2nd International Symposium on
Information Geometry and Its Applications
December 12-16, 2005, Tokyo
Pages 58-67

REFERENTIAL DUALITY AND REPRESENTATIONAL DUALITY ON STATISTICAL MANIFOLDS

JUN ZHANG

1. Background and Goals

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

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

where $\mathrm{E}_\mu\{\cdot\} = \int_{\mathcal{X}} \{\cdot\}\, d\mu$ denotes the expectation with respect to the background measure μ. We assume that $\mathcal{M}_\mu$ is a differentiable manifold under a suitably chosen topology [1]. A parametric family of density functions, called a parametric statistical model, $p(\cdot|\theta)$, is the association, for each n-dimensional vector $\theta = [\theta^1, \cdots, \theta^n]$, of a density function $\theta \mapsto p(\cdot|\theta)$ such that

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

forms a manifold. In classical information geometry, there are two core interrelated concepts: the first is the notion of a divergence function (or functional), defined on $\mathcal{M}_\theta \times \mathcal{M}_\theta$ (or on $\mathcal{M}_\mu \times \mathcal{M}_\mu$) that is non-negative and vanishes only on the "diagonal" points (see below); it measures the directed (asymmetric) "distance" between two points at large on the manifold. The second is the notion of a metric and a pair of dual connections on $\mathcal{M}_\theta$ (or $\mathcal{M}_\mu$), locally measuring distance as well as angles and defining parallelism, respectively. These two core concepts — divergence and geometry — are intimately related, as we review below.

1.1. Divergence function(al)s and the geometry they induce.

Take the familiar example of *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)

$$K(p,q) = \mathrm{E}_\mu \left\{ q - p - p \log \frac{q}{p} \right\} = K^*(q,p) \,,$$

with a unique, global minimum of zero when $p = q$. More generally, one-parameter families of divergence have been introduce, such as the family of α-*divergence*

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

and the family of Jensen difference [2]

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

It is easily seen that

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

Eguchi [3], working in the setting of parametric statistical models, showed that any divergence function $\mathcal{D}(p,q)$ (that is differentiable in its arguments up to third order and that satisfies $\mathcal{D}(p,q) \geq 0$ with $p = q$ as the only global minimum achieving the equality with vanishing first order derivatives) induces a Riemannian metric g and a pair of connections Γ, Γ^* (given in local coordinates):

$$g_{ij} = -(\partial_i)_p\,(\partial_j)_q\,\mathcal{D}(p,q)\big|_{p=q}\,; \tag{3}$$

$$\Gamma_{ij,k} = -(\partial_i)_p\,(\partial_j)_p\,(\partial_k)_q\mathcal{D}(p,q)\big|_{p=q}\,; \tag{4}$$

$$\Gamma^*_{ij,k} = -(\partial_i)_q\,(\partial_j)_q\,(\partial_k)_p\mathcal{D}(p,q)\big|_{p=q}\,, \tag{5}$$

where $(\partial_i)_p$ denotes $\partial/\partial\theta^i$ applied to the expression $p(\cdot|\theta)$ only. Explicitly calculated using α-divergence $\mathcal{A}^{(\alpha)}$, they are

$$\begin{aligned}
g_{ij}(\theta) &= \mathrm{E}_\mu\left\{p(\zeta|\theta)\,\frac{\partial \log p(\zeta|\theta)}{\partial\theta^i}\,\frac{\partial \log p(\zeta|\theta)}{\partial\theta^j}\right\}, \\
\Gamma^{(\alpha)}_{ij,k}(\theta) &= \mathrm{E}_\mu\left\{\left(\frac{1-\alpha}{2}\,\frac{\partial \log p(\zeta|\theta)}{\partial\theta^i}\,\frac{\partial \log p(\zeta|\theta)}{\partial\theta^j} + \frac{\partial^2 \log p(\zeta|\theta)}{\partial\theta^i\partial\theta^j}\right)\frac{\partial p(\zeta|\theta)}{\partial\theta^k}\right\}, \\
\Gamma^{*(\alpha)}_{ij,k}(\theta) &= \mathrm{E}_\mu\left\{\left(\frac{1+\alpha}{2}\,\frac{\partial \log p(\zeta|\theta)}{\partial\theta^i}\,\frac{\partial \log p(\zeta|\theta)}{\partial\theta^j} + \frac{\partial^2 \log p(\zeta|\theta)}{\partial\theta^i\partial\theta^j}\right)\frac{\partial p(\zeta|\theta)}{\partial\theta^k}\right\}.
\end{aligned}$$

