Reference Duality and Representation Duality in Information Geometry

Jun Zhang

Department of Psychology and Department of Mathematics
University of Michigan, Ann Arbor, Michigan, USA
junz@umich.edu

Abstract. Classical information geometry prescribes, on the parametric family of probability functions $\mathscr{M}_\theta$: (i) a Riemannian metric given by the Fisher information; (ii) a pair of dual connections (giving rise to the family of α-connections) that preserve the metric under parallel transport by their joint actions; and (iii) a family of (non-symmetric) divergence functions (α-divergence) defined on $\mathscr{M}_\theta \times \mathscr{M}_\theta$, which induce the metric and the dual connections. The role of α parameter, as used in α-connection and in α-embedding, is not commonly differentiated. For instance, the case with $\alpha = \pm 1$ may refer either to dually-flat (*e*- or *m*-) connections or to exponential and mixture families of density functions. Here we illuminate that there are two distinct types of duality in information geometry, one concerning the referential status of a point (probability function, normalized or denormalized) expressed in the divergence function ("reference duality") and the other concerning the representation of probability functions under an arbitrary monotone scaling ("representation duality"). They correspond to, respectively, using α as a mixture parameter for constructing divergence functions or as a power exponent parameter for monotone embedding of probability functions. These two dualities are coupled into referential-representational biduality for manifolds of denormalized probability functions with α-Hessian structure (i.e, transitively flat α-geometry) and for manifolds induced from homogeneous divergence functions with (α, β)-parameters but one-parameter family of $(\alpha \cdot \beta)$-connections.

Keywords: divergence function, embedding function, metric, affine connection, Fisher information, alpha-connection
PACS: 02.40.Hw; 89.79.Cf; 87.19.Io

INTRODUCTION

Information geometry is, narrowly speaking, the differential geometric study of the manifold of probability measures or probability density functions [3]. Let $(\mathscr{X}, \mu)$ be a measure space with σ-algebra built upon the atoms, $d\zeta$, of $\mathscr{X}$. Let $\widetilde{\mathscr{M}_\mu}$ and $\mathscr{M}_\mu$ denote, respectively, the space of denormalized and normalized probability functions, $p : \mathscr{X} \to \mathbb{R}_+ (\equiv \mathbb{R}^+ \cup \{0\})$, defined on the sample space, $\mathscr{X}$, with background measure $d\mu = \mu(d\zeta)$:

$$\widetilde{\mathscr{M}_\mu} = \{p(\zeta) : p(\zeta) > 0, \forall \zeta \in \mathscr{X}\}, \;\; \mathscr{M}_\mu = \left\{p(\zeta) \in \tilde{\mathscr{M}}, \mathrm{E}_\mu\{p(\zeta)\} = 1\right\} .$$

Here, and throughout this paper, $\mathrm{E}_\mu\{\cdot\} = \int_{\mathscr{X}} \{\cdot\}\, d\mu$ denotes the expectation of a measurable function (in curly brackets) with respect to the background measure, μ. We do not differentiate probability measures on discrete support or probability density functions on continuous support, and use the generic term of probability function.

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

$$\mathscr{M}_\theta = \{p(\zeta|\theta) \in \mathscr{M}_\mu : \theta \in \Theta \subset \mathbb{R}^n\} \subset \mathscr{M}_\mu \tag{1}$$

with the so-called Fisher-Rao metric g_{ij} as its Riemannian metric:

$$g_{ij}(\theta) = \mathrm{E}_\mu \left\{ p(\zeta|\theta) \frac{\partial \log p(\zeta|\theta)}{\partial \theta^i} \frac{\partial \log p(\zeta|\theta)}{\partial \theta^j} \right\} \tag{2}$$

and a family of α-connections with coefficients $\Gamma^{(\alpha)}$ ($\alpha \in \mathbb{R}$):

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \left(\frac{1-\alpha}{2} \frac{\partial \log p(\zeta|\theta)}{\partial \theta^i} \frac{\partial \log p(\zeta|\theta)}{\partial \theta^j} + \frac{\partial^2 \log p(\zeta|\theta)}{\partial \theta^i \partial \theta^j} \right) \frac{\partial p(\zeta|\theta)}{\partial \theta^k} \right\} . \tag{3}$$

Recall that, in general, a metric g is a bilinear map on the tangent space, and an affine connection Γ is used to define parallel transport of tangent vectors. Introducing "conjugacy" of a pair of connections, $\Gamma \longleftrightarrow \Gamma^*$, as defined by their jointly preserving the metric when each acts on one of the two tangent vectors; that is, when each tangent vector undergo parallel transport according to Γ or Γ^* respectively:

$$\frac{\partial g_{ij}}{\partial \theta^k} = \Gamma_{ki,j}(\theta) + \Gamma^*_{kj,i}(\theta) . \tag{4}$$

Any Riemannian manifold with its metric g and the family of connections $\Gamma^{(\alpha)}$ in the form of (2) and (3) is called α-structure, denoted by $\{\mathscr{M}_\theta, g, \Gamma^{(\pm\alpha)}\}$. Amari [1, 2] gave dualistic interpretation of $\alpha \longleftrightarrow -\alpha$ as conjugate connections on the manifold $\mathscr{M}_\theta$:

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

The so-called e-connection ($\alpha = 1$) vanishes (*i.e.*, its components becomes identically zero) on the manifold of the exponential family of probability functions under natural parameters, whereas the so-called m-connection ($\alpha = -1$) vanishes on the manifold of the mixture family of probability functions under mixture parameters; this is in addition to the fact that $\Gamma^{(\pm 1)}$ have zero curvatures for both exponential and mixture families under either natural or expectation parameterization. Such α-geometry has been extended non-parametric (infinite-dimensional) statistical manifolds [26, 29].

