

Nonparametric Information Geometry: Referential Duality and Representational Duality on Statistical Manifolds

Jun Zhang
525 East University
University of Michigan
Ann Arbor 48109
junz@umich.edu

Abstract

Classic parametric information geometry endows the manifold \mathcal{M}_θ of parametric probability density functions with a Riemannian structure in terms of (i) a Riemannian metric given by the Fisher information; (ii) a pair of dual connections (which form a parametric family of α -connection) 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}_θ (that of parametric probability density functions) to the manifold \mathcal{M} of non-parametric measurable functions on the sample space, 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; they specialize to Fisher information proper and α -divergence proper under appropriate contexts. The infinite-dimensional manifold \mathcal{M} has zero curvature for all α values, such that the curvature of \mathcal{M}_θ can be interpreted as arising from an embedding 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 submanifold is dually flat for $\alpha = \pm 1$, generalizing the results of Amari's α -embedding. The present analysis illuminates two different senses of duality in information geometry, one concerning the referential status of a point (function) in expressing the divergence function ("referential duality") and the other concerning its representation under an arbitrary monotone scaling ("representational duality").

KEYWORDS: convex function, Riemannian metric, α -connection, α -divergence, α -family, natural parameter, expectation parameters, Legendre-Fenchel duality

AMS2000 Classification – Primary: 58B20, 52A41; Secondary: 62A01

1 Introduction

Information geometry is the differential geometric study of the manifold of probability measures or probability density functions (see the monograph by Amari and Nagaoka, 2000). Its role in understanding asymptotic inference was summarized in (Amari, 1985; Barndorff-Nielsen, Cox, and Reid, 1986; Murray and Rice, 1993; Kass, 1989; Kass and Vos, 1997). Recently, information geometric methods have been applied to many areas of interest to statisticians, such as the study of estimating functions (Amari and Kumon, 1988; Amari and Kawanabe, 1997) and nuisance parameter (Eguchi, 1991), the dependency of Bayesian predictive distribution on prior selection (Komaki, 1996; Takeuchi, 1997), the class of invariant priors for Bayesian inference (Takeuchi and Amari, 2005; Matsuzoe, Takeuchi and Amari, in press), principle component analysis (Higuchi and Eguchi, 1998), independent component analysis and blind source separation (Amari and Cardoso, 1997; Amari, 1999, 2000; Minami and Eguchi, 2002), hierarchical analysis (Amari, 2001), information recovery (Marriott and Vos, 2004), etc. Information geometry also has also been used for deepening the understanding of machine learning and neural computation algorithms, including Boltzmann machine (Amari, Kurata, and Nagaoka, 1992), EM algorithm (Amari, 1995), natural gradient descent method (Amari, 1998; Yang and Amari, 1998; Amari, Park, and Fukumizu, 2000), support vector machine (Amari and Wu, 1999), boosting (Takenouchi and Eguchi, 2004; Murata, Takenouchi, Kanamori and Eguchi, 2004), belief network (Ikeda, Toshiyuki and Amari, 2004), turbo decoding (Ikeda, Tanaka and Amari, 2004), and others.

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 $\theta = [\theta^1, \dots, \theta^n]$. The collection of such probability functions, where each function is indexed by a point θ in \mathbb{R}^n , forms a manifold \mathcal{M}_θ under suitable conditions. Rao (1945) identified Fisher information to be the Riemannian metric for \mathcal{M}_θ . Efron (1975), 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, A.P. David (1975) invoked the differential geometric notion of linear connections on a manifold as preserving parallelism during vector transportation,

and pointed out other possible construction of connections on \mathcal{M}_θ , including the non-flat Levi-Civita connection associated with the Fisher metric. Amari (1982, 1985), 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}_θ . The e -connection ($\alpha = 1$) has vanishing components (i.e., becomes identically zero) on the manifold of the natural parameters of the exponential family of probability functions, whereas the m -connection ($\alpha = -1$) vanishes on the manifold of the mixture parameter of the mixture family of probability functions. So not only have $\Gamma^{(\pm 1)}$ zero curvatures for exponential/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 recently been generalized to non-parametric probability density functions using the tools of infinite-dimensional analysis (Gibilisco and Pistone, 1998; Grasselli, 2001, 2005; Cena, 2003), and to quantum systems using the tools of noncommutative algebra (Hasegawa, 1993, 1995; Petz and Hasegawa, 1996; Hasegawa and Petz, 1997; Petz and Sudár, 1999; Gibilisco and Isola, 1999; Grasselli and Streater, 2001; Jenčová, 2001; Grasselli, 2004). For the latter case, quantum analogues of the Fisher information, α -connections, and α -divergence have all been identified. Given this beautiful mathematical structure, it is important to have a thorough understanding of the mathematical foundation behind the classical information geometry and investigate whether its form (in both parametric and non-parametric cases) can be further generalized. The goal of the present paper is to investigate and extend the links among three inter-connected mathematical topics that underly information geometry, namely, (i) divergence functions measuring the asymmetric distance of any two points (density functions) on the manifold (the referential duality); (ii) convex analysis and the associated Legendre-Fenchel transformation linking 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.

The Riemannian manifold of parametric statistical models is a special kind, one that involves the theory of conjugate (a.k.a. dual) connections; historically, such

mathematically theory was independently developed to investigate hypersurface immersion (see Simon, Schwenk-Schellschmidt, and Viesel, 1991; Nomizu and Sasaki, 1994). Lauritzen (1987a) characterized the general differential geometric context under which this one-parameter family of α -connections arise, as well as the meaning of conjugacy for a pair of connections for statistical manifolds (Lauritzen, 1987b). Eguchi (1983, 1992) provided a generic way for inducing the metric and a pair of conjugate connections from an arbitrary divergence (contrast) function, whereas Kurose (1990, 1994) and then Matsuzoe (1998, 1999) elucidated their affine differential geometric relevance. This is the framework that the current exposition will adhere to. In the rest of this Introduction, the basic results of parametric information geometry will be reviewed. This includes not only the metric and conjugate connections of the Riemannian manifold \mathcal{M}_θ , but also how they are induced from the divergence function defined on $\mathcal{M}_\theta \times \mathcal{M}_\theta$. Here 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 the function space.

Our 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 sample space (with suitable measurability and smoothness constraints). When finite-dimensional affine embedding is allowed, our formulae reduce to the familiar ones associated with the exponential family. 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.

Our approach assumes that the Banach space of measurable functions considered here is properly normed and has the suitable topology to form an infinite-dimensional

manifold. Defining such topology is known to be difficult, and involves subtleties of infinite-dimensional analysis — the reader is referred to the pioneering work of Pistone and Sempi (1995) using the theory of Orlicz space for characterizing the coordinates of the exponential statistical manifold and a more recent discussion of it (Zhang and Hasto, in press). Such topological issues are assumed to have been already resolved for our current purpose.

1.1 Parametric information geometry revisited

1.1.1 Riemannian manifold, Fisher metric, and α -connections

Let \mathcal{M}_μ denote the space of probability density functions $p : \mathcal{X} \rightarrow R_+(\equiv R^+ \cup \{0\})$ defined on the sample space \mathcal{X} with background measure $d\mu = \mu(d\zeta)$

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

Here and throughout this paper, $\mathbb{E}_\mu\{\cdot\} = \int_{\mathcal{X}}\{\cdot\}d\mu$ denotes the expectation with respect to the background measure μ . We also denote $\mathbb{E}_p\{\cdot\} = \int_{\mathcal{X}}\{\cdot\}pd\mu$.

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 $\theta = [\theta^1, \dots, \theta^n]$. The space of parametric statistical models, under certain regularity conditions, forms a Riemannian manifold

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

with the so-called Fisher metric (Rao, 1945)

$$g_{ij}(\theta) = \mathbb{E}_\mu \left\{ p(\zeta|\theta) \frac{\partial \log p(\zeta|\theta)}{\partial \theta^i} \frac{\partial \log p(\zeta|\theta)}{\partial \theta^j} \right\} , \quad (1)$$

and α -connections (Amari, 1982; Chentsov, 1972/1982)

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

the α -connections satisfying the dualistic relation

$$\Gamma_{ij,k}^{*(\alpha)}(\theta) = \Gamma_{ij,k}^{(-\alpha)}(\theta) ; \quad (3)$$

here $*$ denotes conjugate connection (see below). 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 $\Gamma \leftrightarrow \Gamma^*$ 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 (Lauritzen, 1987b). Any Riemannian manifold with its metric g and conjugate connections Γ, Γ^* given in the form of (1)–(3) is called a *statistical manifold* and is denoted as $\{\mathcal{M}_\theta, g, \Gamma^{(\pm\alpha)}\}$.