We note $\Gamma^{*(\alpha)}_{ij,k}(\theta) = \Gamma^{(-\alpha)}_{ij,k}(\theta)$. For Jensen difference $\mathcal{J}^{(\alpha)}$, the metric is also $g_{ij}(\theta)$, but the connections are $\Gamma^{(\mp\alpha)}_{ij,k}(\theta)$, i.e., conjugate to those induced by $\mathcal{A}^{(\alpha)}$.

The metric g and the pair of connections $\Gamma^{(\alpha)}, \Gamma^{*(\alpha)}$ as derived above satisfy

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

By definition, such Γ, Γ^* are said to be "dual" [4], [5] or "conjugate" [6] with respect to the metric g, and that $(\mathcal{M}_\theta, g, \Gamma, \Gamma^*)$ forms a "statistical manifold." In this sense, the Eguchi relations (3)–(5) generate, from an arbitrary divergence function $\mathcal{D}$, the dualistic geometry of a statistical manifold.

The goal of the present work is to investigate the notion of duality in information geometry by elucidating the precise meaning of the α-parameter under different contexts, and to extend information geometric formulation to the manifold of Banach space functions in general (i.e., without the normalization and positivity constraints associated with probability density functions). In particular, we give explicit expressions of Fisher metric and α-connections for an infinite-dimensional manifold $\mathcal{M}$ of suitably normed Banach space functions. This is achieved through first constructing some generalized expressions of divergence functions and then exploiting the coordinate-free version of the Eguchi relations (where covariant derivatives are denoted as ∇, ∇^*):

$$g(u,v) = -(d_u)_p\,(d_v)_q\,\mathcal{D}(p,q)\big|_{p=q}\,; \tag{6}$$

$$g(\nabla_w u, v) = -(d_w)_p\,(d_u)_p\,(d_v)_q\,\mathcal{D}(p,q)\big|_{p=q}\,; \tag{7}$$

$$g(u, \nabla^*_w v) = -(d_w)_q\,(d_v)_q\,(d_u)_p\,\mathcal{D}(p,q)\big|_{p=q}\,, \tag{8}$$

where d_u, d_v, d_w are directional derivatives along respective tangent directions denoted by $u, v, w \in \mathcal{T}_p(\mathcal{M})$. As an additional consequence of our approach, we clarify two different senses of duality in information geometry, namely, a duality related to the choice of the reference versus the comparison point on the manifold ("referential duality") and a duality related to the choice of a monotone function from a pair of conjugate ones to scale the probability density functions ("representational duality") — their meanings will be made precise in the following.

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

2.1. Fundamental convex inequality and divergence. Recall the notion of a (strictly) convex function $f : \mathbf{R} \to \mathbf{R}$ defined by:

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

for all $\gamma, \delta \in \mathbf{R}$ satisfying $\gamma \neq \delta$ and any $\alpha \in (-1, 1)$, with equality replacing the inequality when $\gamma = \delta$. Treating γ, δ as the values of two functions $p, q : \mathcal{X} \to \mathbf{R}$ evaluated at any particular sample point $\zeta \in \mathcal{X}$, i.e., $\gamma = p(\zeta)$, $\delta = q(\zeta)$, allows us to define the following family of divergence functionals on the set $\mathcal{B}_{f,\rho}$ of ζ-functions (here a ζ-function is one mapping $\mathcal{X} \to \mathbf{R}$, and $\rho : \mathbf{R} \to \mathbf{R}$ is strictly increasing)

$$\mathcal{B}_{f,\rho} = \{p(\zeta) : \mathrm{E}_\mu\{f(\rho(p))\} < \infty\}.$$

Lemma 1. *Let f, ρ be two functions as defined above. For any two ζ-functions $p, q : \mathcal{X} \to \mathbf{R}$ and any $\alpha \in \mathbf{R}$,*