In a broader sense, a *statistical manifold* $\{\mathscr{M}, g, \Gamma, \Gamma^*\}$ is a differentiable manifold equipped with a Riemannian metric g and a pair of torsion-free connections $\Gamma \equiv \Gamma^{(1)}, \Gamma^* \equiv \Gamma^{(-1)}$ compatible with g in the sense of (4), without necessarily requiring g and Γ, Γ^* to take the forms of (2) and (3). Historically, the notion of dual (conjugate) connections appeared independently in the investigation of affine hypersurface immersion (see [17, 23]), where the compatibility of the metric structure and the affine connection structure generalizes that of Levi-Civita coupling. There, α-connections arise as convex mixture of the pair of conjugate connections [14].

It is known that the α-geometry $\{\mathscr{M}_\theta, g, \Gamma^{(\pm\alpha)}\}$ can be induced from a parametric family of divergence functions called "α-divergence" [2], which measure non-symmetric distance of any two points (probability functions) on the manifold $\mathscr{M}_\theta$. On the other hand, α-connections can also arise as the so-called α-embedding of parametric family of probability functions [2, 3], see Equations (7) and (8) in the next Section, where an α-affine family generalizes the notion of exponential and mixture families. Zhang [25, 26, 29] obtained further generalizations of the α-geometry for a pair of monotone embeddings (called ρ- and τ-embeddings there), with corresponding families of (denormalized) probability functions parameterized by natural and expectation parameters that are linked via Legendre–Fenchel transformation.

There have been several different usages of α-parameter in Amari's formulation of information geometry: (i) parameterizing the convex mixture of connections (α-connections); (ii) parameterizing the divergence functions (α-divergences); (iii) parameterizing monotone embedding of probability functions (α-embedding). Below, we carefully scrutinize these usages of α, by illuminating how convex mixing and monotone embedding interact in a divergence function and in the resulting structure of α-connections. Recapitulating [25**?** , 26, 29], it will be argued that there are two senses of duality in information geometry, a reference duality related to the reference/comparison status of a pair of points (functions) and a representation dualtiy related to monotone-scaled representations of them. We remark that such (bi)dualistic structure of the α-geometry is preserved in the infinite-dimensional setting as well [26, 29].

REPRESENTATION DUALITY: EMBEDDING AND DUAL PARAMETERIZATIONS

Monotone embedding

Recall that the α-embedding function [2] is defined as $l^{(\alpha)} : \mathbb{R}^+ \to \mathbb{R}$

$$l^{(\alpha)}(t) = \begin{cases} \log t & \alpha = 1 \\ \frac{2}{1-\alpha} t^{(1-\alpha)/2} & \alpha \neq 1 \end{cases} . \tag{6}$$

The α-embedding (or representation) of a probability function plays an important role in Tsallis statistics; see [16, 18, 4]. The Fisher-Rao metric and α-connections, under such α-representation, have the following expressions:

$$g_{ij}(\theta) = \mathrm{E}_\mu \left\{ \frac{\partial l^{(\alpha)}(p(\cdot|\theta))}{\partial \theta^i} \frac{\partial l^{(-\alpha)}(p(\cdot|\theta))}{\partial \theta^j} \right\} , \tag{7}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathrm{E}_\mu \left\{ \frac{\partial^2 l^{(\alpha)}(p(\cdot|\theta))}{\partial \theta^i \partial \theta^j} \frac{\partial l^{(-\alpha)}(p(\cdot|\theta))}{\partial \theta^k} \right\} . \tag{8}$$

The notion of α-embedding can be extended to monotone embedding in general [25]. Assume that $f : \mathbb{R} \to \mathbb{R}$ is a strictly convex function, with convex conjugate f^* given by:

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

The conjugate function f^*, being a function $\mathbb{R} \to \mathbb{R}$, is also strictly convex and satisfies $(f^*)^* = f$ and $(f^*)' = (f')^{-1}$. The Legendre–Fenchel inequality reads

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

For later application, the notion of conjugate monotone representations of (possibly denormalized) probability function $p(\cdot)$ is introduced [25]:

Definition 1. *Conjugate representations (Zhang, 2004). For a strictly increasing function $\rho : \mathbb{R} \to \mathbb{R}$, we call ρ-representation of probability function p the mapping $p \mapsto \rho(p)$. For a strictly increasing function $\tau : \mathbb{R} \to \mathbb{R}$, we say that τ-representation of the probability function, $p \mapsto \tau(p)$, is conjugate to ρ-representation with respect to a smooth and strictly convex function, $f : \mathbb{R} \to \mathbb{R}$, if:*

$$\tau(p) = f'(\rho(p)) = ((f^*)')^{-1}(\rho(p)) \longleftrightarrow \rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p))\,. \quad (9)$$

As an example, we may set $\rho(p) = l^{(\alpha)}(p)$ to be the α-*representation* given by Equation (6), and the conjugate representation is the $(-\alpha)$-representation $\tau(p) = l^{(-\alpha)}(p)$:

$$\rho(t) = l^{(\alpha)}(t) \longleftrightarrow \tau(p) = l^{(-\alpha)}(p) \quad (10)$$

In this case:

$$f(t) = \frac{2}{1+\alpha}\left(\left(\frac{1-\alpha}{2}\right)t\right)^{\frac{2}{1-\alpha}}\,, \quad f^*(t) = \frac{2}{1-\alpha}\left(\left(\frac{1+\alpha}{2}\right)t\right)^{\frac{2}{1+\alpha}} \quad (11)$$

so that

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

are both linear in p.

