Divergence Function, Duality, and Convex Analysis

Jun Zhang
junz@umich.edu
Department of Psychology, University of Michigan, Ann Arbor, MI 48109, U. S. A

From a smooth, strictly convex function Φ: $\mathbf{R}^n \to \mathbf{R}$, a parametric family of divergence function $\mathcal{D}_\Phi^{(\alpha)}$ may be introduced:

$$\mathcal{D}_\Phi^{(\alpha)}(x,y) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}\,\Phi(x) + \frac{1+\alpha}{2}\,\Phi(y) - \Phi\left(\frac{1-\alpha}{2}x + \frac{1+\alpha}{2}y\right)\right)$$

for $x, y \in \operatorname{int} \operatorname{dom}(\Phi) \subset \mathbf{R}^n$, and for $\alpha \in \mathbf{R}$, with $\mathcal{D}_\Phi^{(\pm 1)}$ defined through taking the limit of α. Each member is shown to induce an α-independent Riemannian metric, as well as a pair of dual α-connections, which are generally nonflat, except for $\alpha = \pm 1$. In the latter case, $\mathcal{D}_\Phi^{(\pm 1)}$ reduces to the (nonparametric) Bregman divergence, which is representable using Φ and its convex conjugate Φ^* and becomes the canonical divergence for dually flat spaces (Amari, 1982, 1985; Amari & Nagaoka, 2000). This formulation based on convex analysis naturally extends the information-geometric interpretation of divergence functions (Eguchi, 1983) to allow the distinction between two different kinds of duality: referential duality ($\alpha \leftrightarrow -\alpha$) and representational duality ($\Phi \leftrightarrow \Phi^*$). When applied to (not necessarily normalized) probability densities, the concept of conjugated representations of densities is introduced, so that $\pm\alpha$-connections defined on probability densities embody both referential and representational duality and are hence themselves bidual. When restricted to a finite-dimensional affine submanifold, the natural parameters of a certain representation of densities and the expectation parameters under its conjugate representation form biorthogonal coordinates. The alpha representation (indexed by β now, $\beta \in [-1, 1]$) is shown to be the only measure-invariant representation. The resulting two-parameter family of divergence functionals $\mathcal{D}^{(\alpha,\beta)}$, $(\alpha, \beta) \in [-1, 1] \times [-1, 1]$ induces identical Fisher information but bidual alpha-connection pairs; it reduces in form to Amari's alpha-divergence family when $\alpha = \pm 1$ or when $\beta = 1$, but to the family of Jensen difference (Rao, 1987) when $\beta = -1$.

Neural Computation **16, 159–195** (2004)

1 Introduction

Divergence functions play an important role in many areas of neural computation like learning, optimization, estimation, and inference. Roughly, they measure the directed (asymmetric) difference of two probability density functions in the infinite-dimensional functional space, or two points in a finite-dimensional vector space that defines parameters of a statistical model. An example is the Kullback-Leibler (KL) divergence (cross-entropy) between two probability densities p and q, here expressed in its extended form (i.e., without requiring p, q to be normalized),

$$K(p, q) = \int \left(q - p - p \log \frac{q}{p} \right) d\mu = K^*(q, p), \tag{1.1}$$

which reaches the unique, global minimum value of zero on the diagonal of the product manifold (i.e., $p = q$). Many learning algorithms and/or proof for their convergence rely on properties of the KL divergence; the common ones are Boltzmann machine (Ackley, Hinton, & Sejnowski, 1985; Amari, 1991; Amari, Kurata, & Nagaoka 1992), the em algorithm and its comparison with EM algorithm (Amari, 1995), and the wake-sleep algorithm of the Helmholtz machine (Ikeda, Amari, & Nakahara, 1999).

Another class of divergence functions widely used in optimization and convex programming literature is the so-called Bregman divergence (see below). It plays an essential role in unifying the class of projection and alternating minimization algorithms (Lafferty, Della Pietra, & Della Pietra, 1997; Della Pietra, Della Pietra, & Lafferty, 2002; Bauschke & Combettes, 2002; Bauschke, Borwein, & Combettes, 2002). Parametric families of Bregman divergence were used in blind source separation (Mihoko & Eguchi, 2002) and for boosting machines (Lebanon & Lafferty, 2002; Eguchi, 2002).