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

is non-negative and equals zero if and only $p(\zeta) = q(\zeta)$ almost surely.

The family (parameterized by α) of divergence functional $\mathcal{D}^{(\alpha)}$, in which representational duality is embodied as $\mathcal{D}^{(\alpha)}_{f,\rho}(p,q) = \mathcal{D}^{(-\alpha)}_{f,\rho}(q,p)$, was first introduced in [7]. The function ρ is invoked to implement the notion of conjugate-scaled representations, see the next subsection.

2.2. Conjugate-scaled representations of measurable functions. Recall that for a strictly convex function $f : \mathbf{R} \to \mathbf{R}$, the Fenchel conjugate $f^* : \mathbf{R} \to \mathbf{R}$ is given by

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

with $(f^*)' = (f')^{-1}$. Now we introduce the notion of ρ-representation of a ζ-function $p(\cdot)$ as a mapping $p \mapsto \rho(p)$ using a strictly increasing function $\rho : \mathbf{R} \to \mathbf{R}$. 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, strictly convex function $f : \mathbf{R} \to \mathbf{R}$ if

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

(Here and below, $A \longleftrightarrow B$ is taken to mean "identity A, or equivalently identity B holds.") For example, take the α-embedding function $l^{(\alpha)}$ defined as

$$l^{(\alpha)}(t) = \begin{cases} \log t & \alpha = 1 \\ t & \alpha = -1 \\ \frac{2}{1-\alpha} t^{\frac{1-\alpha}{2}} & \alpha \neq \pm 1 \end{cases} \tag{9}$$

and introduce an auxiliary strictly convex function $\lambda^{(\alpha)}$ defined by

$$\lambda^{(\alpha)}(t) = \begin{cases} e^t & \alpha = 1 \\ t \log t - t & \alpha = -1 \\ \frac{2}{1+\alpha} \left(\frac{1-\alpha}{2} t\right)^{\frac{2}{1-\alpha}} & \alpha \neq \pm 1 \end{cases}.$$

Then, with respect to $f(t) = \lambda^{(\alpha)}(t) \longleftrightarrow f^*(t) = \lambda^{(-\alpha)}(t)$, the scaling functions $\rho(p) = l^{(\alpha)}(p) \longleftrightarrow \tau(p) = l^{(-\alpha)}(p)$ form a conjugate pair; this will be called the *canonical α-scaling*. In this case, the divergence functional $\mathcal{D}^{(\alpha)}_{f,\rho}(p,q)$ or $\mathcal{D}^{(\alpha)}_{f^*,\tau}(p,q)$ is homogeneous in p, q, since

$$f(\rho(p)) = \begin{cases} p & \alpha = 1 \\ p\log p - p & \alpha = -1 \\ \frac{2}{1+\alpha} p & \alpha \neq \pm 1 \end{cases} \longleftrightarrow f^*(\tau(p)) = \begin{cases} p\log p - p & \alpha = 1 \\ p & \alpha = -1 \\ \frac{2}{1-\alpha} p & \alpha \neq \pm 1 \end{cases} .$$