1.1.2 Exponential family, mixture family, and their generalization

An exponential family is defined as

$$p^{(e)}(\zeta|\theta) = \exp \left(F_0(\zeta) + \sum_i \theta^i F_i(\zeta) - \Phi(\theta) \right) \quad (4)$$

where θ is its natural parameter, $F_i(\zeta)$ ($i = 1, \dots, n$) is a set of linearly independent functions on the support \mathcal{X} , and the cumulant generating function (“potential function”) $\Phi(\theta)$ is

$$\Phi(\theta) = \log E_\mu \left\{ \exp \left(F_0(\zeta) + \sum_i \theta^i F_i(\zeta) \right) \right\} . \quad (5)$$

Substituting (4) into (1) and (2), the Fisher metric and the α -connections are simply

$$g_{ij}(\theta) = \frac{\partial^2 \Phi(\theta)}{\partial \theta^i \partial \theta^j} \quad (6)$$

and

$$\Gamma_{ij,k}^{(\alpha)}(\theta) = \frac{1-\alpha}{2} \frac{\partial^3 \Phi(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k} , \quad (7)$$

whereas the Riemannian curvature tensor (of an α -connection) is given by (Amari, 1985, p.106)

$$R_{ij\mu\nu}^{(\alpha)}(\theta) = \frac{1-\alpha^2}{4} \sum_{l,k} (\Phi_{il\nu} \Phi_{jk\mu} - \Phi_{il\mu} \Phi_{jk\nu}) \Phi^{lk} , \quad (8)$$

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 $\Gamma_{ij,k}^{(1)}$ vanishes, due to (7), on the

manifold formed by $p^{(e)}(\cdot|\theta)$ in which the natural parameter θ serves as the coordinates.

On the other hand, the mixture family

$$p^{(m)}(\zeta|\theta) = \sum_i \theta^i F_i(\zeta) , \quad (9)$$

when viewed as a manifold of 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 $\Gamma_{ij,k}^{(-1)}$. 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 (Amari, 1985; Amari and Nagaoka, 2000) of density functions $p(\zeta|\theta)$ which satisfy

$$l^{(\alpha)}(p) = F_0(\zeta) + \sum_i \theta^i F_i(\zeta) \quad (10)$$

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

$$l^{(\alpha)}(t) = \begin{cases} \log t & \alpha = 1 \\ \frac{2}{1-\alpha} t^{(1-\alpha)/2} & \alpha \neq 1 \end{cases} . \quad (11)$$

In such a case, the density functions form the so-called α -affine manifold. The Fisher metric and α -connections, under such α -representation, have the following expressions:

$$g_{ij}(\theta) = E_\mu \left\{ \frac{\partial l^{(\alpha)}(p(\zeta|\theta))}{\partial \theta^i} \frac{\partial l^{(-\alpha)}(p(\zeta|\theta))}{\partial \theta^j} \right\} , \quad (12)$$

$$\Gamma_{ij,k}^{(\alpha)}(\theta) = E_\mu \left\{ \frac{\partial^2 l^{(\alpha)}(p(\zeta|\theta))}{\partial \theta^i \partial \theta^j} \frac{\partial l^{(-\alpha)}(p(\zeta|\theta))}{\partial \theta^k} \right\} . \quad (13)$$

Clearly, on an α -affine manifold, $\Gamma^{(\alpha)}$ is identically zero by virtue of the definition (10) — $\pm\alpha$ -connections are dually flat for the α -family.

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, α -divergence

Divergence functions are distance-like quantities; they measure the directed (asymmetric) 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* (a.k.a. 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)

$$K(p, q) = \int \left(q - p - p \log \frac{q}{p} \right) d\mu = K^*(q, p), \quad (14)$$

with a unique, global minimum of zero when $p = q$. For the exponential family (4), expression (14) takes the form of the so-called *Bregman divergence* (Bregman, 1967) defined on $\Theta \times \Theta \subseteq \mathbb{R}^n \times \mathbb{R}^n$:

$$B_\Phi(\theta_p, \theta_q) = \Phi(\theta_p) - \Phi(\theta_q) - \langle \theta_p - \theta_q, \partial\Phi(\theta_q) \rangle, \quad (15)$$

where Φ is the potential function (5), ∂ is the gradient operator, and $\langle \cdot, \cdot \rangle$ denotes the standard inner product of two vectors. The Bregman divergence (15) 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 as¹

$$\mathcal{A}^{(\alpha)}(p, q) = \frac{4}{1 - \alpha^2} \mathbb{E}_\mu \left\{ \frac{1 - \alpha}{2} p + \frac{1 + \alpha}{2} q - p^{\frac{1-\alpha}{2}} q^{\frac{1+\alpha}{2}} \right\}, \quad (16)$$

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

$$\begin{aligned} \lim_{\alpha \rightarrow -1} \mathcal{A}^{(\alpha)}(p, q) &= K(p, q) = K^*(q, p); \\ \lim_{\alpha \rightarrow 1} \mathcal{A}^{(\alpha)}(p, q) &= K^*(p, q) = K(q, p). \end{aligned}$$

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

¹Traditionally (see Amari, 1982, 1985), the term $\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q$ is replaced by 1 in the integrand of (16), and the term $q - p$ is absent in the integrand of (14); this is trivially true when p, q are probability densities with normalization of 1. Zhu and Rohwer (1995, 1997), 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 term “divergence function” is appropriate. However, if the underlying space is an infinite-dimensional function space, that is, the manifold \mathcal{M}_μ of non-parametric probability densities p and q , then the term “divergence functional” ought to be used. The latter implicitly defines a divergence function (through pullback) if the probability densities are embedded into a finite-dimensional submanifold \mathcal{M}_θ in the case of a parametric statistical model $p(\cdot|\theta)$. As an example, for the exponential family (4), the Kullback-Leiber divergence (14) in terms of p and q implicitly defines a divergence in terms of θ_p, θ_q , i.e., the Bregman divergence (15). 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 (or “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 (1) by its second order properties, and a pair of conjugate connections Γ, Γ^* in the forms of (2) and (3) by its third order properties — the relations were first formulated by Eguchi (1983), which we are going to recall 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 $p \in \mathcal{M}$ is a semi-positive bilinear form $\langle \cdot, \cdot \rangle$ 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 , and $\Sigma(\mathcal{M})$ denotes the collection of all vector fields on \mathcal{M} . Then

$$g(u, v) = \langle u, v \rangle$$

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 the point p . A pair of connections ∇, ∇^* are said to be *conjugate* to each other if

$$d_w g(u, v) = \langle \nabla_w u, v \rangle + \langle u, \nabla_w^* v \rangle, \quad (17)$$

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

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}^* . \quad (18)$$

The “contravariant” form Γ_{ij}^l of the affine connection defined by

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^l \partial_l$$

is related to the “covariant” form $\Gamma_{ij,k}$ through

$$\sum_l g_{lk} \Gamma_{ij}^l = \Gamma_{ij,k} .$$

The Riemannian metric and conjugate connections on a statistical manifold can be induced by a divergence function $\mathcal{D} : \mathcal{M} \times \mathcal{M} \rightarrow R_+$ that satisfies

- (i) $\mathcal{D}(p, q) \geq 0 \forall p, q \in \mathcal{M}$ with equality holding iff $p = q$;
- (ii) $(d_u)_p \mathcal{D}(p, q)|_{p=q} = (d_v)_q \mathcal{D}(p, q)|_{p=q} = 0$, 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 (1983) showed that any such divergence function \mathcal{D} satisfying (i) and (ii) will induce a Riemannian metric g and a pair of connections ∇, ∇^* via

$$g(u, v) = -(d_u)_p (d_v)_q \mathcal{D}(p, q)|_{p=q} ; \quad (19)$$

$$\langle \nabla_w u, v \rangle = -(d_w)_p (d_u)_p (d_v)_q \mathcal{D}(p, q)|_{p=q} ; \quad (20)$$

$$\langle u, \nabla_w^* v \rangle = -(d_w)_q (d_v)_q (d_u)_p \mathcal{D}(p, q)|_{p=q} . \quad (21)$$

In index-laden component forms, they are

$$g_{ij} = -(\partial_i)_p (\partial_j)_q \mathcal{D}(p, q)|_{p=q} ; \quad (22)$$

$$\Gamma_{ij,k} \equiv \langle \nabla_{\partial_i} \partial_j, \partial_k \rangle = -(\partial_i)_p (\partial_j)_p (\partial_k)_q \mathcal{D}(p, q)|_{p=q} ; \quad (23)$$

$$\Gamma_{ij,k}^* \equiv \langle \partial_k, \nabla_{\partial_i}^* \partial_j \rangle = -(\partial_i)_q (\partial_j)_q (\partial_k)_p \mathcal{D}(p, q)|_{p=q} . \quad (24)$$

