural and expectation parameters is akin to generalizing L_p space to L_Φ (*i.e.*, Orlicz) space, which is an entirely different matter. Future research will further clarify these fundamental relations between convexity, conjugacy, and duality in non-parametric (and parametric) information geometry.

6. Conclusions

We constructed an extension of parametric information geometry to the non-parametric setting by studying the manifold $\mathcal{M}$ of non-parametric functions on sample space (without positivity and normalization constraints). The generalized Fisher information and α-connections on $\mathcal{M}$ are induced by an α-parameterized family of divergence functions, reflecting the fundamental convex inequality associated with any smooth and strictly convex function. Parametric models are recovered as submanifolds of $\mathcal{M}$. We also generalize Amari's α-embedding to an affine submanifold under arbitrary monotonic embedding, and show that its natural and expectation parameters form biorthogonal coordinates, and such a submanifold is dually flat for $\alpha = \pm 1$. Our analysis illuminates two different types of duality in information geometry, one concerning the referential status of a point (measurable function) expressed in the divergence function ("referential duality") and the other concerning its representation under an arbitrary monotone scaling ("representational duality").

Acknowledgments

This paper is based on work presented to the Second International Conference of Information Geometry and Its Applications (IGAIA2), Tokyo, Japan in 2005 and appeared in preliminary form as [47]. The author appreciates two anonymous reviewers for constructive feedback. The author has been supported by research grants NSF 0631541 and ARO W911NF-12-1-0163 during revision and final production of this work.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Amari, S.; Nagaoka, H. *Method of Information Geometry*; Oxford University Press: Oxford, UK, 2000.
2. Amari, S. *Differential Geometric Methods in Statistics*; Springer-Verlag: New York, NY, USA, 1985.
3. Barndorff-Nielsen, O.E. *Parametric Statistical Models and Likelihood*; Springer-Verlag: Heidelberg, Germany, 1988.
4. Barndorff-Nielsen, O.E.; Cox, R.D.; Reid, N. The role of differential geometry in statistical theory. *Int. Stat. Rev.* **1986**, *54*, 83–96.

5. Kass, R.E. The geometry of asymptotic inference (with discussion). *Stat. Sci.* **1989**, *4*, 188–234.
6. Kass, R.E.; Vos, P.W. *Geometric Foundation of Asymptotic Inference*; John Wiley and Sons: New York, NY, USA, 1997.
7. Murray, M.K.; Rice, J.W. *Differential Geometry and Statistics*; Chapman & Hall: London, UK, 1993.
8. Amari, S.; Kumon, M. Estimation in the presence of infinitely many nuisance parameters — Geometry of estimating functions. *Ann. Stat.* **1988**, *16*, 1044–1068.
9. Henmi, M.; Matsuzoe, H. Geometry of pre-contrast functions and non-conservative estimating functions. In Proceedings of the International Workshop on Complex Structures, Integrability, and Vector Fields, Sofia, Bulgaria, 13–17 September 2010; Volume 1340, pp. 32–41.
10. Matsuzoe, H.; Takeuchi, J.; Amari. S. Equiaffine structures on statistical manifolds and Bayesian statistics. *Differ. Geom. Its Appl.* **2006**, *109*, 567–578.
11. Takeuchi, J.; Amari, S. α-Parallel prior and its properties. *IEEE Trans. Inf. Theory* **2005**, *51*, 1011–1023.
12. Amari, S. Natural gradient works efficiently in learning. *Neural Comput.* **1988**, *10*, 251–276.
13. Yang, H.H.; Amari, S. Complexity issues in natural gradient descent method for training multilayer perceptrons. *Neural Comput.* **1998**, *10*, 2137–2157.
14. Amari, S. I.; Wu, S. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks* **1999**, *12*, 783–789.
15. Murata, N.; Takenouchi, T.; Kanamori, T.; Eguchi, S. Information geometry of U-Boost and Bregman divergence. *Neural Comput.* **2004**, *16*, 1437–1481.
16. Ikeda, S.; Tanaka, T.; Amari, S. Information geometry of turbo and low-density parity-check codes. *IEEE Trans. Inf. Theory* **2004**, *50*, 1097–1114.
17. Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
18. Efron, B. Defining the curvature of a statistical problem (with application to second order efficiency) (with discussion). *Ann. Stat.* **1975**, *3*, 1189–1242.
19. Dawid, A.P. Discussion to Efron's paper. *Ann. Stat.* **1975**, *3*, 1231–1234.
20. Amari, S. Differential geometry of curved exponential families—Curvatures and information loss. *Ann. Stat.* **1982**, *10*, 357–385.
21. Cena, A. *Geometric Structures on the Non-Parametric Statistical Manifold*. Ph.D. Thesis, UniversitÀ Degli Studi di Milano, Milano, Italy, 2003.
22. Gibilisco, P.; Pistone, G. Connections on non-parametric statistical manifolds by Olicz space geometry. *Infin. Dimens. Anal. QU.* **1998**, *1*, 325–347.
23. Grasselli, M. Dual connections in nonparametric classical information geometry. *Ann. Inst. Stat. Math.* **2010**, *62*, 873–896.
24. Pistone, G.; Sempi, C. An infinite dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Stat.* **1995**, *33*, 1543–1561.
25. Zhang, J.; Hasto, P. Statistical manifold as an affine space: A functional equation approach. *J. Math. Psychol.* **2006**, *50*, 60–65.