2.3. Geometry induced by the $\mathcal{D}^{(\alpha)}$-divergence. We assume that the set $B_{f,\rho}$ of ζ-functions, under a suitable topology, forms a manifold $\mathcal{M}$ of infinite dimensions. A point x_0 on $\mathcal{M}$ is a specific ζ-function $p : \zeta \mapsto p(\zeta)$ defined for all $\zeta \in \mathcal{X}$. Any function $p \to \mathrm{F}(p)$ on $\mathcal{M}$ is referred to as a *ζ-functional*, because it takes in a ζ-function p and outputs a number. The set of smooth ζ-functionals on $\mathcal{M}$ is denoted as $\mathcal{F}(\mathcal{M})$. A curve on $\mathcal{M}$ passing through a 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", that is, $p(\zeta|t)$ is a ζ-function "parameterized" by t — a one-parameter family of ζ-functions is formed as t varies. More generally, $p(\zeta|\theta)$, where $\theta = [\theta^1, \cdots, \theta^n] \in \Theta \subseteq \mathbf{R}^n$, is a ζ-function indexed by n parameters $\theta^1, \cdots, \theta^n$. As θ varies, $p(\zeta|\theta)$ represents an embedding $\widetilde{\mathcal{M}}_\theta \subset \mathcal{M}$ where

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

They are referred to as parametric models (and parametric statistical model if $p(\zeta|\theta)$ is normalized and positively valued).

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. Consider the ζ-functional in the following form:

$$\mathrm{F}(p) = \int_{\mathcal{X}} f(p(\zeta))\, d\mu = \mathrm{E}_\mu\{f(p)\} .$$

When the tangent vector v operates on $\mathrm{F}(p)$, utilizing the expansion $p(\zeta|t) = p(\zeta) + v(\zeta)t + o(t^2)$, we have

$$d_v(\mathrm{F}(p)) = \lim_{t\to 0} \frac{\mathrm{F}(p(\zeta|t)) - \mathrm{F}(p(\zeta|0))}{t} = \int_{\mathcal{X}} f'(p(\zeta))\, v(\zeta)\, d\mu = \mathrm{E}_\mu\{f'(p)\, v\} ,$$

the outcome being another ζ-functional of both $p(\zeta)$ and $v(\zeta)$, and linear in the latter.

A vector field in the infinite-dimensional setting, as a cross-section of $\mathcal{T}M$, takes in a ζ-function and outputs a ζ-function. We denote a vector field as $u(\zeta|p) \in \Sigma(\mathcal{M})$, where the variable following the "|" sign indicates the dependency on the point $p(\zeta)$, an element of the base manifold $\mathcal{M}$. Take, for example, $u(\zeta|p) = \rho(p(\zeta))$ as the vector field. Its directional derivative $d_v u$ in the direction of v (which is a ζ-function) at the base point p (which is another ζ-function) is

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

Note that $d_v u$ is another ζ-function; that is why we can write $d_v u(\zeta|p)$ also as $(d_v u)(\zeta)$. 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)

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

When operating on a vector field, say $u(\zeta|p)$, ∇_w is defined as (see [8])

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

where $\mathrm{B} : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \Sigma(\mathcal{M})$ is a tensor field which is bilinear in the two vector fields w and u (which are p-dependent ζ-functions); it is the infinite-dimensional counterpart of the Christoffel symbol Γ (used for finite dimensions).

With the above preparations, we now apply (6)–(8) to $\mathcal{D}^{(\alpha)}_{f,\rho}$.

Theorem 1. *At any given $p \in \mathcal{M}$ and for any vector fields $u, v \in \Sigma(\mathcal{M})$, the induced metric tensor field $g : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \mathcal{F}(\mathcal{M})$ and the induced covariant derivatives: $\nabla, \nabla^* : \Sigma(\mathcal{M}) \times \Sigma(\mathcal{M}) \to \Sigma(\mathcal{M})$ are*

$$\begin{aligned} g(u,v) &= E_\mu\{g(p(\zeta))\, u(\zeta|p)\, v(\zeta|p)\}\,; \\ \nabla^{(\alpha)}_w u &= (d_w u)(\zeta) + B^{(\alpha)}(p(\zeta)) u(\zeta|p) w(\zeta|p)\,, \end{aligned}$$

where

$$\begin{aligned} g(t) &= f''(\rho(t))(\rho'(t))^2 = \rho'(t)\,\tau'(t)\,; \\ B^{(\alpha)}(t) &= \frac{1-\alpha}{2}\frac{f'''(\rho(t))\rho'(t)}{f''(\rho(t))} + \frac{\rho''(t)}{\rho'(t)} = \frac{d}{dt}\left(\frac{1+\alpha}{2}\log\rho'(t) + \frac{1-\alpha}{2}\log\tau'(t)\right). \end{aligned}$$