Equations (19)–(21) in coordinate-free form, or (22)–(24) 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 (17) or (18), respectively. These relations are the stepping stones going from a divergence function defining (generally) asymmetric distances between

a pair of points on a manifold at large to the dual Riemannian geometric structure on the same manifold in the small. Below we are particularly interested in the coordinate-free form, because once we construct divergence functional on the infinite-dimensional function space (the Kullback-Leibler divergence being a special example), then we may 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 by Zhu and Rohwer, 1995; 1997), the family of Jensen difference related to the Shannon entropy function (Rao, 1987), both specializing to Kullback-Leibler divergence as a limiting case. The latter class include the Bregman divergence (Bregman 1967), also called “geometric divergence” (Kurose, 1994), which turns out to be identical to the “canonical divergence” (Amari and Nagaoka, 2000) 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 (Amari, 1982; 1985). Murata et al. (2004) 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 $\alpha \leftrightarrow -\alpha$ in such divergence functions corresponds to an exchange of the two points the divergence functions measured (generally in an asymmetric fashion), while $\alpha \leftrightarrow -\alpha$ in the induced connections corresponds to the conjugacy operation for the metric-specified pairing of two connections operating on the dual vector spaces. Hence, our approach to divergence functions from convex analysis will address both 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 (1) and the α -connections (2) (or equivalently, (12) and (13) under α -embedding) while still staying within the framework of Lauritzen (1987a) 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 (Gibilisco and Pistone, 1998; Grasselli, 2001; 2005). Our approach exploits the coordinate-free version of the Eguchi relations (19)–(21) 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 (the coordinate-free analogue of the index-laden Christoffel symbol Γ for the finite-dimensional case). 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 an ambient space. From this perspective, statistical curvature (curvature of a statistical manifold) can be viewed as an embedding curvature, that is, curvature arising out of restricting to the submanifold \mathcal{M}_μ of normalized and positive-valued functions (i.e., non-parametric statistical manifold), or to the finite-dimensional submanifold \mathcal{M}_θ (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 David (1975), 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 (1982, 1985) generalized this observation to any α -family (i.e., a density function that is affine under α -embedding): the α -connection is flat (indeed $\Gamma_{ij,k}^{(\alpha)}$ vanishes) for the α -affine manifold (which specializes 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 α -connection and the α parameter in α -embedding are one and the same, and thereby to 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}_μ 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 statistical model can be affine embedded under any ρ -representation, where a strictly increasing function $\rho : \mathbb{R} \rightarrow \mathbb{R}$ generalizes the α -embedding function (11). In such cases, there exists a strictly convex potential function, akin to (5) for the exponential statistical model, that will reduce the Fisher metric and α -connections to the forms of (6) and (7). 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 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 scales (representational duality). Roughly speaking, the (± 1) -duality reflects the former, whereas the e/m -duality reflects the latter. In traditional analysis, they are 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 Lauritzen's (1987a) 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 a set of measurable functions mapping \mathcal{X} to \mathbb{R} which, under a suitable topology and with integrability and smoothness assumptions, form a manifold \mathcal{M} of infinite dimension such that each point on \mathcal{M} is a function $p : \mathcal{X} \rightarrow \mathbb{R}$ on the sample space \mathcal{X} ; here p is assumed to belong

to some Banach space under a suitable norm (e.g. Orlicz space, as adopted by Pistone and Sempi, 1995; Gibilisco and Pistone, 1998; Pistone and Rogantin, 1999; Grasselli, 2001, 2005; Cena, 2003). Compared with the original setting of Pistone and Sempi (1995), 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} as an ambient space to embed a statistical manifold \mathcal{M}_μ 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}_μ as the projection of a natural connection on \mathcal{M} is then “borrowed” over from the finite-dimensional setting to this infinite-dimensional setting (the rigorous proof of the correspondence, however, is beyond the scope of the current exposition).

The structure of the rest of the paper is as follows. Section 2 will deal with information geometry under infinite-dimensional setting and Section 3 under 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.

2 Information Geometry on the Infinite-Dimensional Function Space

In this section, we first review basic apparatus of differentiable manifold with particular emphasis paid on 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 inter-related 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 $F : \mathcal{U} \rightarrow \mathbb{R}$ 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 $\{x(t) \in \mathcal{U} : t \in I \subset \mathbb{R}\}$, 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 $v = \left. \frac{dx}{dt} \right|_{t=0}$ (without loss of generality, we assume $0 \in I$). We use $T_{x_0}(\mathcal{M})$ to denote the space of all tangent vectors at a given x_0 ; it is obviously a vector space, called the tangent space (associated with the point x_0). 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 number, denoted $d_v F$, called the direction derivative of F :

$$d_v F = \lim_{t \rightarrow 0} \frac{1}{t} (F(x(t)) - F(x_0)) .$$

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

$$d_v u = \lim_{t \rightarrow 0} \frac{1}{t} (u(x(t)) - u(x_0)) .$$

In our setting, $\mathcal{U} \subset \mathcal{M}$ represents a collection of measurable functions with common support, i.e., smooth functions (satisfying certain regularization conditions) defined on the sample space \mathcal{X} with a background measure μ . A point x_0 on the manifold is a specific measurable function $p : \zeta \mapsto p(\zeta)$ defined for all $\zeta \in \mathcal{X}$, the sample space, which is assumed to have (in general uncountably) infinite dimension. We call any function that maps $\mathcal{X} \rightarrow \mathbb{R}$ a ζ -function. Implicitly assumed here is that, under a suitable topology, and with certain restrictions (including measurability and smoothness assumptions), the collection of ζ -functions form a manifold denoted as \mathcal{M} above. On this manifold, any function $p \rightarrow F(p)$ is referred to, in the following, as a ζ -functional, because it takes in a ζ -function p and outputs a number. The set of ζ -functionals on \mathcal{M} is denoted as $\mathcal{F}(\mathcal{M})$.² A curve on \mathcal{M} passing through a

²The terms “function”, “ ζ -function”, “ ζ -functional”, can be at times very confusing to a seasoned mathematician, who would prefer to use the universal terminology of “function” while carefully

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 ”, so $p(\zeta|t)$ is a ζ -function “parameterized” by t — a one-parameter family of ζ -functions is formed as t varies.³ More generally, $p(\zeta|\theta)$, where $\theta = [\theta^1, \dots, \theta^n] \in \Theta \subseteq \mathbb{R}^n$, is a ζ -function indexed by n parameters $\theta^1, \dots, \theta^n$. As θ varies, $p(\zeta|\theta)$ represents an embedding $\widetilde{\mathcal{M}}_\theta \subset \mathcal{M}$ where

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

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

In the infinite-dimensional setting, the tangent vector v , defined as

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

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

$$d_v(F(p)) = \lim_{t \rightarrow 0} \frac{F(p(\zeta|t)) - F(p(\zeta|0))}{t} ,$$

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:

$$F(p) = \int_{\mathcal{X}} f(p(\zeta)) d\mu = E_\mu\{f(p(\zeta))\}$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ 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

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

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

A vector field in the current setting, as a cross-section of \mathcal{TM} , takes in a ζ -function and generates a ζ -function. We denote a vector field as $u(\zeta|p) \in \Sigma(\mathcal{M})$, where the specifying its domain and range. However, aiming at a broader audience of statisticians, we try to be intuitive and therefore at times deliberately set apart a “function” from a “functional”. In our usage, ζ -function always refers to a real-valued function defined on the sample space (e.g. density functions, random-variable functions), and ζ -functional refers to a mapping from one or more ζ -functions to a real number. On the other hand, the word “mapping” or “map” is the general term to refer the correspondence between a domain and a range.

³Intuitively, the value of $p(\zeta|t)$ is determined by first specifying t , the member of the family, and then ζ , the sample point. So it is really a function from $\mathcal{X} \times I \rightarrow \mathbb{R}$.

variable following the “|” sign indicates that u depends on the point $p(\zeta)$, an element of the base manifold \mathcal{M} . The following are examples of vector field (when $\mathcal{X} = \mathbb{R}$): $\zeta p(\zeta)$, $p(2\zeta)$, $p(\zeta)p(\zeta + 2)$, $\int p(\zeta)d\mu + \zeta^2$ (the last one being a constant vector field defined for all points of the base manifold). Though 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 $\rho : \mathbb{R} \rightarrow \mathbb{R}$.

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 , which is another ζ -function, is

$$d_v u(\zeta|p) = \lim_{t \rightarrow 0} \frac{u(\zeta|p(\zeta|t)) - u(\zeta|p(\zeta))}{t} .$$

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 $\rho(p(\zeta))$ in the direction of $v(\zeta)$ is

$$d_v \rho(p(\zeta|t)) = \lim_{t \rightarrow 0} \frac{\rho(p(\zeta|t)) - \rho(p)}{t} = \rho'(p(\zeta)) v(\zeta) .$$

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

$$\nabla_w F(p) = d_w F(p) .$$

When operating on a vector field, say $u(\zeta|p)$, ∇_w is defined as (see Lang, 1995)

$$(\nabla_w u)(\zeta) = (d_w u)(\zeta) + B(\zeta|w(\zeta|p), u(\zeta|p)) \quad (25)$$

where $B : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \rightarrow \Sigma(\mathcal{M})$ 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^* formally as

$$(\nabla_w^* u)(\zeta) = (d_w u)(\zeta) + B^*(\zeta|w(\zeta), u(\zeta)) .$$

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

$$R(u, v, w) = R(u, v)w = \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u, v]} w , \quad (26)$$

where

$$[u, v] = d_u v - d_v u .$$

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

$$T(u, v) = \nabla_u v - \nabla_v u - [u, v] . \quad (27)$$

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 $\mathcal{D} : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}_+$ maps two ζ -functions to a non-negative real number. To the extent that ζ -functions can be parameterized by finite-dimensional vectors $\theta \in \Theta \subseteq \mathbb{R}^n$, a divergence *functional* on $\mathcal{M} \times \mathcal{M}$ will implicitly induce a divergence *function* on the parameter space $\Theta \times \Theta \subseteq \mathbb{R}^n \times \mathbb{R}^n$ (technically, this is called “pullback”). In this section, we will discuss the general form of 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 $f : \mathbb{R} \rightarrow \mathbb{R}$. We recall the *fundamental convex inequality* that defines a strictly convex function f :

$$f\left(\frac{1-\alpha}{2}\gamma + \frac{1+\alpha}{2}\delta\right) \leq \frac{1-\alpha}{2}f(\gamma) + \frac{1+\alpha}{2}f(\delta) , \quad (28)$$