The motivation for introducing the auxiliary function f in the definition of conjugate representations will be clear in the discussion of divergence functions. For the moment, it suffices to note that strictly increasing functions from $\mathbb{R} \to \mathbb{R}$ form a group, with functional composition as group composition operation and the functional inverse as the group inverse operation. That is, (i) for any two strictly increasing functions, ρ_1, ρ_2, their functional composition $\rho_2 \circ \rho_1$ is strictly increasing; (ii) the functional inverse, ρ^{-1}, of any strictly increasing function, ρ, is also strictly increasing; (iii) there exists a strictly increasing function, ι, the identity function, such that $\rho \circ \rho^{-1} = \rho^{-1} \circ \rho = \iota$. From this perspective, $f' = \tau \circ \rho^{-1}$, $(f^*)' = \rho \circ \tau^{-1}$, encountered above, are themselves two mutually inverse, strictly increasing functions.

α-Geometry under monotone embedding

The parametric family of functions, $p(\zeta|\theta)$, forms a finite-dimensional manifold $\mathscr{M}_\theta$ with coordinates θ as *natural parameter* of the parametric model, see (1). The following

generalized α-geometry were derived in [25] based on an analysis of divergence functions and the geometries they induce. For convenience, we denote $\rho(\zeta,\theta) \equiv \rho(p(\zeta|\theta))$, $\tau(\zeta,\theta) \equiv \tau(p(\zeta|\theta))$.

Proposition 2. *(Zhang, 2004). For parametric models $p(\zeta|\theta) \in \mathscr{M}_\theta$, the metric tensor takes the form:*

$$g_{ij}(\theta) = E_\mu \left\{ f''(\rho(\zeta,\theta)) \frac{\partial \rho(\zeta,\theta)}{\partial \theta^i} \frac{\partial \rho(\zeta,\theta)}{\partial \theta^j} \right\} \tag{12}$$

and the α-connections take the form:

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\mu \left\{ \frac{1-\alpha}{2} f'''(\rho(\zeta,\theta)) A_{ijk} + f''(\rho(\zeta,\theta)) B_{ijk} \right\} \tag{13}$$

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = E_\mu \left\{ \frac{1+\alpha}{2} f'''(\rho(\zeta,\theta)) A_{ijk} + f''(\rho(\zeta,\theta)) B_{ijk} \right\} . \tag{14}$$

where:

$$A_{ijk}(\zeta,\theta) = \frac{\partial \rho(\zeta,\theta)}{\partial \theta^i} \frac{\partial \rho(\zeta,\theta)}{\partial \theta^j} \frac{\partial \rho(\zeta,\theta)}{\partial \theta^k} , \quad B_{ijk}(\zeta,\theta) = \frac{\partial^2 \rho(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(\zeta,\theta)}{\partial \theta^k} . \tag{15}$$

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

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 2 reduce to the Fisher information and α-connections of the exponential family in Equations (2) and (3).

With conjugare representations, the geometric quantities of Proposition 2 have dualistic expressions:

Corollary 3. *(Zhang, 2004). Under conjugate representations, the metric and α-connections of (12) - (15) are:*

$$g_{ij}(\theta) = E_\mu \left\{ \frac{\partial \rho(\zeta,\theta)}{\partial \theta^i} \frac{\partial \tau(\zeta,\theta)}{\partial \theta^j} \right\} \tag{16}$$

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial^2 \tau(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(\zeta,\theta)}{\partial \theta^k} + \frac{1+\alpha}{2} \frac{\partial^2 \rho(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \tau(\zeta,\theta)}{\partial \theta^k} \right\} \tag{17}$$

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = E_\mu \left\{ \frac{1+\alpha}{2} \frac{\partial^2 \tau(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \rho(\zeta,\theta)}{\partial \theta^k} + \frac{1-\alpha}{2} \frac{\partial^2 \rho(\zeta,\theta)}{\partial \theta^i \partial \theta^j} \frac{\partial \tau(\zeta,\theta)}{\partial \theta^k} \right\} \tag{18}$$

Affine submanifolds

Recall that an exponential family of probability functions is defined as:

$$p^{(e)}(\zeta|\theta) = \exp\left(F_0(\zeta) + \sum_i \theta^i F_i(\zeta) - \phi(\theta) \right) \tag{19}$$

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

$$\phi(\theta) = \log \mathrm{E}_\mu \left\{ \exp\left(F_0(\zeta) + \sum_i \theta^i F_i(\zeta) \right) \right\} . \tag{20}$$

On the other hand, the mixture family

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

can be viewed as a manifold charted by its mixture parameter θ satisfying $\sum_i \theta^i = 1$, along with the constraints $\int_X F_i(\zeta) d\mu = 1$. The exponential family and the mixture family are special cases of the α-family [2, 3] of probability functions, $p(\zeta|\theta)$, whose denormalization satisfies (with constant κ):

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

With respect to general monotone embedding, we have the notion of a ρ-affine family. A parametric model, $p(\zeta|\theta) \in \widetilde{\mathscr{M}_\mu}$, is said to be *$\rho$-affine* if its ρ-representation can be embedded into a finite-dimensional affine space

$$\rho(p(\zeta|\theta)) = \sum_i \theta^i F_i(\zeta) . \tag{21}$$

For any denormalized probability function, $p(\zeta)$, the projection of its τ-representation onto the functions $F_i(\zeta)$,

$$\eta_i = \int_{\mathscr{X}} \tau(p(\zeta))\, F_i(\zeta)\, d\mu \tag{22}$$

forms a vector $\eta = [\eta_1, \cdots, \eta_n] \subseteq \mathbb{R}^n$. We call η the expectation parameter of $p(\zeta)$, and the functions $F(\zeta) = [F_0(\zeta), F_1(\zeta), \cdots, F_n(\zeta)]$ the affine basis functions.