Furthermore, $\nabla^{*(\alpha)}_w u = \nabla^{(-\alpha)}_w u$.

Note that the $g(p)$ term and the $B^{(\alpha)}(p)$ term depend on p, the point on the base manifold, at which the metric and covariant derivatives are evaluated.

It is immediately evident from Theorem 1 that the Riemannian metric induced from $\mathcal{D}^{(\alpha)}_{f,\rho}(p,q)$ and the one induced from $\mathcal{D}^{(\alpha)}_{f^*,\tau}(p,q)$ are the one and the same for all α values, while the connections (covariant derivatives) induced from these two families of divergence are mutually conjugate in the sense of $\alpha \leftrightarrow -\alpha$. (Here and below, $a \leftrightarrow b$ indicates the exchange of the two variables a and b.) This implies that the conjugacy in the pair of connections is related to both referential duality (duality related to $p \leftrightarrow q$ in the expression of divergence) and representational duality (duality related to the $\rho \leftrightarrow \tau$ in the induced geometry).

With respect to the α-family of covariant derivatives $\nabla^{(\alpha)}$, it can be shown that (i) the Riemann curvature tensor $R^{(\alpha)}(u,v,w) \equiv 0$; (ii) the torsion tensor $T^{(\alpha)}(u,v) \equiv 0$. In other words, the ambient manifold $\mathcal{M}$ has zero-curvature and zero-torsion for all α. As such, any curvature on the manifold $\mathcal{M}_\mu$ of non-parametric probability density functions or the manifold $\mathcal{M}_\theta$ of parameterized densities may be interpreted as arising from embedding of or restriction to a lower dimensional space.

2.4. Canonical divergence. When $\lim_{\alpha\to\pm1} \mathcal{D}^{(\alpha)}_{f,\rho}(p,q)$ or $\lim_{\alpha\to\pm1} \mathcal{D}^{(\alpha)}_{f^*,\tau}(p,q)$,

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

The canonical divergence functional $\mathcal{A} : \mathcal{M} \times \mathcal{M} \to \mathbf{R}_+$ is defined as

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

such that

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

We remark that under canonical α-scaling, $\mathcal{A}_f$ is simply the α-divergence proper $\mathcal{A}^{(\alpha)}$: $\mathcal{A}_f(\rho(p),\tau(q)) = \mathcal{A}^{(\alpha)}(p,q)$.

2.5. Finite-dimensional parametric models. The manifold $\widetilde{\mathcal{M}}_\theta$ of parametric models is indexed by $\theta = [\theta^1, \cdots, \theta^n] \in \Theta \subseteq \mathbf{R}^n$, called the *natural parameter*. The tangent vector fields u, v, w of $\mathcal{M}$ in the directions that are also tangent for $\widetilde{\mathcal{M}}_\theta$ (or $\mathcal{M}_\theta$) take the form

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

For convenience, we denote $\rho_p \equiv \rho(p(\zeta|\theta))$, $\tau_p \equiv \tau(p(\zeta|\theta))$. The divergence functional then becomes a divergence function on $\Theta \times \Theta$.