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,

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

where

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

(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,

$$\begin{aligned} l^{(\alpha)}(\gamma, \delta) &= \frac{f(\delta) - f(\epsilon)}{\delta - \epsilon} = \frac{1}{\delta - \gamma} \frac{2}{1 - \alpha} \left(f(\delta) - f\left(\frac{1 - \alpha}{2}\gamma + \frac{1 + \alpha}{2}\delta\right) \right), \\ \tilde{l}^{(\alpha)}(\gamma, \delta) &= \frac{f(\gamma) - f(\epsilon)}{\gamma - \epsilon} = \frac{1}{\delta - \gamma} \frac{2}{1 + \alpha} \left(f\left(\frac{1 - \alpha}{2}\gamma + \frac{1 + \alpha}{2}\delta\right) - f(\gamma) \right), \end{aligned}$$

with skew symmetry

$$l^{(-\alpha)}(\gamma, \delta) = -\tilde{l}^{(\alpha)}(\delta, \gamma), \quad \tilde{l}^{(-\alpha)}(\gamma, \delta) = -l^{(\alpha)}(\delta, \gamma).$$

As $\alpha : -1 \rightarrow 1$ (i.e., as point ϵ moves from γ to δ , the two fixed ends), both $l^{(\alpha)}(\gamma, \delta)$ and $\tilde{l}^{(\alpha)}(\gamma, \delta)$ are increasing function of α , but the chord theorem dictates that the latter is always \leq the former. In fact, their difference has a non-negative value

$$\begin{aligned} 0 &\leq l^{(\alpha)}(\gamma, \delta) - \tilde{l}^{(\alpha)}(\gamma, \delta) = l^{(-\alpha)}(\gamma, \delta) - \tilde{l}^{(-\alpha)}(\gamma, \delta) \\ &= \frac{1}{\delta - \gamma} \frac{4}{1 - \alpha^2} \left(\frac{1 - \alpha}{2} f(\gamma) + \frac{1 + \alpha}{2} f(\delta) - f\left(\frac{1 - \alpha}{2}\gamma + \frac{1 + \alpha}{2}\delta\right) \right). \end{aligned} \quad (29)$$

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 $p, q : \mathcal{X} \rightarrow \mathbb{R}$ 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.

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

$$\mathcal{D}_{f,\rho}^{(\alpha)}(p, q) = \frac{4}{1 - \alpha^2} \mathbb{E}_\mu \left\{ \frac{1 - \alpha}{2} f(\rho(p)) + \frac{1 + \alpha}{2} f(\rho(q)) - f\left(\frac{1 - \alpha}{2}\rho(p) + \frac{1 + \alpha}{2}\rho(q)\right) \right\} \quad (30)$$

is non-negative and equals zero if and only

$$p(\zeta) = q(\zeta) \quad a.s.$$

Proof. See Section 4.

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

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

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. Also, in order to ensure that the integrals in (30) are well defined, we require p, q to be elements of the set

$$\mathcal{B} = \{p(\zeta) : E_\mu\{f(\rho(p))\} < \infty\} .$$

$\mathcal{D}^{(\alpha)}$ -divergence was first introduced in Zhang (2004a). It generalized the familiar α -divergence (16) — take $\rho(p) = \log p$ and $f(p) = e^p$, then $\mathcal{D}_{f,\rho}^{(\alpha)}(p, q) = \mathcal{A}^{(\alpha)}(p, q)$, the latter in turn specializes to the Kullback-Leibler divergence as $\alpha \rightarrow \pm 1$.

2.2.2 Conjugate scaled representations of measurable functions

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

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

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

$$f(\delta) + f^*(\lambda) - \gamma \lambda \geq 0$$

can be cast as the Young's inequality

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

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 $\rho : \mathbb{R} \rightarrow \mathbb{R}$. 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 $f : \mathbb{R} \rightarrow \mathbb{R}$ if

$$\tau(p) = f'(\rho(p)) = ((f^*)')^{-1}(\rho(p)) \iff \rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p)) . \quad (31)$$

As an example, we may let $\rho(p) = l^{(\alpha)}(p)$ to be the α -representation where $l^{(\alpha)}$ is given by (11), and the conjugate representation is the $(-\alpha)$ -representation $\tau(p) = l^{(-\alpha)}(p)$:

$$\rho(t) = l^{(\alpha)}(t) \longleftrightarrow \tau(p) = l^{(-\alpha)}(p) . \quad (32)$$

In this case

$$f(t) = \frac{2}{1+\alpha} \left(\left(\frac{1-\alpha}{2} \right) t \right)^{\frac{2}{1-\alpha}} , \quad f^*(t) = \frac{2}{1-\alpha} \left(\left(\frac{1+\alpha}{2} \right) t \right)^{\frac{2}{1+\alpha}} , \quad (33)$$

so that

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

both linear in p . More generally, strictly increasing functions from $\mathbb{R} \rightarrow \mathbb{R}$ form a group, with functional composition as group composition operation and functional inverse as 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,

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

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

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

This is to say

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

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

$$M_{\rho}^{(\alpha)}(\gamma, \delta) = \rho^{-1} \left(\frac{1-\alpha}{2} \rho(\gamma) + \frac{1+\alpha}{2} \rho(\delta) \right) \quad (34)$$

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)

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

2.2.3 Canonical divergence

The use of a pair of strictly increasing functions f, f^* allow us to define, in parallel with $\mathcal{D}_{f,\rho}^{(\alpha)}(p, q)$ given in (30), the conjugate family $\mathcal{D}_{f^*,\tau}^{(\alpha)}(p, q)$. The two families turn out to have the same form when $\alpha = \pm 1$; this is the so-called *canonical divergence*.

Taking the limit $\alpha \rightarrow -1$, the inequality (29) becomes

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

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

$$\begin{aligned} \mathcal{D}_{f,\rho}^{(-1)}(p, q) &= \mathbb{E}_\mu \{ f(\rho(q)) - f(\rho(p)) - (\rho(q) - \rho(p))f'(\rho(p)) \} \\ &= \mathbb{E}_\mu \{ f^*(\tau(p)) - f^*(\tau(q)) - (\tau(p) - \tau(q))(f^*)'(\tau(q)) \} = \mathcal{D}_{f^*,\tau}^{(-1)}(q, p) ; \\ \mathcal{D}_{f,\rho}^{(1)}(p, q) &= \mathbb{E}_\mu \{ f(\rho(p)) - f(\rho(q)) - (\rho(p) - \rho(q))f'(\rho(q)) \} \\ &= \mathbb{E}_\mu \{ f^*(\tau(q)) - f^*(\tau(p)) - (\tau(q) - \tau(p))(f^*)'(\tau(p)) \} = \mathcal{D}_{f^*,\tau}^{(1)}(q, p) . \end{aligned}$$

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

$$\mathcal{A}_f(\rho(p), \tau(q)) = \mathbb{E}_\mu \{ f(\rho(p)) + f^*(\tau(q)) - \rho(p) \tau(q) \} = \mathcal{A}_{f^*}(\tau(q), \rho(p)) \quad (35)$$

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 $\alpha = 1 \leftrightarrow \alpha = -1$ and between $(f, \rho) \leftrightarrow (f^*, \tau)$:

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

We can see that under conjugate $(\pm\alpha)$ -representations (32), \mathcal{A}_f is simply the α -divergence proper $\mathcal{A}^{(\alpha)}$:

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

In fact,

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

is an expression of Hölder'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 (30) generalizes the α -divergence in a sensible way. Now we investigate the metric and conjugate connections such divergence functionals induce; this is accomplished by invoking the Eguchi relations (19)–(21).

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 $g : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \rightarrow \mathcal{F}(\mathcal{M})$ is given by

$$g(u, v) = \mathbb{E}_\mu \{ g(p(\zeta)) u(\zeta|p) v(\zeta|p) \} \quad (36)$$

where

$$g(t) = f''(\rho(t))(\rho'(t))^2 ; \quad (37)$$

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

$$\nabla_w^{(\alpha)} u = (d_w u)(\zeta) + B^{(\alpha)}(p(\zeta)) u(\zeta|p) w(\zeta|p) \quad (38)$$

where

$$B^{(\alpha)}(t) = \frac{1 - \alpha}{2} \frac{f'''(\rho(t))\rho'(t)}{f''(\rho(t))} + \frac{\rho''(t)}{\rho'(t)} ; \quad (39)$$

- (iii) the family of conjugate covariant derivatives is

$$\nabla_w^{*(\alpha)} u = (d_w u)(\zeta) + B^{(-\alpha)}(p(\zeta)) u(\zeta|p) w(\zeta|p) .$$

Proof. See Section 4.

Note that the $g(\cdot)$ term in (36) and the $B^{(\alpha)}(\cdot)$ term in the covariant derivatives (38) 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 (36) and $B^{(\alpha)}(\cdot)$ in expressing the covariant derivatives (38) can be expressed in dualistic forms:

$$g(t) = \rho'(t) \tau'(t) \quad (40)$$

and

$$B^{(\alpha)}(t) = \frac{d}{dt} \left(\frac{1+\alpha}{2} \log \rho'(t) + \frac{1-\alpha}{2} \log \tau'(t) \right). \quad (41)$$

Proof. See Section 4.

Corollary 3 makes it immediately evident that the Riemannian metrics induced by $\mathcal{D}_{f,\rho}^{(\alpha)}(p, q)$ and by $\mathcal{D}_{f^*,\tau}^{(\alpha)}(p, q)$ 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 $\alpha \leftrightarrow -\alpha$. 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 (38) 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}_μ of non-parametric probability density functions and the manifold \mathcal{M}_θ of parametric density functions, and any curvature on \mathcal{M}_μ or \mathcal{M}_θ may be interpreted as arising from embedding or restriction to a lower dimensional space.