The above notion of ρ-affinity is a generalization of α-affine manifolds [2, 3], where ρ- and τ-representations are just α- and $(-\alpha)$-representations, respectively. Note that elements of the ρ-affine manifold may not be a probability model; rather, after denormalization, probability models can become ρ-affine.

Biorthogonality of natural and expectation parameters

Lemma 4. *(Zhang, 2004). When a parametric model is ρ-affine,*

(i) the function $\Phi(\theta)$ is strictly convex, where

$$\Phi(\theta) = \int_{\mathscr{X}} f(\rho(p(\zeta|\theta)))\, d\mu \; ; \tag{23}$$

(ii) define

$$\tilde{\Phi}(\theta) = \int_{\mathscr{X}} f^*(\tau(p(\zeta|\theta)))\, d\mu \tag{24}$$

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

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

$$\frac{\partial\Phi(\theta)}{\partial\theta^i} = \eta_i \longleftrightarrow \frac{\partial\Phi^*(\eta)}{\partial\eta_i} = \theta^i \tag{25}$$

where η *is given by (22).*

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

$$\Phi^*(\eta) = \langle \eta, (\partial\Phi)^{-1}(\eta)\rangle - \Phi((\partial\Phi)^{-1}(\eta)) \tag{26}$$

where $\partial\Phi$ stands for the gradient (sub-differential) of Φ, and $\langle\,,\,\rangle$ denotes the standard bilinear form. The function Φ^*, which is also convex, has Φ as its conjugate $(\Phi^*)^* = \Phi$. The Hessian (second derivatives) of a strictly convex function (Φ and Φ^*) is positive-definite. The Legendre–Fenchel inequality (26) 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 (25) is satisfied.

Note that while $\Phi(\theta)$ in Lemma 4 can be viewed as the generalized cumulant generating function (or partition function), $\Phi^*(\eta)$ is the generalized entropy function. It is important to realize that, while $f(\cdot)$ is strictly convex, $\mathscr{F}(p) = \int_{\mathscr{X}} f(p(\zeta|\theta))\, d\mu$ is not at all convex in θ in general, when p is not ρ-affine and does not satisfy (21).

Proposition 5. *(Zhang, 2004; Zhang, 2007; Zhang and Matsuzoe, 2009). The family of* ρ*-affine denormalized probability functions generates* α*-Hessian manifold, with*

(i) Riemannian metric tensor

$$g_{ij}(\theta) = \Phi_{ij}\,;$$

(ii) a family of affine connections

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

(iii) Riemann curvature tensor $R^{(\alpha)}$ *for* $\Gamma^{(\alpha)}$ *as*

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

(iv) equiaffine parallel volume form $\Omega^{(\alpha)}$ *for* $\Gamma^{(\alpha)}$ *as*

$$\Omega^{(\alpha)}(\theta) = \det[\Phi_{ij}(\theta)]^{\frac{1-\alpha}{2}}\,.$$

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

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

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

For ρ-affine family, the natural parameter, $\theta \in \Theta$ and the expectation parameter $\eta \in \Xi$ 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 $g_{ij}(\theta)$. Biorthogonal coordinates (being "good" coordinates on the manifold) leads to the vanishing of affine connection coefficients $\Gamma^{*(-1)}_{ij,k}(\eta) = 0$ or $\Gamma^{(1)}_{ij,k}(\theta) = 0$. This is the well-studied "dually flat" parametric statistical manifold [1, 2, 3], under which divergence functions have a unique, canonical form. When $\Gamma^{(\pm 1)}$ is dually flat, $\Gamma^{(\alpha)}$ is called "α-transitively flat" [24]. Therefore, the α-Hessian structure stated by Proposition 4 is the full α-geometry of the Hessian manifold, where $\alpha = \pm 1$ leads to the vanishing of the curvature tensor, $R^{(\pm 1)}_{ij\mu\nu}(\theta) = 0$.

REFERENCE DUALITY IN DIVERGENCE FUNCTIONS

Reference duality is to be understood in the context of divergence functions and the resulting statistical manifold they induce. In general, a divergence function (also called "contrast function") is non-negative for all p, q and vanishes only at its global minimum when $p = q$; it is assumed to be smooth, with vanishing first derivatives at those extremal points. In general, a divergence function is not symmetric, hence, demonstrating an effect of choosing either p or q as the reference point. For technical convenience, we also assume the divergence functions to be negative semi-definite for its mixed second derivatives at $p = q$.

Take, for instance, the *Kullback-Leibler divergence* between two probability densities, $p, q \in \mathscr{M}_\mu$, here expressed in its extended form (i.e., without requiring p and q to be normalized):

$$K(p,q) = \int \left(q - p - p \log \frac{q}{p} \right) d\mu , \tag{27}$$

$$K^*(p,q) = \int \left(p - q - q \log \frac{p}{q} \right) d\mu , \tag{28}$$

both expressions with a unique, global minimum of zero when $p = q$. Each version of the KL divergence is directed, meaning $K(p,q) \neq K(q,p); K^*(p,q) \neq K^*(q,p)$. The effect of exchanging the two points p, q (and hence demonstrating the reference duality) is seen as switching from K to K^*:

$$K(p,q) = K^*(q,p) .$$

Hence, K and K^* are, in this sense, "dual" divergence functions.