Divergence function or functional[1] is an essential subject in information geometry, the differential geometric study of the manifold of (parametric or nonparametric) probability distributions (Amari, 1982, 1985; Eguchi, 1983, 1992; Amari & Nagaoka, 2000). As first demonstrated by Eguchi (1983), a well-defined divergence function (also called a contrast function) with vanishing first order (in the vicinity of $p = q$) term will induce a Riemannian metric g by its second-order properties and a pair of dual (also called conjugate) connections (Γ, Γ^*) by its third-order properties, where the dual connections jointly preserve the metric under parallel transport. A manifold

[1] Strictly speaking, when the underlying space is a finite-dimensional vector space, for example, that of parameters for a statistical or neural network model, then the term *function* is appropriate. However, if the underlying space is an infinite-dimensional function space, for example, that of nonparametric probability densities, then the term *functional* ought to be used. The latter implicitly defines a divergence function (through pullback) if the probability densities are embedded into a finite-dimensional space as a statistical model.

endowed with $\{g, \Gamma, \Gamma^*\}$ is known as a statistical manifold; conditions for its affine realization through its embedding into a higher-dimension space have been clarified (Kurose, 1994; Matsuzoe, 1998, 1999; Uohashi, Ohara, & Fujii, 2000).

1.1 Alpha-, Bregman, and Csiszar's f-Divergence and Their relations.

Amari (1982, 1985) introduced and investigated an important parametric family of divergence functionals, called *α-divergence*[2]

$$\mathcal{A}^{(\alpha)}(p, q) = \frac{4}{1-\alpha^2} \int \left(\frac{1-\alpha}{2} p + \frac{1+\alpha}{2} q - p^{\frac{1-\alpha}{2}} q^{\frac{1+\alpha}{2}} \right) d\mu, \quad \alpha \in \mathrm{R}. \tag{1.2}$$

The α-divergence, which specializes to $K(p, q)$ for $\alpha = -1$ and $K^*(p, q)$ for $\alpha = 1$ (by taking the limit of α), induces on the statistical manifold the family of α-connections (Chentsov, 1982; Amari, 1982). The duality between $\alpha \leftrightarrow -\alpha$ is reflected in that $\pm\alpha$-connections form a pair of dual connections that jointly preserve the metric and that an α-connection is flat if and only $(-\alpha)$-connection is flat (Amari, 1985; Lauritzen, 1987). As a special case, the exponential family ($\alpha = 1$) and the mixture family ($\alpha = -1$) of densities generate dually flat statistical manifolds.

Alpha divergence is a special case of the measure-invariant f-divergence introduced by Csiszár (1967), which is associated with any convex function f_c: $\mathrm{R}_+ \to \mathrm{R}_+$ satisfying $f_c(1) = 0$, $f_c'(1) = 0$:

$$\mathcal{F}_{f_c}(p, q) = \int p\, f_c\left(\frac{q}{p}\right) d\mu, \tag{1.3}$$

where $\mathrm{R}_+ \equiv \mathrm{R}^+ \cup \{0\}$. This can be seen by f_c taking the following family of convex functions[3] indexed by a parameter α,

$$f^{(\alpha)}(t) = \frac{4}{1-\alpha^2} \left(\frac{1-\alpha}{2} + \frac{1+\alpha}{2} t - t^{\frac{1+\alpha}{2}} \right), \quad \alpha \in \mathrm{R}. \tag{1.4}$$

[2] This form of α-divergence first appeared in Zhu and Rohwer (1995, 1997), where it was called the δ-divergence, $\delta = (1-\alpha)/2$. The term $\frac{1-\alpha}{2} p + \frac{1+\alpha}{2} q$ in equation 1.2 after integration, is simply 1 for normalized densities; this was how Amari (1982, 1985) introduced α-connection as a specific family of Csiszár's f-divergence. See note 3.