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 $\rho(t) = f^{-1}(\varepsilon t) \leftrightarrow f(t) = \varepsilon \rho^{-1}(t)$, so that the divergence functional becomes

$$\mathcal{D}_\rho^{(\alpha)}(p, q) = \frac{4\varepsilon}{1-\alpha^2} \int_{\mathcal{X}} \left\{ \frac{1-\alpha}{2} p + \frac{1+\alpha}{2} q - \rho^{-1} \left(\frac{1-\alpha}{2} \rho(p) + \frac{1+\alpha}{2} \rho(q) \right) \right\} d\mu. \quad (42)$$

Now the second term in the integrand is just the quasi-linear mean $M_\rho^{(\alpha)}$ introduced in (34), where ρ is strictly increasing and concave here. As an example, take $\rho(p) = \log p$, $c = 1$, 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 (16), while

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

is an immediate generalization of the extended Kullback-Leibler divergence in (14).

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

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

then (see Zhang, 2004a) $\rho(p) = l^{(\beta)}(p)$, so (42) becomes a two-parameter family

$$\mathcal{D}^{(\alpha, \beta)}(p, q) \equiv \frac{4}{1 - \alpha^2} \frac{2}{1 + \beta} \mathbb{E}_\mu \left\{ \frac{1 - \alpha}{2} p + \frac{1 + \alpha}{2} q - \left(\frac{1 - \alpha}{2} p^{\frac{1-\beta}{2}} + \frac{1 + \alpha}{2} q^{\frac{1-\beta}{2}} \right)^{\frac{2}{1-\beta}} \right\} . \quad (43)$$

Here $(\alpha, \beta) \in [-1, 1] \times [-1, 1]$, and $\varepsilon = 2/(1+\beta)$ in (42) 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 Csiszar (1967). 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 (16) with

$$\begin{aligned} \lim_{\alpha \rightarrow -1} \mathcal{D}^{(\alpha, \beta)}(p, q) &= \mathcal{A}^{(-\beta)}(p, q) , \\ \lim_{\alpha \rightarrow 1} \mathcal{D}^{(\alpha, \beta)}(p, q) &= \mathcal{A}^{(\beta)}(p, q) , \\ \lim_{\beta \rightarrow 1} \mathcal{D}^{(\alpha, \beta)}(p, q) &= \mathcal{A}^{(\alpha)}(p, q) , \\ \lim_{\beta \rightarrow -1} \mathcal{D}^{(\alpha, \beta)}(p, q) &= J^{(\alpha)}(p, q) \end{aligned}$$

where $J^{(\alpha)}$ denotes the Jensen difference discussed by Rao (1987)

$$\begin{aligned} J^{(\alpha)}(p, q) \equiv & \frac{4}{1 - \alpha^2} \mathbb{E}_\mu \left(\frac{1 - \alpha}{2} p \log p + \frac{1 + \alpha}{2} q \log q \right. \\ & \left. - \left(\frac{1 - \alpha}{2} p + \frac{1 + \alpha}{2} q \right) \log \left(\frac{1 - \alpha}{2} p + \frac{1 + \alpha}{2} q \right) \right) . \end{aligned}$$

$J^{(\alpha)}$ reduces to the Kullback-Leibler divergence (14) when $\alpha \rightarrow \pm 1$. Lastly, we note that in $\mathcal{D}^{(\alpha, \beta)}$, when either α or β equals 0, the Levi-Civita connection results.

With respect to the geometry induced from the (α, β) -divergence, 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

$$\begin{aligned} g(u, v) &= \int_{\mathcal{X}} \frac{1}{p} u v d\mu , \\ \nabla_u^{(\alpha, \beta)} v &= d_u v - \frac{1 + \alpha\beta}{2p} u v , \\ \nabla_u^{*(\alpha, \beta)} v &= d_u v - \frac{1 - \alpha\beta}{2p} u v , \end{aligned}$$

where $u, v \in \Sigma(\mathcal{M})$.

Proof. Immediate upon substituting (32) and (33) to (37) and (39). \diamond

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

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

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

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

The same is true when $\lim_{\alpha \rightarrow 1} \nabla^{(\alpha, \beta)}$ (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 $\theta = [\theta^1, \dots, \theta^n] \in \Theta \subseteq \mathbb{R}^n$. 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 $\theta_p, \theta_q \in \Theta$. 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

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

where $p(\cdot|\theta)$ is a mapping from $\Theta \rightarrow \mathcal{M}$, taking the value of θ as the input and generating a ζ -function as the output. We also denote the manifold of parametric statistical model as

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

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}_θ). 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}_θ) take the form

$$u = \frac{\partial p(\zeta|\theta)}{\partial \theta^i}, \quad v = \frac{\partial p(\zeta|\theta)}{\partial \theta^k}, \quad w = \frac{\partial p(\zeta|\theta)}{\partial \theta^j}. \quad (44)$$

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

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

$$g_{ij}(\theta) = \mathbb{E}_\mu \left\{ f''(\rho) \frac{\partial \rho}{\partial \theta^i} \frac{\partial \rho}{\partial \theta^j} \right\} \quad (45)$$

and the α -connections take the form

$$\Gamma_{ij,k}^{(\alpha)}(\theta) = \mathbb{E}_\mu \left\{ \frac{1-\alpha}{2} f'''(\rho) A_{ijk} + f''(\rho) B_{ijk} \right\}, \quad (46)$$

$$\Gamma_{ij,k}^{*(\alpha)}(\theta) = \mathbb{E}_\mu \left\{ \frac{1+\alpha}{2} f'''(\rho) A_{ijk} + f''(\rho) B_{ijk} \right\}. \quad (47)$$

where

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

Proof. See Section 4.

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

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 reduces to the Fisher information and α -connections of the exponential family in (1) and (2).

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

$$g_{ij}(\theta) = E_\mu \left\{ \frac{\partial \rho}{\partial \theta^i} \frac{\partial \tau}{\partial \theta^j} \right\} , \quad (48)$$

$$\Gamma_{ij,k}^{(\alpha)}(\theta) = E_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial^2 \tau}{\partial \theta^i \partial \theta^j} \frac{\partial \rho}{\partial \theta^k} + \frac{1+\alpha}{2} \frac{\partial^2 \rho}{\partial \theta^i \partial \theta^j} \frac{\partial \tau}{\partial \theta^k} \right\} , \quad (49)$$

$$\Gamma_{ij,k}^{*(\alpha)}(\theta) = E_\mu \left\{ \frac{1+\alpha}{2} \frac{\partial^2 \tau}{\partial \theta^i \partial \theta^j} \frac{\partial \rho}{\partial \theta^k} + \frac{1-\alpha}{2} \frac{\partial^2 \rho}{\partial \theta^i \partial \theta^j} \frac{\partial \tau}{\partial \theta^k} \right\} . \quad (50)$$

Proof. See Section 4.

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

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

with

$$\tilde{\Gamma}_{ij,k}^{(\alpha)}(\theta) = \Gamma_{ij,k}^{(-\alpha)}(\theta) , \quad \tilde{\Gamma}_{ij,k}^{*(\alpha)}(\theta) = \Gamma_{ij,k}^{(\alpha)}(\theta) .$$

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

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

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}_θ), we have the two-parameter family of divergence *functions* $D^{(\alpha, \beta)}(\theta_p, \theta_q)$ defined by

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

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

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

$$\begin{aligned} g_{ij}(\theta) &= \mathbb{E}_p \left\{ \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right\} \\ \Gamma_{ij,k}^{(\alpha, \beta)}(\theta) &= \mathbb{E}_p \left\{ \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} \frac{\partial \log p}{\partial \theta^k} + \frac{1 - \alpha\beta}{2} \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \frac{\partial \log p}{\partial \theta^k} \right\} , \\ \Gamma_{ij,k}^{*(\alpha, \beta)}(\theta) &= \mathbb{E}_p \left\{ \frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} \frac{\partial \log p}{\partial \theta^k} + \frac{1 + \alpha\beta}{2} \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \frac{\partial \log p}{\partial \theta^k} \right\} . \end{aligned}$$

Proof. See Section 4.

This two-parameter family of affine connections $\Gamma_{ij,k}^{(\alpha, \beta)}(\theta)$, 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 Lauritzen (1987a))

$$\Gamma_{ij,k}^{(\alpha, \beta)}(\theta) = \Gamma_{ij,k}^{(-\alpha, -\beta)}(\theta)$$

with biduality compactly expressed as

$$\Gamma_{ij,k}^{*(\alpha, \beta)}(\theta) = \Gamma_{ij,k}^{(-\alpha, \beta)}(\theta) = \Gamma_{ij,k}^{(\alpha, -\beta)}(\theta) . \quad (51)$$

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

$$\rho(p(\zeta|\theta)) = \sum_i \theta^i \lambda_i(\zeta) . \quad (52)$$

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

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

$$\eta_i = \int_{\mathcal{X}} \tau(p(\zeta)) \lambda_i(\zeta) d\mu \quad (53)$$

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

The above notion of ρ -affinity is a generalization of α -affine manifolds (Amari, 1985; Amari and Nagaoka, 2000), where ρ - and τ -representations are just α - and $(-\alpha)$ -representations, respectively.