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

$$\mathscr{A}^{(\alpha)}(p,q) = \frac{4}{1-\alpha^2}\,\mathrm{E}_\mu\left\{\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q - p^{\frac{1-\alpha}{2}}\,q^{\frac{1+\alpha}{2}}\right\}, \tag{29}$$

with

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

It is easily seen that the α-divergence family contain dual divergence pairs, with reference duality $p \longleftrightarrow q$ reflected as $\alpha \longleftrightarrow -\alpha$ duality:

$$\mathscr{A}^{(\alpha)}(p,q) = \mathscr{A}^{(-\alpha)}(q,p)\,.$$

$\mathscr{D}^{(\alpha)}$-divergence

A general theory of divergence functions was proposed in Zhang (2004). Starting from a strictly convex functino $f:\mathbb{R}\to\mathbb{R}$ which, by definition, satisfies

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

for all $\gamma,\delta\in\mathbb{R}$, with equality holding, if and only if $\gamma=\delta$, for all $\alpha\in(-1,1)$. This *fundamental convex inequality* applies to any two real numbers, γ,δ. We can treat γ,δ as point evaluations, at any particular sample point, ζ, of two functions $p,q:\mathscr{X}\to\mathbb{R}$, $\gamma=p(\zeta)$, $\delta=q(\zeta)$. This allows us to define the following family of divergence functions.

Lemma 6. *(Zhang, 2004). Let $f:\mathbb{R}\to\mathbb{R}$ be smooth and strictly convex, and $\rho:\mathbb{R}\to\mathbb{R}$ be strictly increasing. For any two (possibly denormalized) probability functions, p,q, and any $\alpha\in\mathbb{R}$:*

$$\mathscr{D}^{(\alpha)}_{f,\rho}(p,q) = \frac{4}{1-\alpha^2}E_\mu\left\{\frac{1-\alpha}{2}f(\rho(p)) + \frac{1+\alpha}{2}f(\rho(q)) - f\left(\frac{1-\alpha}{2}\rho(p) + \frac{1+\alpha}{2}\rho(q)\right)\right\} \tag{30}$$

is non-negative and equals zero, if and only

$$p(\zeta) = q(\zeta) \ \textit{almost everywhere}.$$

Furthermore, evaluated at $p=q$, it has vanishing first derivatives and negative semi-definite mixed second derivatives. Here we require p,q to be elements of the set:

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

Lemma 6 constructed a family (parameterized by α) of divergence functionals, $\mathscr{D}^{(\alpha)}$, with reference duality embodied as:

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

Note that in the above construction, we retain the freedom of the ρ-embedding function (which can be taken to be the identity function if necessary). The reason ρ is introduced is for conjugate representations of probability functions (see the previous section). The two functions f and ρ allows the family of $\mathscr{D}^{(\alpha)}$-divergence to allow for dual divergence under conjugate representations, see the next section.

$\mathscr{D}^{(\alpha)}$-divergence, introduced in [25], generalizes many familiar divergence functions.

(a) α-divergence [2]: There are several ways $\mathscr{D}^{(\alpha)}$-divergence reduces to the familiar α-divergence (29) as a special case:
- Take $f(p) = e^p$ and $\rho(p) = \log p$. Then $\mathscr{D}^{(\alpha)}_{f,\rho}(p,q) = \mathscr{A}^{(\alpha)}(p,q)$.
- Take $\alpha = 1$, $\rho(p) = l^{(\beta)}(p)$ and $\tau(p) = l^{(-\beta)}(p)$, that is the alpha-representation (11) with (10) using β parameter. Then $\mathscr{D}^{(1)}_{f,\beta}(p,q) = \mathscr{A}^{(\beta)}(p,q)$.
- Take $\alpha = -1$, $\rho(p) = l^{(-\beta)}(p)$ and $\tau(p) = l^{(\beta)}(p)$, the minus-alpha-representation. Then $\mathscr{D}^{(-1)}_{f,\beta}(p,q) = \mathscr{A}^{(\beta)}(p,q)$.

(b) U-divergence [15]: It is defined with any strictly convex function $U : \mathbb{R} \to \mathbb{R}$:

$$\mathscr{D}_U(p,q) = \mathrm{E}_\mu\{U((U')^{-1}(q)) - U((U')^{-1}(p)) - p\cdot((U')^{-1}(q) - (U')^{-1}(p))\}\,.$$

That $\mathscr{D}^{(\alpha)}$-divergence includes U-divergence as a special case can be seen by taking $f(p) = U(p), \rho(p) = (U')^{-1}(p), \alpha \to 1$.

(c) β-divergence [6]: This family is defined as

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

It was well known that β-divergence is a special case of U-divergence, when taking

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

(d) (α,β)-divergence of Cichocki *et al.* [8]: the following two-parameter family was introduced

$$D^{\alpha,\beta}_{AB} = -\frac{1}{\alpha\beta}\,\mathrm{E}_\mu\left\{p^\alpha q^\beta - \frac{\alpha}{\alpha+\beta}p^{\alpha+\beta} - \frac{\beta}{\alpha+\beta}q^{\alpha+\beta}\right\} \tag{31}$$

and called (α,β)-divergence (which is different from the usage of the terminology in [25]). Essentially, it is α-divergence under β- (power) embedding:

$$D^{\alpha,\beta}_{AB} = (\alpha+\beta)^2\,\mathscr{A}^{\frac{\beta-\alpha}{\alpha-\beta}}(p^{\alpha+\beta}, q^{\alpha+\beta})\,.$$

Clearly, by taking $f(t) = e^t, \rho(t) = (\alpha+\beta)\log t$ and renaming $\frac{\beta-\alpha}{\alpha+\beta}$ as α, $D_{AB}^{\alpha,\beta}$ is a special case of $\mathscr{D}_{f,\rho}^{(\alpha)}(p,q)$.

(For the meaning of $\mathscr{A}$, see Equation (32) later). Note that the above divergence functions reduce to Kullback-Leibler divergence under appropriate choice of α, β.