[3] Note that this form differs slightly with the function given in Amari (1985) and Amari and Nagaoka (2000) by the additional term $\frac{1-\alpha}{2} + \frac{1+\alpha}{2} t$. This term is needed for the form of α-divergence expressed in equation 1.2, which is "extended" from the original definition given in Amari (1982, 1985) to allow denormalized densities, in much the same way that extended Kullback-Leibler divergence (see equation 1.1) differs from its original form (without the $p - q$ or $q - p$ term). This additional term in $f^{(\alpha)}$ allows the condition $f^{(\alpha)}(1) = 0$ to be satisfied. It also provides a unified treatment for the $\alpha = \pm 1$ case, since $\lim_{\alpha \to 1} f^{(\alpha)}(t) = 1 - t + t \log t$, $\lim_{\alpha \to -1} f^{(\alpha)}(t) = t - 1 - \log t$.

Eguchi (1983) showed that any f-divergence induced a statistical manifold with a metric proportional to Fisher information with the constant of proportionality $f_c''(1)$ and an equivalent α-connection,

$$\alpha = 3 + 2f_c'''(1)/f_c''(1). \tag{1.5}$$

We note in passing that for a general, smooth, and strictly convex function f: R → R, we can always induce a measure-invariant divergence by using $f_c(t) = g(t)$ in equation 1.3, where

$$g(t) \equiv f(t) - f(1) - f'(1)(t-1). \tag{1.6}$$

That the right-hand side of the above is nonnegative can be easily proved by showing that $t = 1$ is a global minimum with $g(1) = g'(1) = 0$.

Another kind of divergence function, called Bregman divergence, is defined for any two points $x = [x^1, \ldots, x^n], y = [y^1, \ldots, y^n]$ in an n-dimensional vector space R^n (Bregman, 1967). It is induced by a smooth and strictly convex function Φ: $\mathrm{R}^n \to \mathrm{R}$:

$$\mathrm{B}_\Phi(x, y) = \Phi(y) - \Phi(x) - \langle y - x, \nabla\Phi(x)\rangle, \tag{1.7}$$

where ∇ is the gradient (or, more strictly, subdifferential $\partial\Phi$ if differentiability condition is removed) operator and $\langle\cdot,\cdot\rangle$ denotes the standard inner product of two vectors. It is also called (actually proportional to) *geometric divergence* (Kurose, 1994), proposed in the context of affine realization of a statistical manifold. Bregman divergence $\mathrm{B}_\Phi(x, y)$ specializes to the KL divergence upon setting $\Phi(x) = \sum_i e^{x^i}$, introducing new variables $x^i = \log p^i, y^i = \log q^i$, and changing $\int d\mu$ into $\sum_i$. More generally, as observed by Eguchi (2002), Csiszár's f-divergence (see equation 1.3) is naturally related (but not identical) to Bregman divergence (see equation 1.7), having taken $\Phi(x) = \sum_i f(x^i)$ with $y^i = q^i/p^i$ and $x^i = 1$. In this case (with a slight abuse of notation),

$$\mathcal{F}_f(p, q) = \sum_i p^i \,\mathrm{B}_f\left(\frac{q^i}{p^i}, 1\right).$$

It is now known (Kass & Vos, 1997) that Bregman divergence is essentially the canonical divergence (Amari & Nagaoka, 2000) on a dually flat manifold equipped with a pair of biorthogonal coordinates induced from a pair of "potential functions" under the Legendre transform (Amari, 1982, 1985). It is a distance-like measure on a finite-dimensional Riemannian manifold that is essentially flat and is very different from the α-divergence (see equation 1.2) that is defined over the space of positively valued, infinite-dimensional functions on sample space (i.e., positive measures) and is generally nonflat. However, if the positive measures p and q can be affinely embedded into some finite-dimensional submanifold, the Legendre potentials for α-divergence could exist. Technically, this corresponds to the so-called α-affine manifold, where the embedded α-representation of the densities ($\alpha \in \mathrm{R}$),

$$l^{(\alpha)}(p) = \begin{cases} \log p & \alpha = 1 \\ \frac{2}{1-\alpha}\, p^{(1-\alpha)/2} & \text{else}, \end{cases} \tag{1.8}$$

can be expressed as a linear combination of a countable set of basis functions of the infinite-dimensional functional space (the definition of α-affinity). If and only if such embedding is possible for a certain value of α, a potential function (and its dual) can be found so that equation 1.2 becomes 1.7. In general, Bregman divergence and α-divergence are very different in terms of both the dimensionality and the flatness of the underlying manifold that they are defined on, though both may induce dual connections.