3.2.1 Biorthogonality of natural and expectation parameters

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

(i) the function

$$\Phi(\theta) = \int_{\mathcal{X}} f(\rho(p(\zeta|\theta))) d\mu \quad (54)$$

is strictly convex;

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

$$D_{\Phi}^{(\alpha)}(\theta_p, \theta_q) = \frac{4}{1-\alpha^2} \left(\frac{1-\alpha}{2} \Phi(\theta_p) + \frac{1+\alpha}{2} \Phi(\theta_q) - \Phi \left(\frac{1-\alpha}{2} \theta_p + \frac{1+\alpha}{2} \theta_q \right) \right) ; \quad (55)$$

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

$$\begin{aligned} g_{ij}(\theta) &= \Phi_{ij} ; & \Gamma_{ij,k}^{(\alpha)}(\theta) &= \frac{1-\alpha}{2} \Phi_{ijk} = \Gamma_{ij,k}^{*(-\alpha)}(\theta) ; \\ R_{ij\mu\nu}^{(\alpha)}(\theta) &= \frac{1-\alpha^2}{4} \sum_{l,k} (\Phi_{il\nu} \Phi_{jk\mu} - \Phi_{il\mu} \Phi_{jk\nu}) \Phi^{lk} = R_{ij\mu\nu}^{*(\alpha)}(\theta) . \end{aligned}$$

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

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

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

Proof. See Section 4.

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

$$\Phi^*(\eta) = \langle \eta, (\partial\Phi)^{-1}(\eta) \rangle - \Phi((\partial\Phi)^{-1}(\eta)), \quad (56)$$

where $\partial\Phi$ stands for the gradient (sub-differential) of Φ , and $\langle \cdot, \cdot \rangle$ 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 semi-definite. The Legendre-Fenchel inequality (56) can be expressed using dual variables θ, η as

$$\Phi(\theta) + \Phi^*(\eta) - \sum_i \eta_i \theta^i \geq 0;$$

where equality holds if and only if

$$\theta = (\partial\Phi^*)(\eta) = (\partial\Phi)^{-1}(\eta) \longleftrightarrow \eta = \partial\Phi(\theta) = (\partial\Phi^*)^{-1}(\theta). \quad (57)$$

COROLLARY 10. For ρ -affine manifold,

(i) define

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

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 :

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

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

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

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 $\Phi^*(\eta)$ as the generalized entropy function. For an exponential family, the two are well known to form one-to-one correspondence; either can be used to that index 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 vanishes, i.e., $R_{ij\mu\nu}^{(\pm 1)}(\theta) = 0$. In this case, there exists a coordinate system under which either $\Gamma_{ij,k}^{*(-1)}(\theta) = 0$ or $\Gamma_{ij,k}^{(1)}(\theta) = 0$. This is the well-studied “dually flat” parametric statistical manifold (Amari, 1982, 1985; Amari and Nagaoka, 2000), under which divergence functions have a unique, canonical form.

COROLLARY 11. When $\alpha \rightarrow \pm 1$, $D_{\Phi}^{(\alpha)}$ reduces to the Bregman divergence (15)

$$\begin{aligned} D_{\Phi}^{(-1)}(\theta_p, \theta_q) &= D_{\Phi}^{(1)}(\theta_q, \theta_p) = \Phi(\theta_q) - \Phi(\theta_p) - \langle \theta_q - \theta_p, \partial\Phi(\theta_p) \rangle = B_{\Phi}(\theta_q, \theta_p) , \\ D_{\Phi}^{(1)}(\theta_p, \theta_q) &= D_{\Phi}^{(-1)}(\theta_q, \theta_p) = \Phi(\theta_p) - \Phi(\theta_q) - \langle \theta_p - \theta_q, \partial\Phi(\theta_q) \rangle = B_{\Phi}(\theta_p, \theta_q) , \end{aligned}$$

or equivalently, to the canonical divergence functions

$$\begin{aligned} D_{\Phi}^{(1)}(\theta_p, (\partial\Phi)^{-1}(\eta_q)) &= \Phi(\theta_p) + \Phi^*(\eta_q) - \langle \theta_p, \eta_q \rangle \equiv A_{\Phi}(\theta_p, \eta_q) , \quad (58) \\ D_{\Phi}^{(-1)}((\partial\Phi)^{-1}(\theta_p), \theta_q) &= \Phi(\theta_q) + \Phi^*(\eta_p) - \langle \eta_p, \theta_q \rangle \equiv A_{\Phi^*}(\eta_p, \theta_q) . \end{aligned}$$

Proof. Immediate by substitution using the definition (56). \diamond

The canonical divergence $A_{\Phi}(\theta_p, \eta_q)$ based on the Legendre-Fenchel inequality was introduced by Amari (1982, 1985), where the functions Φ, Φ^* , the cumulant generating functions of an exponential family, were referred to as dual “potential” functions. This form (58) 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 $D_{\Phi}^{(-1)} \leftrightarrow D_{\Phi}^{(1)}$ and between $D_{\Phi^*}^{(-1)} \leftrightarrow D_{\Phi^*}^{(1)}$, the other between $D_{\Phi}^{(-1)} \leftrightarrow D_{\Phi^*}^{(-1)}$ and between $D_{\Phi}^{(1)} \leftrightarrow D_{\Phi^*}^{(1)}$. 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,

$$D_{\Phi}^{(-1)}(\theta_p, \theta_q) = D_{\Phi^*}^{(-1)}(\partial\Phi(\theta_q), \partial\Phi(\theta_p)) = D_{\Phi^*}^{(1)}(\partial\Phi(\theta_p), \partial\Phi(\theta_q)) = D_{\Phi}^{(1)}(\theta_q, \theta_p) .$$

The biduality is compactly reflected in the canonical divergence as

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

4 Proofs

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

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

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 (28) 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 Φ guarantees

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

or explicitly

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

This, along with $1 - \alpha^2 < 1$ 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 (28) to the three points γ, λ and their convex mixture $\delta = \frac{2}{1-\alpha} \lambda + \frac{-1-\alpha}{1-\alpha} \gamma$. Finally, continuity of $[\mathbb{F}^{(\alpha)}](\gamma, \delta)$ 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 (30), note that $(d_u)_p$ means that the functional derivative is with respect to p only (point q is treated as fixed)

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

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

$$(d_v)_q \left((d_u)_p \mathcal{D}_{f,\rho}^{(\alpha)}(p, q) \right) = - \int_{\mathcal{X}} f'' \left(\frac{1-\alpha}{2} \rho(p) + \frac{1+\alpha}{2} \rho(q) \right) \rho'(p) \rho'(q) u v d\mu . \quad (59)$$

Setting $p = q$ and invoking (19) yields (36) with (37).

Next, applying $(d_w)_p$ to (59), and realizing that u, v are both vector fields,

$$\begin{aligned} (d_w)_p \left((d_v)_q (d_u)_p \mathcal{D}_{f,\rho}^{(\alpha)}(p, q) \right) &= \\ &- \int_{\mathcal{X}} \frac{1-\alpha}{2} \left(f''' \left(\frac{1-\alpha}{2} \rho(p) + \frac{1+\alpha}{2} \rho(q) \right) (\rho'(p))^2 \rho'(q) u v w d\mu \right. \\ &- \left. \int_{\mathcal{X}} f'' \left(\frac{1-\alpha}{2} \rho(p) + \frac{1+\alpha}{2} \rho(q) \right) \rho'(q) v (\rho''(p) u w + \rho'(p) (d_w u)) d\mu \right) . \end{aligned}$$

Setting $p = q$, invoking (20) and

$$g(\nabla_w u, v) = \int f''(\rho) (\rho')^2 (\nabla_w u)(\zeta|p) v(\zeta|p) d\mu , \quad (60)$$

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

$$f''(\rho) (\rho')^2 \nabla_w^{(\alpha)} u = \frac{1-\alpha}{2} f'''(\rho) (\rho')^3 u w + f''(\rho) \rho' v (\rho'' u w + \rho' (d_w u)) .$$

where ρ is the shorthand for $\rho(p(\zeta))$. Remember that $\nabla_w^{(\alpha)} u$ is a ζ -function, the above equation yields

$$\nabla_w^{(\alpha)} u = d_w u + \frac{1-\alpha}{2} \frac{f'''(\rho)}{f''(\rho)} \rho' u w + \frac{\rho''}{\rho'} u w = d_w u + \left(\frac{1-\alpha}{2} \frac{f'''(\rho)}{f''(\rho)} \rho' + \frac{\rho''}{\rho'} \right) u w$$

Thus we obtain (38) with (39). The expression for $\nabla^{*(\alpha)}$ is analogous. \diamond

PROOF OF COROLLARY 3. From the identities

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

we obtain (40) and (41) 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 (Lang, 1995, p.226) appears to miss some terms. From (25),

$$d_u(\nabla_v w) = d_u(d_v w) + B(d_u v, w) + B(v, d_u w) + (d_u B)(v, w)$$

so that

$$\nabla_u(\nabla_v w) = d_u(d_v w) + B(d_u v, w) + B(v, d_u w) + (d_u B)(v, w) + B(u, d_v w) + B(u, B(v, w)) ;$$

here $d_u B = B' u$ 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 $u \rightarrow v$ in the above.