Theorem 2. *The metric tensor and the affine connections on the manifold $\widetilde{\mathcal{M}}_\theta$ of parametric models take the form*

$$\begin{aligned}
(10) \quad g_{ij}(\theta) &= E_\mu\left\{f''(\rho_p)\frac{\partial\rho_p}{\partial\theta^i}\frac{\partial\rho_p}{\partial\theta^j}\right\} = E_\mu\left\{\frac{\partial\rho_p}{\partial\theta^i}\frac{\partial\tau_p}{\partial\theta^j}\right\}\\
(11) \quad \Gamma^{(\alpha)}_{ij,k}(\theta) &= E_\mu\left\{\left(\frac{1-\alpha}{2}f'''(\rho_p)\frac{\partial\rho_p}{\partial\theta^i}\frac{\partial\rho_p}{\partial\theta^j} + f''(\rho_p)\frac{\partial^2\rho_p}{\partial\theta^i\partial\theta^j}\right)\frac{\partial\rho_p}{\partial\theta^k}\right\},\\
&= E_\mu\left\{\frac{1-\alpha}{2}\frac{\partial^2\tau_p}{\partial\theta^i\partial\theta^j}\frac{\partial\rho_p}{\partial\theta^k} + \frac{1+\alpha}{2}\frac{\partial^2\rho_p}{\partial\theta^i\partial\theta^j}\frac{\partial\tau_p}{\partial\theta^k}\right\},
\end{aligned}$$

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

Just as in the infinite-dimensional case, if we construct the divergence function $\mathcal{D}^{(\alpha)}_{f^*,\tau}(\theta_p, \theta_q)$ on $\Theta \times \Theta$, then the induced metric will be the same as the g given by (10), while the induced connection will be conjugate to the one given by (11), i.e., $\Gamma^{*(\alpha)} = \Gamma^{(-\alpha)}$ — in other words, $\Gamma \leftrightarrow \Gamma^*$ reflects, in addition to referential duality $\theta_p \leftrightarrow \theta_q$, representational duality between ρ-scaling and τ-scaling of a ζ-function $\rho_p \leftrightarrow \tau_p$.

3. Two Special Cases and Their Induced Geometries

3.1. Case I: Homogeneous (α, β)-divergence. Under the canonical α-scaling (see Section 2.2) but expressed using the symbol β, the divergence functional becomes a two-parameter family

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

where $(\alpha, \beta) \in [-1, 1] \times [-1, 1]$. This homogeneous divergence (invariant against a change of background measure) is called the (α, β)-*divergence*; it belongs to the general class of f-divergence studied by [9]. Note that the α parameter encodes referential duality, and the β parameter encodes representational duality. When *either* $\alpha = \pm 1$ *or* $\beta = 1$, the one-parameter version of the generic alpha-connection results. The family $\mathcal{D}^{(\alpha,\beta)}$ is then a generalization of the α-divergence (1) and the Jensen difference (2) with

$$\begin{aligned}
\lim_{\alpha\to-1}\mathcal{D}^{(\alpha,\beta)}(p,q) &= \mathcal{A}^{(-\beta)}(p,q), \quad \lim_{\alpha\to1}\mathcal{D}^{(\alpha,\beta)}(p,q) = \mathcal{A}^{(\beta)}(p,q),\\
\lim_{\beta\to1}\mathcal{D}^{(\alpha,\beta)}(p,q) &= \mathcal{A}^{(\alpha)}(p,q), \quad \lim_{\beta\to-1}\mathcal{D}^{(\alpha,\beta)}(p,q) = \mathcal{J}^{(\alpha)}(p,q).
\end{aligned}$$