Canonical divergence

When $\alpha \to \pm 1$, the divergence function $\mathscr{D}_{f,\rho}^{(\pm 1)}(p,q)$ takes the form:

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

where conjufate representations (9) are used. The canonical divergence function, $\mathscr{A} : \mathscr{M} \times \mathscr{M} \to \mathbb{R}_+$, is defined (with the aid of a pair of conjugate representations) as:

$$\mathscr{A}_f(\rho(p), \tau(q)) = \mathrm{E}_\mu\{f(\rho(p)) + f^*(\tau(q)) - \rho(p)\,\tau(q)\} \tag{32}$$

where $\int_{\mathscr{X}} f(\rho(p))d\mu$ can be called the (generalized) cumulant generating functional and $\int_{\mathscr{X}} f^*(\tau(p))d\mu$, the (generalized) entropy functional. Reference duality is reflected by:

$$\mathscr{A}_f(\rho(p), \tau(q)) \;=\; \mathscr{A}_{f^*}(\tau(q), \rho(p))\,.$$

Thus, switching reference $p \longleftrightarrow q$ can be achieved through $\alpha = 1 \longleftrightarrow \alpha = -1$ or through $(f, \rho) \longleftrightarrow (f^*, \tau)$:

$$\mathscr{D}_{f,\rho}^{(1)}(p,q) = \mathscr{D}_{f,\rho}^{(-1)}(q,p) \;=\; \mathscr{D}_{f^*,\tau}^{(1)}(q,p) \;=\; \mathscr{D}_{f^*,\tau}^{(-1)}(p,q)\,.$$

We can see that under conjugate $(\pm\alpha)$-representations (10), $\mathscr{A}_f$ is simply the α-divergence proper $\mathscr{A}^{(\alpha)}$:

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

In fact:

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

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

Divergence on ρ-affine family

For the exponential family (19), the expression (27) takes the form of the so-called *Bregman divergence* [7] defined on $\Theta \times \Theta \subseteq \mathbb{R}^n \times \mathbb{R}^n$:

$$\mathrm{B}_\phi(\theta_p, \theta_q) = \phi(\theta_p) - \phi(\theta_q) - \langle \theta_p - \theta_q, \partial\phi(\theta_q) \rangle \tag{33}$$

where ϕ is the potential function (20), ∂ is the gradient operator and $\langle \cdot, \cdot \rangle$ denotes the standard bilinear form.

In general, when p is ρ-affine, because of Lemma 4, $\mathscr{D}^{(\alpha)}_{f,\rho}(p,q)$ of (30) becomes

$$\mathscr{D}^{(\alpha)}_{f,\rho}(p,q) = \frac{4}{1-\alpha^2}\left\{\frac{1-\alpha}{2}\Phi(\theta_p) + \frac{1+\alpha}{2}\Phi(\theta_q) - \Phi\left(\frac{1-\alpha}{2}\theta_p + \frac{1+\alpha}{2}\theta_q\right)\right\} \equiv D^{(\alpha)}_\Phi \ .$$

When $\alpha \to \pm 1$, $\mathscr{D}^{(\alpha)}_{f,\rho}$ reduces to the Bregman divergence:

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

with reference duality revealed as

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

REFERENTIAL-REPRESENTATIONAL BIDUALITY IN α-GEOMETRY

Linking $\mathscr{D}^{(\alpha)}$- and $D^{(\alpha)}$-divergence to α-geometry

A divergence function $\mathscr{D}$ will induce a Riemannian metric, g by its second order properties and a pair of conjugate connections, Γ, Γ^* by its third order properties; these relations were first formulated by Eguchi [12, 13].

$$g_{ij} = -\frac{\partial}{\partial\theta^i_p}\frac{\partial}{\partial\theta^j_q}\mathscr{D}(p,q)\bigg|_{p=q} \ ,$$

$$\Gamma_{ij,k} = -\frac{\partial}{\partial\theta^i_p}\frac{\partial}{\partial\theta^j_p}\frac{\partial}{\partial\theta^k_q}\mathscr{D}(p,q)\bigg|_{p=q} \ , \ \Gamma^*_{ij,k} = -\frac{\partial}{\partial\theta^i_q}\frac{\partial}{\partial\theta^j_q}\frac{\partial}{\partial\theta^k_p}\mathscr{D}(p,q)\bigg|_{p=q} \ .$$

Applying these relations, we can show:

Proposition 7. *(Zhang, 2004).*

1. *The family of $\mathscr{D}^{(\alpha)}_{f,\rho}$-divergence induces the α-geometry given in Proposition 2 (and its dualistic expressions of Collorary 3);*

2. *The family of $D_\Phi^{(\alpha)}$-divergence induces the α-Hessian geometry given in Proposition 5.*

We remark that α-Hessian is a special kind of α-geometry; this parallels the fact that $D^{(\alpha)}$-divergence is a special case of $\mathscr{D}^{(\alpha)}$-divergence, where the probability functions are ρ-affine, and Φ is related to f,ρ via (23).

The link between a divergence function, which has the freedom of a choice of convex function for its construction and a choice of monotone function for embedding probability functions, and the resulting geometry, reveals the interaction between reference duality and representation duality. It is in this sense that we say the α-geometry reflect *referential-representational biduality.*

As an application, consider dualistic expressions of Corollary 3 and its inducing (dual) divergence functions. If we construct the divergence function, $\mathscr{D}^{(\alpha)}_{f^*,\tau}(\theta_p,\theta_q)$, then the induced metric, $\tilde{g}_{ij}$, and the induced conjugate connections, $\tilde{\Gamma}^{(\alpha)}_{ij,k},\tilde{\Gamma}^{*(\alpha)}_{ij,k}$, will be related to those induced from $\mathscr{D}^{(\alpha)}_{f,\rho}(\theta_p,\theta_q)$ (and denoted without the ˜) via:

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

with:

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

So, the difference between using $\mathscr{D}^{(\alpha)}_{f,\rho}(\theta_p,\theta_q)$ and $\mathscr{D}^{(\alpha)}_{f^*,\tau}(\theta_p,\theta_q)$ reflects a conjugacy in the ρ- and τ-representations of $p(\zeta|\theta)$. Corollary 3 says that the conjugacy in the connection pair $\Gamma \longleftrightarrow \Gamma^*$ reflects, in addition to the referential duality $\theta_p \longleftrightarrow \theta_q$, the representational duality between ρ-representation and τ-representation of a probability function:

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

Divergence from quasi-linear means

Recall the construction of $\mathscr{D}^{(\alpha)}_{f,\rho}$ in a previous section. If $f' = \tau\circ\rho^{-1}$ 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:

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

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

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

is the quasi-linear mean of two numbers γ, δ. Therefore, the following is also a divergence function

$$\frac{4}{1-\alpha^2}\mathrm{E}_\mu\left\{\tau^{-1}\left(\frac{1-\alpha}{2}\tau(p(\zeta)) + \frac{1+\alpha}{2}\tau(q(\zeta))\right) - \rho^{-1}\left(\frac{1-\alpha}{2}\rho(p(\zeta)) + \frac{1+\alpha}{2}\rho(q(\zeta))\right)\right\}.$$

Homogeneous (α,β)-divergence

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

$$\mathscr{D}_\rho^{(\alpha)}(p,q) = \frac{4\varepsilon}{1-\alpha^2}\int_{\mathscr{X}}\left\{\frac{1-\alpha}{2}p(\zeta) + \frac{1+\alpha}{2}q(\zeta) - \rho^{-1}\left(\frac{1-\alpha}{2}\rho(p(\zeta)) + \frac{1+\alpha}{2}\rho(q(\zeta))\right)\right\}d\mu \tag{34}$$

As an example, take $\rho(p) = \log p$, $\varepsilon = 1$; then $M_\rho^{(\alpha)}(p,q) = p^{\frac{1-\alpha}{2}}q^{\frac{1+\alpha}{2}}$, and $\mathscr{D}_\rho^{(\alpha)}(p,q)$ is the α-divergence (29), while

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

is an immediate generalization of the KL divergence in (27) and (28).

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

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

then (see [25]) $\rho(p) = l^{(\beta)}(p)$; so (34) becomes a two-parameter family

$$\mathscr{D}^{(\alpha,\beta)}(p,q) \equiv \frac{4}{1-\alpha^2}\frac{2}{1+\beta}\mathrm{E}_\mu\left\{\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q - \left(\frac{1-\alpha}{2}p^{\frac{1-\beta}{2}} + \frac{1+\alpha}{2}q^{\frac{1-\beta}{2}}\right)^{\frac{2}{1-\beta}}\right\} \tag{35}$$

Here $(\alpha,\beta) \in [-1,1]\times[-1,1]$, and $\varepsilon = 2/(1+\beta)$ in Equation (34) is chosen to make $\mathscr{D}^{(\alpha,\beta)}(p,q)$ well defined for $\beta = -1$. We call this family (α,β)-*divergence*[26] [1] ; it belongs to the general class of f-divergence studied by [11]. 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

[1] This usage of the term "(α,β)-divergence" is different from the later use by [8] who refer to another two-parametric family in the form of (31). See discussions in the previous subsection.

results. The family, $\mathscr{D}^{(\alpha,\beta)}$, is then a generalization of Amari's α-divergence (29) with:

$$\lim_{\alpha\to-1}\mathscr{D}^{(\alpha,\beta)}(p,q)=\mathscr{A}^{(-\beta)}(p,q)\;,\quad \lim_{\alpha\to1}\mathscr{D}^{(\alpha,\beta)}(p,q)=\mathscr{A}^{(\beta)}(p,q)\;,$$
$$\lim_{\beta\to1}\mathscr{D}^{(\alpha,\beta)}(p,q)=\mathscr{A}^{(\alpha)}(p,q)\;,\quad \lim_{\beta\to-1}\mathscr{D}^{(\alpha,\beta)}(p,q)=J^{(\alpha)}(p,q)$$

where $J^{(\alpha)}$ denotes the Jensen difference discussed by [20]:

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

$J^{(\alpha)}$ reduces to Kullback-Leibler divergence (27) when $\alpha\to\pm1$. Lastly, we note that $\mathscr{D}^{(\alpha,\beta)}$, when either α or β equals zero, leads to the Levi-Civita connection.

With respect to the geometry induced from the (α,β)-divergence of Equation (35), we have the following result.

Corollary 8. *(Zhang, 2004). The metric and affine connections for the parametric (α,β)-manifold are:*

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

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:

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

Setting $\lim_{\beta\to1}\Gamma^{(\alpha,\beta)}$ yields Amari's one-parameter family of α-connections.

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

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

with biduality compactly expressed as

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

DISCUSSIONS

Our analysis above illuminates two different types of duality in information geometry, one concerning the choice of a reference point (probability function, normalized or denormalized) in the divergence function ("reference duality") and the other concerning the choice of a monotone scale in representing probability functions ("representation duality"). To tease apart the two, we study the conjugate ρ- and τ-representations and the associated ρ-affine family of probability function. Our investigation demonstrated an intimate connection between convex analysis and information geometry. The divergence functions are associated with the fundamental inequality of a convex function, $f : \mathbb{R} \to \mathbb{R}$ (or $\Phi : \mathbb{R}^n \to \mathbb{R}$), with the convex mixture coefficient as the α-parameter in the induced geometry. Reference duality is associated with $\alpha \longleftrightarrow -\alpha$, and representation duality is associated with the convex conjugacy $f \longleftrightarrow f^*$ (or $\Phi \longleftrightarrow \Phi^*$). To illustrate these differences, we introduce the (α, β)-divergence, with bidualistic structure extending that of the α-divergence, with α and β representing reference duality and representation duality, respectively. Interestingly, the induced Fisher metric is independent of α, β while the induced alpha-connection uses $\alpha\beta$ as a single parameter.