26. Zhang, J.; Matsuzoe, H. Dualistic Differential Geometry Associated with a Convex Function. In *Advances in Applied Mathematics and Global Optimization*; Gao D.Y., Sherali, H.D., Eds.; Springer: New York, NY, USA, 2009; Volume III, Chapter 13, pp. 439–466.
27. Nomizu, K.; Sasaki, T. *Affine Differential Geometry—Geometry of Affine Immersions*; Cambridge University Press: Cambridge, MA, USA, 1994.
28. Simon U.; Schwenk-Schellschmidt, A.; Viesel, H. *Introduction to the Affine Differential Geometry of Hypersurfaces*; University of Tokyo Press: Tokyo, Japan, 1991.
29. Lauritzen, S. Statistical manifolds. In *Differential Geometry in Statistical Inference*; Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., Rao, C.R., Eds.; IMS: Hayward, CA, USA, 1987; Volume 10, pp. 163–216.
30. Lauritzen, S. Conjugate connections in statistical theory. In Proceedings of the Workshop on Geometrization of Statistical Theory; Dodson, C.T.J., Ed.; University of Lancaster: Lancaster, UK, 1987; pp. 33–51.
31. Kurose, T. Dual connections and affine geometry. *Math. Z* **1990**, *203*, 115–121.
32. Kurose, T. On the divergences of 1-conformally flat statistical manifolds. *Tôhoko Math. J.* **1994**, *46*, 427–433.
33. Matsuzoe H. On realization of conformally-projecively flat statistical manifolds and the divergences. *Hokkaido Math. J.* **1998**, *27*, 409–421.
34. Matsuzoe H. Geometry of contrast functions and conformal geometry. *Hokkaido Math. J.* **1999**, *29*, 175–191.
35. Calin, O.; Matsuzoe, H.; Zhang, J. Generalization of conjugate connections. In *Trends in Differential Geometry, Complex Analysis, and Mathematical Physics*; In Proceedings of the 9th International Workshop on Complex Structures, Integrability, and Vector Fields, Sofia, Bulgaria, 25–29 August 2008; pp. 24–34.
36. Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Stat.* **1983**, *11*, 793–803.
37. Eguchi, S. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.* **1985**, *15*, 341–391.
38. Eguchi, S. Geometry of minimum contrast. *Hiroshima Math. J.* **1992**, *22*, 631–647.
39. Chentsov, N.N. *Statistical Decision Rules and Optimal Inference*; AMS: Providence, RI, USA, 1982.
40. Naudts, J. Generalised exponential families and associated entropy functions. *Entropy* **2008**, *10*, 131–149.
41. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Phys.* **1967**, *7*, 200–217.
42. Zhu, H.Y.; Rohwer, R. Bayesian invariant measurements of generalization. *Neural Process. Lett.* **1995**, *2*, 28–31.