Theorem 3. *The metric g and covariant derivatives $\nabla^{(\alpha,\beta)}$ associated with the (α, β)-divergence are given by*

$$\begin{aligned}
g(u,v) &= E_\mu\left\{\frac{1}{p}uv\right\},\\
\nabla^{(\alpha,\beta)}_u v &= d_u v - \frac{1+\alpha\beta}{2p}uv,
\end{aligned}$$

with $\nabla^{*(\alpha,\beta)} = \nabla^{(-\alpha,\beta)} = \nabla^{(\alpha,-\beta)}$, *and* $u, v \in \Sigma(\mathcal{M})$. *Their parametric counterparts are*

$$\begin{aligned} g_{ij}(\theta) &= E_\mu \left\{ p \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right\} \\ \Gamma^{(\alpha,\beta)}_{ij,k}(\theta) &= E_\mu \left\{ \left(\frac{\partial^2 \log p}{\partial \theta^i \partial \theta^j} + \frac{1-\alpha\beta}{2} \frac{\partial \log p}{\partial \theta^i} \frac{\partial \log p}{\partial \theta^j} \right) \frac{\partial p}{\partial \theta^k} \right\} . \end{aligned}$$

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

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; setting $\lim_{\beta \to 1} \nabla^{(\alpha,\beta)}$ or $\lim_{\alpha \to 1} \nabla^{(\alpha,\beta)}$ yields the one-parameter family of α-connections. Note, due to (11), the two ways to reduce to the alpha-connections (indexed by β here to avoid confusion): (i) take $\alpha = 1$, and $\rho(p) = l^{(\beta)}(p)$ and $\tau(p) = l^{(-\beta)}(p)$; or (ii) take $\alpha = -1$, and $\rho(p) = l^{(-\beta)}(p)$ and $\tau(p) = l^{(\beta)}(p)$.

3.2. **Case II: 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 sample space $\mathcal{X} \ni \zeta$ such that

$$\rho(p(\zeta|\theta)) = \sum_i \theta^i \lambda_i(\zeta) ;$$

here the parameter $\theta = [\theta^1, \cdots, \theta^n] \in \Theta$ is its natural parameter, and the functions $\lambda_1(\zeta), \cdots, \lambda_n(\zeta)$ are the affine basis functions.

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

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

forms a vector $\eta = [\eta_1, \cdots, \eta_n] \in \Xi \subseteq \mathbf{R}^n$; η is the expectation parameter of $p(\zeta)$.

The above notion of ρ-affinity is a generalization of α-affine manifolds [5], [10], defined as (recall 9)

$$l^{(\alpha)}(p) = \sum_i \theta^i \lambda_i(\zeta) ;$$

here ρ- and τ-representations are just $l^{(\alpha)}$ and $l^{(-\alpha)}$, respectively.

When a parametric model is ρ-affine, the function

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

can be shown to be strictly convex. Furthermore, define

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

then the function

$$\Phi^*(\eta) \equiv \tilde{\Phi}((\partial\Phi)^{-1}(\eta))$$

can be shown to be the Fenchel conjugate of $\Phi(\theta)$ (here ∂ is the gradient operator). The convex functions Φ, Φ^* form a pair of "potentials" to induce η, θ:

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

Theorem 4. *For* ρ-*affine manifold,*

(i) *The divergence functional $\mathcal{D}^{(\alpha)}_{f,\rho}(p,q)$ then takes the form of the divergence function $D^{(\alpha)}_{\Phi}(\theta_p,\theta_q)$ given by*

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

(ii) *The metric tensor g_{ij}, the affine connections $\Gamma^{(\alpha)}_{ij,k}$ and their Riemann curvature tensors $R^{(\alpha)}_{ij\mu\nu}$ take the forms*

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

(iii) *The manifold is equiaffine, with α-parallel volume form $\omega^{(\alpha)}$ given by*

$$\omega^{(\alpha)}=(\det\Phi_{ij})^{\frac{1-\alpha}{2}}\,.$$

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

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

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

Note that the expressions for $\Gamma^{(\alpha)}$ and $R^{(\alpha)}$ in the form of (ii) and for $\omega^{(\alpha)}$ in the form of (iii) were previously given, respectively, by [5] (p.106) and by [11], both for the exponential family (manifold). Here their applicability is generalized to any ρ-affine manifold.

A special case arises when $\alpha=\pm1$, where the connections are curvature-free $R^{(\pm1)}_{ij\mu\nu}(\theta)=0$. This is the well-studied "dually flat" parametric statistical manifold [5], [10], under which divergence functions have a unique, canonical form.