The kind of reference duality (originating from non-symmetric status of a referent and a comparison object), while common in behavioral-psychological contexts [27], has always been implicitly acknowledged in statistics. Formal investigation of such non-symmetry between a reference and a comparison probability function leads to the framework of preferred point geometry [9, 10, 31, 32]. Preferred point geometry reformulates Amari's [1] expected geometry and Barndorff-Nelsen's [5] observed geometry by studying the product manifold $\mathscr{M}_{\theta_p} \times \mathscr{M}_{\theta_q}$ formed by an ordered pair of probability functions (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 reference duality awaits future exploration.

ACKNOWLEDGMENTS

The writing of this paper is supported by research grant ARO W911NF-12-1-0163.

REFERENCES

1. Amari, S. Differential geometry of curved exponential families—curvatures and information loss. *Ann. Stat.* **1982**, *10*, 357–385.
2. Amari, S. *Differential Geometric Methods in Statistics*, Lecture Notes in Statistics 28; Springer-Verlag: New York, NY, USA, 1985.
3. Amari, S.; Nagaoka, H. *Method of Information Geometry*; Oxford University Press: Oxford, UK, 2000.
4. Amari. S.; Ohara, A. Geometry of q-exponential family of probability distributions. *Entropy* **2011**, *13*, 1170–1185.
5. Barndorff-Nielsen, O.E. *Parametric Statistical Models and Likelihood*. Lecture Notes in Statistics, 50. Springer-Verlag, 1988.
6. Basu, A.; Harris, I.R.; Hjort, N.; Jones, M. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **1998**, *85*, 549âĂŞ559.

7. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Phys.* **1967**, *7*, 200–217.
8. Cichocki, A.; Cruces, S; Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
9. Critchley, F.; Marriott, P.; Salmon, M. Preferred point geometry and statistical manifolds. *The Annals of Statistics*, **1993**, *21*, 1197-1224.
10. Critchley, F.; Marriott, P.; Salmon, M. Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence. *The Annals of Statistics*, **1994**, *22*, 1587-1602.
11. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, **1967**, *2*, 229–318.
12. Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Stat.* **1983**, *11*, 793–803.
13. Eguchi, S. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Math. J.* **1985**, *15*, 341–391.
14. Lauritzen, S. Statistical manifolds. In *Differential Geometry in Statistical Inference*; Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., Rao, C.R., Eds.; IMS: Hayward, CA, USA, Volume 10, Lecture Notes, 1987, pp. 163–216.
15. Murata, N.; Takenouchi, T.; Kanamori, T., and Eguchi, S. Information geometry of U-Boost and Bregman divergence. *Neural Computation*, **2004**, *16*, 1437-1481.
16. J. Naudts. Generalised exponential families and associated entropy functions, *Entropy*, **2008**, *10*, 131-149.
17. Nomizu, K.; Sasaki, T. *Affine Differential Geometry—Geometry of Affine Immersions*; Cambridge University Press: Cambridge, MA, USA, 1994.
18. Ohara, A.; Matsuzoe, H.; Amari, S. A dually at structure on the space of escort distributions. *J. Phys. Conf. Ser.* **2010**, *201*, No. 012012.
19. Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
20. Rao, C.R. Differential Metrics in Probability Spaces. In *Differential Geometry in Statistical Inference*; Amari, S., Barndorff-Nielsen, O., Kass, R., Lauritzen, S., Rao, C.R., Eds.; IMS: Hayward, CA, USA, 1987; Voluem 10, Lecture Notes, pp. 217–240.
21. Shima, H. Compact locally Hessian manifolds. *Osaka J. Math.* **1978**, *15*, 509–513.
22. Shima, H.; Yagi, K. Geometry of Hessian manifolds. *Differ. Geom. Its Appl.* **1997**, *7*, 277–290.
23. Simon U.; Schwenk-Schellschmidt, A.; Viesel, H. *Introduction to the Affine Differential Geometry of Hypersurfaces*, Lecture Notes of the Science; University of Tokyo: Tokyo, Japan, 1991.
24. Uohashi, K. "On α-conformal equivalence of statistical submanifolds." Journal of Geometry, **2002**, *75*, 179-184.
25. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195.
26. Zhang, J. Referential Duality and Representational Duality on Statistical Manifolds. In *Proceedings of the Second International Symposium on Information Geometry and Its Applications* , Tokyo, Japan, 12–16 December 2005; pp. 58–67.
27. Zhang, J. Referential duality and representational duality in the scaling of multi-dimensional and infinite-dimensional stimulus space. In *Measurement and Representation of Sensations: Recent Progress in Psychological Theory*; Dzhafarov, E., Colonius, H., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2006.
28. Zhang J. A note on curvature of alpha-connections of a statistical manifold. *Ann. Inst. Stat. Math.* **2007**, *59*, 161–170.
29. Zhang, J. Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. *Entropy*, **2013**, *15*: 5384-5418.
30. Zhang, J.; Matsuzoe, H. Dualistic Differential Geometry Associated with a Convex Function. In *Advances in Applied Mathematics and Global Optimization*; Gao D.Y., Sherali, H.D., Eds.; Springer: New York, NY, USA, 2009; Voluem III, Chapter 13, pp. 439–466.
31. Zhu, H.-T.; Wei, B.-C. Some notes on preferred point α-geometry and α-divergence function. *Statistics and Probability Letters*, **1997**, *33*, 427-437.
32. Zhu, H.-T.; Wei, B.-C. Preferred point α-manifold and Amari's α-connections. *Statistics and Probability Letters*, **1997**, *36*, 219-229.