43. Zhu, H.Y.; Rohwer, R. Measurements of generalisation based on information geometry. In *Mathematics of Neural Networks: Models Algorithms and Applications*; In Proceedings of the Mathematics of Neural Networks and Applications (MANNA 1995), Oxford, UK, 3–7 July 1995; Ellacott, S.W., Mason, J.C., Anderson, I.J., Eds.; Kluwer: Boston, MA, USA, 1997; pp. 394–398.
44. Rao, C.R. Differential Metrics in Probability Spaces. In *Differential Geometry in Statistical Inference*; Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., Rao, C.R., Eds.; IMS: Hayward, CA, USA, 1987; Volume 10, Lecture, pp. 217–240.
45. Pistone G.; Rogantin M.P. The exponential statistical manifold: Mean parameters, orthogonality and space transformations. *Bernoulli* **1999**, *5*, 721–760.
46. Lang, S. *Differential and Riemannian Manifolds*; Springer-Verlag: New York, NY, USA, 1995.
47. Zhang, J. Referential Duality and Representational Duality on Statistical Manifolds. In Proceedings of the Second International Symposium on Information Geometry and Its Applications , Tokyo, Japan, 12–16 December 2005; pp. 58–67.
48. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195.
49. Basu, A.; Harris, I.R.; Hjort, N.; Jones, M. Robust and efficient estimation by minimising a density power divergence. *Biometrika* **1998**, *85*, 549–559.
50. Zhang J. A note on curvature of alpha-connections of a statistical manifold. *Ann. Inst. Stat. Math.* **2007**, *59*, 161–170.
51. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, **1967**, *2*, 229–318.
52. Cichocki, A.; Cruces, S; Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
53. Cichocki, A.; Amari, S. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
54. Zhang, J. Dual scaling between comparison and reference stimuli in multidimensional psychological space. *J. Math. Psychol.* **2004**, *48*, 409–424.
55. Zhang, J. Referential duality and representational duality in the scaling of multi-dimensional and infinite-dimensional stimulus space. In *Measurement and Representation of Sensations: Recent Progress in Psychological Theory*; Dzhafarov, E., Colonius, H., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2006.
56. Critchley, F.; Marriott, P.; Salmon, M. Preferred point geometry and statistical manifolds. *Ann. Stat.* **1993**, *21*, 1197–1224.
57. Critchley, F.; Marriott, P.; Salmon, M. Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence. *Ann. Stat.* **1994**, *22*, 1587–1602.
58. Critchley. F.; Marriott, P.K.; Salmon, M. On preferred point geometry in statistics. *J. Stat. Plan. Inference* **2002**, *102*, 229–245.
59. Marriott, P.; Vos, P. On the global geometry of parametric models and information recovery. *Bernoulli* **2004**, *10*, 639–649.
60. Zhu, H.-T.; Wei, B.-C. Some notes on preferred point α-geometry and α-divergence function. *Stat. Probab. Lett.* **1997**, *33*, 427–437.

61. Zhu, H.-T.; Wei, B.-C. Preferred point α-manifold and Amari's α-connections. *Stat. Probab. Lett.* **1997**, *36*, 219–229.
62. Ohara, A. Geometry of distributions associated with Tsallis statistics and properties of relative entropy minimization. *Phys. Lett. A* **2007**, *370*, 184–193.
63. Ohara, A.; Matsuzoe, H.; Amari, S. A dually at structure on the space of escort distributions. *J. Phys. Conf. Ser.* **2010**, *201*, No. 012012.
64. Amari. S.; Ohara, A. Geometry of q-exponential family of probability distributions. *Entropy* **2011**, *13*, 1170–1185.
65. Amari, S.; Ohara, A.; Matsuzoe, H. Geometry of deformed exponential families: Invariant, dually-flat and conformal geometry. *Physica A* **2012**, *391*, 4308–4319.
66. Shima, H. Compact locally Hessian manifolds. *Osaka J. Math.* **1978**, *15*, 509–513.
67. Shima, H.; Yagi, K. Geometry of Hessian manifolds. *Differ. Geom. Its Appl.* **1997**, *7*, 277–290.
68. Matumoto, T. Any statistical manifold has a contrast function—On the C^3-functions taking the minimum at the diagonal of the product manifold. *Hiroshima Math. J.* **1993**, *23*, 327–332.

© 2013 by the author; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).