Theorem 5. *When $\alpha\to\pm1$, $D^{(\alpha)}_{\Phi}$ reduces to the Bregman divergence B_{Φ}*

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

or equivalently, to the canonical divergence functions

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

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product of two vectors.

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

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

The biduality is compactly reflected in the canonical divergence as

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

4. Summary and Future Directions

This paper constructs a family of divergence functionals, induced by a smooth and strictly convex function, to measure the asymmetric "distance" between two functions defined on the sample space. Subject to an arbitrary monotone scaling, any such divergence functional induces on the Riemannian manifold of nonparameterized functions a metric tensor generalizing the conventional Fisher information metric and a pair of conjugate connections (covariant derivatives) generalizing the conventional ($\pm\alpha$)-connections (note that compared with [12] and [13], the current representation is more explicit). 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 alpha-divergence proper, 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 conjugate potentials exist to induce the natural and expectation parameters as biorthogonal coordinates on the manifold.

Our approach demonstrates 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 inequality defining a convex function $f : \mathbf{R} \to \mathbf{R}$ (or $\Phi : \mathbf{R}^n \to \mathbf{R}$), 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!

It should be noted that, while any divergence function determines uniquely a statistical manifold, 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 [14], it is not unique in general, except when the connections are dually flat (traditionally, $\alpha = \pm1$), for which canonical divergence is uniquely determined. In this sense, there is nothing special about our use of $\mathcal{D}^{(\alpha)}$-divergence apart from it being a generalization of familiar divergence families (including α-divergence in particular). Rather, $\mathcal{D}^{(\alpha)}$-divergence is a vehicle for us to derive the underlying Riemannian geometry with dual connections. 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 — our generalizations of t he Fisher metric and of conjugate α-connections hinge on this miraculous identification.

References

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

[2] C. R. Rao *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, pp. 217-240, Hayward, CA, 1987.

[3] S. Eguchi *Second order efficiency of minimum contrast estimators in a curved exponential family.* Annals of Statistics, 11: 793-803, 1983.

[4] H. Nagaoka and S. Amari *Differential geometry of smooth families of probability distributions.* Mathematical Engineering Technical Report (METR) 82-07, University of Tokyo, 1982.

[5] S. Amari *Differential Geometric Methods in Statistics.* Lecture Notes in Statistics, 28, Springer-Verlag, New York, 1985 (Reprinted in 1990).

[6] S. Lauritzen *Conjugate connections in statistical theory.* In C.T.J. Dodson (Ed.) *Proc. Workshop on Geometrization of Statistical Theory,* pp. 33-51, Univ. of Lancaster, 1987.

[7] J. Zhang *Divergence function, duality, and convex analysis.* Neural Computation, 16: 159-195, 2004

[8] S. Lang *Differential and Riemannian Manifolds.* Springer-Verlag, New York, 1995.

[9] I. Csiszàr *On topical properties of f-divergence.* Studia Mathematicarum Hungarica, 2: 329-339, 1967.

[10] S. Amari and H. Nagaoka *Method of Information Geometry.* AMS monograph, Oxford University Press, 2000.

[11] J. Takeuchi and S. Amari *α-Parallel prior and its properties.* IEEE Transactions on Information Theory, 51: 1011-1023, 2005.

[12] P. Gibilisco and G. Pistone *Connections on non-parametric statistical manifolds by Olicz space geometry.* Infinite Dimensional Analysis, Quantum Probability and Related Topics, 1: 325-347, 1998.

[13] M. Grasselli *Dual connections in nonparametric classical information geometry.* Annals of the Institute for Statistical Mathematics (to appear).

[14] T. Matumoto *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, 1993.

UNIVERSITY OF MICHIGAN,, 525 EAST UNIVERSITY, ANN ARBOR, MICHIGAN, 48109-1109, USA
E-mail address: junz@umich.edu