Now

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

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

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

Substitute them into (26), we get a general expression of Riemann curvature tensor in infinite-dimensional setting

$$R(u, v, w) = B(u, B(v, w)) - B(v, B(u, w)) + (d_u B)(v, w) - (d_v B)(u, w) ,$$

The expression for $T(u, v)$ in (27) becomes

$$T(u, v) = B(u, v) - B(v, u) .$$

In the current case,

$$B(u, v) = B^{(\alpha)}(p(\zeta))u(\zeta|p)v(\zeta|p) .$$

Substituting this into the above, and realizing that $(d_u B)(v, w)$ is simply $(B^{(\alpha)})' u v w$, we immediately have $R^{(\alpha)}(u, v, w) = 0$, as well as $T^{(\alpha)}(u, v) = 0$. \diamond

PROOF OF PROPOSITION 6. Given (44) as the tangent vector fields for parametric models, we note that

$$d_u \rho = \rho' u = \rho' \frac{\partial p}{\partial \theta^i} = \frac{\partial \rho(p)}{\partial \theta^i} , \quad (61)$$

$$d_w \rho = \rho' w = \rho' \frac{\partial p}{\partial \theta^j} = \frac{\partial \rho(p)}{\partial \theta^j} , \quad (62)$$

so (45) follows. Next, from

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

we have

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

Observing $\Gamma_{ij,k} = \langle \nabla_w u, v \rangle$, expression (46) results after substituting the above derived expressions into (38) with (39). \diamond

PROOF OF COROLLARY 7. Applying (61) and (62) to (36) with (40) immediately yields (48). Next, from Corollary 3,

$$B(\alpha) = \frac{1 - \alpha}{2} \frac{\tau''}{\tau'} + \frac{1 + \alpha}{2} \frac{\rho''}{\rho'} .$$

It follows that

$$\begin{aligned} \Gamma_{ij,k}^{(\alpha)} &= \langle \nabla_w^{(\alpha)} u, v \rangle = \left(\frac{1 - \alpha}{2} (\rho' \tau'' u w + \rho' \tau' d_w u) + \frac{1 + \alpha}{2} (\rho'' \tau' u w + \rho' \tau' d_w u) \right) v \\ &= \rho' v \frac{1 - \alpha}{2} d_w (d_u \tau) + \tau' v \frac{1 + \alpha}{2} d_w (d_u \rho) = \frac{1 - \alpha}{2} (d_v \rho) d_w (d_u \tau) + \frac{1 + \alpha}{2} (d_v \tau) d_w (d_u \rho) . \end{aligned}$$

Note that

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

Substituting into (38) with (41) yields (49) and (50). \diamond

PROOF OF PROPOSITION 9. The assumption (52) implies that $\frac{\partial \rho}{\partial \theta^i} = \lambda_i(\zeta)$, so from (45)

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

That the above expression is positive definite is seen by observing

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

for any $\xi = [\xi^1, \dots, \xi^n] \in \mathbb{R}^n$, due to linear independence of the λ_i 's and the strict convexity of f . Hence, $\Phi(\theta)$ is strictly convex in θ , proving (i). An immediate consequence is that expression (55) is non-negative and vanishes if and only if $\theta_p = \theta_q$. This establishes (ii), i.e., $D_{\Phi}^{(\alpha)}(\theta_p, \theta_q)$ is a divergence function. Part (iii) follows from a straight-forward application of Eguchi relations (22)–(24). \diamond

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

$$f^*(\tau(p(\zeta|\theta)) + f(\rho(p(\zeta|\theta)))) = f'(\rho(p(\zeta|\theta))) \rho(p(\zeta|\theta)) .$$

From (54), taking derivative with respect to θ^i while noting that $p(\zeta|\theta)$ satisfies (52) gives

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

and that

$$\begin{aligned} \sum_i \theta^i \frac{\partial \Phi(\theta)}{\partial \theta^i} - \Phi(\theta) &= \int_{\mathcal{X}} \left\{ f' \left(\sum_j \theta^j \lambda_j(\zeta) \right) \left(\sum_i \theta^i \lambda_i(\zeta) \right) - f \left(\sum_j \theta^j \lambda_j(\zeta) \right) \right\} d\mu \\ &= \int_{\mathcal{X}} f^*(\tau(p(\zeta|\theta))) d\mu = \tilde{\Phi}(\theta) . \end{aligned}$$

It follows from (56) that Φ^* as defined in (i) is the conjugate of Φ , and that the relation in (ii) is the basic Legendre-Fenchel duality. Finally, biorthogonality of η and θ as expressed by (iii) also becomes evident on account of (ii). \diamond

5 Summary and Future Directions

This paper constructs a family of divergence functionals, induced by any smooth and strictly convex function, to measure the asymmetric “distance” between two measurable functions defined on the sample space and properly normed. 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 $f : \mathbb{R} \rightarrow \mathbb{R}$ (or $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$), with the convex mixture coefficient as the α -parameter in the induced geometry. Referential duality is associated with $\alpha \leftrightarrow -\alpha$, and representational duality is associated with convex conjugacy $f \leftrightarrow f^*$ (or $\Phi \leftrightarrow \Phi^*$). Thus, our analysis reveals that 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 asymmetric status for a referent and for a comparison object), while common in psychological and behavioral contexts (e.g. Dzhafarov, 2002; Zhang, 2004b, in press), has always been implicitly acknowledged in statistics. Formal investigation of such asymmetry between a reference probability distribution and comparison probability distribution in constructing divergence functions leads to the framework of preferred point geometry (Critchley, Marriott, and Salmon, 1993, 1994, 2002; Zhu and Wei, 1997a, b). Preferred point geometry reformulates Amari's (1982) expected geometry and Barndorff-Nelsen's (1988) 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 Simon, Schwenk-Schellschmidt & Viesel, 1991; Nomizu & Sasaki, 1994). Kurose (1990, 1994), Matsuzoe (1998, 1999), Uohashi, Ohara, and Fujii (2000a,b) and Uohashi (2002) investigated central-affine immersion of statistical manifolds. Since the pseudo-linear (i.e., ρ -affine) manifold (Section 3.2) is nothing but a Hessian manifold (Shima, 1978; Shima and Yagi, 1997), a follow-up study of hypersurface immersion of our generalized dual Riemannian geometry would be an interesting direction to pursue.

It should be noted that, while any divergence function determines uniquely a statistical manifold (conceptualized, according to Lauritzen (1987a), as a manifold with a Riemannian metric and a pair of conjugate connections), 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 (Matsumoto, 1993), it is not unique in general, except when the connections are dually flat (traditionally, $\alpha = \pm 1$), 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; the 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.

Reference

- Amari, S. (1982). Differential geometry of curved exponential families – curvatures and information loss. *Annals of Statistics*, 10: 357-385.
- Amari, S. (1985). *Differential Geometric Methods in Statistics*, Lecture Notes in Statistics, 28, Springer-Verlag, New York. Reprinted in 1990.
- Amari, S. (1995). Information geometry of EM and em algorithms for neural networks. *Neural Networks*, 8:1379-1408.
- Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10: 251-276.
- Amari, S. (1999). Superefficiency in blind source separation. *IEEE Transactions on Signal Processing*, 47:936-944.
- Amari, S. (2000). Estimating functions of independent component analysis for temporally correlated signals. *Neural Computation*, 12: 1155-1179.
- Amari, S. (2001). Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47: 1701-1711.
- Amari, S. and Cardoso, J.-F. (1997). Blind source separation – semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45: 2692-2700.
- Amari, S. and Kawanabe, M. (1997) Information geometry of estimating functions in semiparametric statistical models. *Bernoulli*, 3: 29-54.
- Amari, S. and Kumon, M. (1988). Estimation in the presence of infinitely many nuisance parameters – Geometry of estimating functions. *Annals of Statistics* 16, 1044-1068.
- Amari, S., Kurata, K., and Nagaoka, H. (1992). Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, 3: 260-271.
- Amari, S. and Nagaoka, H. (2000). *Method of Information Geometry*. AMS monograph, Oxford University Press.
- Amari, S., Park, H., and Fukumizu, K. (2000). Adaptive method of realizing natural

- gradient learning for multilayer perceptrons. *Neural Computation*, 12:1399-1409.
- Amari, S. and Wu, S. (1999). Improving support vector machine classifiers by modifying kernel function. *Neural Networks*, 12: 783-789.
- Barndorff-Nielsen, O.E. (1988). *Parametric Statistical Models and Likelihood*. Lecture Notes in Statistics, 50. Springer-Verlag.
- Barndorff-Nielsen, O.E., Cox, R.D., and Reid, N. (1986). The role of differential geometry in statistical theory. *International Statistical Review*, 54: 83-96.
- Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Physics*, 7: 200-217.
- Cena, A. (2003). *Geometric Structures on the Non-Parametric Statistical Manifold*. Doctoral Dissertation, Università Degli Studi Di Milano.
- Chentsov, N. N. (1972/1982). *Statistical Decision Rules and Optimal Inference*. AMS, Rhode Island, USA, 1982. (Originally published in Russian, Nauka, Moscow, 1972).
- Critchley, F., Marriott, P., and Salmon, M. (1993). Preferred point geometry and statistical manifolds. *The Annals of Statistics*, 21: 1197-1224.
- Critchley, F., Marriott, P., and Salmon, M. (1994). Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence. *The Annals of Statistics*, 22: 1587-1602.
- Critchley, F., Marriott, P.K., and Salmon, M. (2002). On preferred point geometry in statistics. *JSPI*, 102: 229-245.
- Dawid, A. P. (1975). Discussion to Efron's paper. *Annals of Statistics*, 3: 1231-1234.
- Dzhafarov, E.N. (2002). Multidimensional Fechnerian scaling: Pairwise comparisons, regular minimality, and nonconstant self-similarity. *Journal of Mathematical Psychology*, 46, 583-608.
- Efron, B. (1975). Defining the curvature of a statistical problem (with application to second order efficiency) (with discussion). *Annals of Statistics*, 3: 1189-1242.

- Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. *Annals of Statistics*, 11: 793-803.
- Eguchi, S. (1991). A geometric look at nuisance parameter effect of local powers in testing hypothesis. *Ann. Inst. Statist. Math.*, 43: 245-260.
- Eguchi, S. (1992). Geometry of minimum contrast. *Hiroshima Mathematical Journal*, 22: 631-647.
- Gibilisco, P. and Pistone, G. (1998). Connections on non-parametric statistical manifolds by Olicz space geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 1: 325-347.
- Gibilisco, P. and Isola, T. (1999). Connections on statistical manifolds of density operators by geometry of noncommutative L^p -spaces. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 2: 169-178.
- Grasselli, M.R. (2001). *Classical and Quantum Information Geometry*, Ph.D. thesis, King's College London.
- Grasselli, M. (2004). Duality, monotonicity and the Wigner-Yanase-Dyson metrics. *Infinite Dimensional Analysis, Quantum Probability and Related Topics* 7, 215-232.
- Grasselli, M. (2005). Dual connections in nonparametric classical information geometry, to appear in the Annals of the Institute for Statistical Mathematics.
- Grasselli, M. and Streater, R.F. (2001). On the uniqueness of the Chentsov metric in quantum information geometry. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, 4, No. 2, 173-182, 2001.
- Hasegawa, H. (1993). α -Divergence of the non-commutative information geometry. *Reports on mathematical physics*. 33: 87-93.
- Hasegawa, H. (1995). Non-commutative extension of the information geometry. In Belavkin V.P., Hirota, O. and Hudson, R.L. (Eds) *Proceedings of International Workshop on Quantum Communication, Computing, and Measurement, Nottingham*. Plenum Press, New York.
- Hasegawa, H. and Petz D. (1997). Non-commutative extension of information ge-

- ometry II. In Hirota et al (Eds) *Proceedings of International Workshop on Quantum Communication, Computing, and Measurement*. Plenum Press, New York.
- Higuchi, I. and Eguchi, S. (1998) The influence function of principal component analysis by self-organizing rule. *Neural Computation* 10, 1435-1444.
- Ikeda, S., Tanaka, T., and Amari, S. (2004). Information geometry of turbo and low-density parity-check codes. *IEEE transaction on Information Theory*, 50: 1097-1114.
- Ikeda, S. Toshiyuki, T., and Amari, S. (2004). Stochastic reasoning, free energy, and information geometry. *Neural Computation*, 16: 1779-1810.
- Jenčová, A. (2001). Geometry of quantum states: dual connections and divergence functions. *Reports on Mathematical Physics*, 47: 121-138.
- Kass, R.E. (1989). The geometry of asymptotic inference (with discussion), *Statistical Science*, 4: 188-234.
- Kass, R. E. and Vos, P. W.(1997) *Geometric Foundation of Asymptotic Inference*. John Wiley and Sons: New York.
- Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, 83, 299-313.
- Kurose, T. (1990). Dual connections and affine geometry. *Mathematische Zeitschrift*, 203:115-121.
- Kurose, T. (1994). On the divergences of 1-conformally flat statistical manifolds. *Töhoko Mathematical Journal*, 46:427-433.
- Lang, S. (1995). *Differential and Riemannian Manifolds*. Springer-Verlag, New York.
- Lauritzen, S. (1987a). Statistical manifolds. In Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., and Rao, C.R. (Eds.) *Differential Geometry in Statistical Inference*, IMS Lecture Notes, Vol. 10, Hayward, CA (pp. 163-216).
- Lauritzen, S. (1987b). Conjugate connections in statistical theory. In C.T.J. Dodson (Ed.) *Proc. Workshop on Geometrization of Statistical Theory*. Univ. of Lancaster,

pp.33-51.

Marriott, P. and Vos, P. (2004). On the global geometry of parametric models and information recovery. *Bernoulli*, 10, 639649

Matsuzoe, H. (1998). On realization of conformally-projectively flat statistical manifolds and the divergences. *Hokkaido Mathematical Journal*, 27: 409-421.

Matsuzoe, H. (1999). Geometry of contrast functions and conformal geometry. *Hiroshima Mathematical Journal*, 29: 175-191.

Matsuzoe, H., Takeuchi, J., and Amari, S. (in press). Equiaffine structures on statistical manifolds and Bayesian statistics. *Differential Geometry and its Applications*.

Matumoto, T. (1993). Any statistical manifold has a contrast function – On the C^3 -functions taking the minimum at the diagonal of the product manifold. *Hiroshima Mathematical Journal*, 23: 327-332.

Minami, M. and Eguchi, S. (2002). Robust blind source separation by beta-divergence. *Neural Computation* 14, 1859-1886.

Murata, N., Takenouchi, T., Kanamori, T., and Eguchi, S. (2004). Information geometry of U-Boost and Bregman divergence. *Neural Computation*, 16: 1437-1481.

Murray, M.K. and Rice, J.W. (1993). *Differential Geometry and Statistics*. Chapman & Hall, London.

Nomizu, K. and Sasaki, T. (1994). *Affine Differential Geometry – Geometry of Affine Immersions*. Cambridge University Press.

Petz, D. and Hasegawa, H. (1996). On the Riemannian metric of α -entropies of density matrices. *Letters in Mathematical Physics*, 38:221-225.

Petz, D. and Sudár, Cs. (1999). Extending the Fisher metric to density matrices. In Barndorff-Nielsen, O.E. and Vendel Jensen(?), E.B. (Eds) *Geometry in Present Days Science*, World Scientific Publishing (pp 21-34).

Pistone, G. and Sempi, C. (1995). An infinite dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Annals of Statistics*, 33:1543-1561.

- Pistone G., Rogantin M. P. (1999). The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, 5, pp.721-760.
- Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37: 81-91.
- Rao, C.R. (1987). Differential metrics in probability spaces. In Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., and Rao, C.R. (Eds.) *Differential Geometry in Statistical Inference*, IMS Lecture Notes, Vol. 10, Hayward, CA (pp. 217-240).
- Shima, H. (1978). Compact locally Hessian manifolds. *Osaka Journal of Mathematics*, 15: 509-513.
- Shima, H. and Yagi, K. (1997). Geometry of Hessian manifolds. *Differential Geometry and its Applications*, 7: 277-290.
- Simon U., Schwenk-Schellschmidt, A., and Viesel, H. (1991). *Introduction to the Affine Differential Geometry of Hypersurfaces*. Lecture notes of the Science University of Tokyo.
- Takenouchi, T. and Eguchi, S. (2004). Robustifying AdaBoost by adding the naive error rate. *Neural Computation*, 16: 767-787.
- Takeuchi, J. (1997). Characterization of the Bayes estimator and the MDL estimator for exponential families. *IEEE Transactions on Information Theory*, 43: 1165-1174.
- Takeuchi, J. and Amari, S. (2005). α -Parallel prior and its properties. *IEEE Transaction on Information Theory* (to appear).
- Uohashi, K., Ohara, A., and Fujii, T. (2000a). 1-Conformally flat statistical submanifolds. *Osaka Journal of Mathematics*, 37: 501-507.
- Uohashi, K., Ohara, A., and Fujii, T. (2000b). Foliations and divergences of flat statistical manifolds. *Hiroshima Math J.*, 30: 403-414.
- Uohashi, K. (2002). On α -conformal equivalence of statistical manifolds. *J. Geom.*, 75, 179-184.
- Yang, H. H., and Amari, S. (1998). Complexity issues in natural gradient descent method for training multilayer perceptrons. *Neural Computation*, 10, 2137-2157.

- Zhang, J. (2004a). Divergence function, duality, and convex analysis. *Neural Computation*, 16: 159-195.
- Zhang, J. (2004b). Dual scaling between comparison and reference stimuli in multi-dimensional psychological space. *Journal of Mathematical Psychology*, 48:409-424.
- Zhang, J. (in press). Referential duality and representational duality in the scaling of multi-dimensional and infinite-dimensional stimulus space. In Dzhafarov, E. and Colonius, H. (Eds.) *Measurement and representation of sensations: Recent progress in psychological theory*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Zhang, J. and Hasto, P. (in press). Statistical manifold as an affine space: A functional equation approach. *Journal of Mathematical Psychology*.
- Zhu, H.-T. and Wei, B.-C. (1997a). Some notes on preferred point α -geometry and α -divergence function. *Statistics and Probability Letters*. 33: 427-437.
- Zhu, H.-T. and Wei, B.-C. (1997b). Preferred point α -manifold and Amari's α -connections. *Statistics and Probability Letters*. 36: 219-229.
- Zhu, H.Y. and Rohwer, R. (1995). Bayesian invariant measurements of generalization. *Neural Processing Letter*, 2: 28-31.
- Zhu, H.Y. and Rohwer, R. (1997) Measurements of generalisation based on information geometry. In S.W. Ellacott, J.C. Mason, and I.J. Anderson (Eds.) *Mathematics of Neural Networks: Models Algorithms and Applications*, pp 394-398. Kluwer, Boston. Proc. Math. of Neural Networks and Appl. Conf (MANNA), Oxford, 3-7 July 1995.