Given the fundamental importance of α-connections in information geometry, it is natural to ask whether the parameter α may arise other than from the context of α-embedding of density functions. How is the $\alpha \leftrightarrow -\alpha$ duality related to the pair of biorthogonal coordinates and the canonical divergence they define? Does there exist an even more general expression of divergence functions that would include the α-divergence, the Bregman divergence, and the f-divergence as special cases yet would still induce the dual $\pm\alpha$-connections? The existence of a divergence function on a statistical manifold given the Riemannian metric and a pair of dual, torsion-free connections was answered positively by Matumoto (1993). Here, the interest is to find explicit forms for such general divergence functions, in particular, measure-invariant ones.

The goal of this article is to introduce a unifying perspective for the α-, Bregman, and Csiszar's f-divergence as arising from certain fundamental convex inequalities and to clarify two different senses of duality embodied by divergence function and functionals and the statistical manifolds they define: referential duality and representational duality.

Starting from the definition of a smooth, strictly convex function $\Phi: \mathrm{R}^n \rightarrow \mathrm{R}$, a parametric family of divergence $\mathcal{D}_\Phi^{(\alpha)}(x, y)$, $\alpha \in \mathrm{R}$, over points $x, y \in S = \operatorname{int}\operatorname{dom}(\Phi)$ can be introduced that will be shown to induce a single Riemannian metric with a parametric family of affine connections indexed by α, the convex mixture parameter. These α-connections are nonflat unless $\alpha = \pm 1$, when $\mathcal{D}_\Phi^{(\pm 1)}$ turns out to be the Bregman divergence, which can be cast into the canonical form using a pair of convex conjugate functions Φ and Φ^* (Amari's potential functions) that obey Legendre-Fenchel duality. The biorthogonal coordinates x and u, originally introduced for a dually flat manifold, can now be extended to define the divergence function on any nonflat manifold ($\alpha \neq \pm 1$) as well. A distinction is drawn between two kinds of duality ("biduality") of statistical manifolds, in the sense of mutual references $x \leftrightarrow y$, as reflected by $\alpha \leftrightarrow -\alpha$ duality, and in the sense of conjugate representations $u = \nabla\Phi(x) \leftrightarrow x = (\nabla\Phi^*)(u) = (\nabla\Phi)^{-1}(u)$, as reflected by $\Phi \leftrightarrow \Phi^*$ duality. In the infinite-dimensional case, representational duality is achieved through conjugate representations of any (not necessarily normalized) density function; here, conjugacy is with respect to a strictly convex function defined on the real line f: R → R. Our notion of conjugate representations of density functions, which generalizes the notion of alpha representation (see equation 1.8), proves to be useful in characterizing the affine embedding of a density function into a finite-dimensional submanifold; the natural and expectation parameters become the pair of biorthogonal coordinates, and this case completely reduces to the one discussed earlier.

Of particular practical importance is that our analysis provides a two-parameter family of measure-invariant divergence function(al) $\mathcal{D}^{(\alpha,\beta)}(p, q)$ under the alpha representation $l^{(\beta)}(p)$ of densities (indexed by β now, $\beta \in [-1, 1]$, with α reserved to index convex mixture), which induce identical Fisher information metric and a family of alpha connections where the product $\alpha\beta$ serves as the "alpha" parameter. The two indices themselves, $(\alpha, \beta) \in [-1, 1] \times [-1, 1]$, precisely express referential duality and representational duality, respectively. Interestingly, at the level of divergence functional, $\mathcal{D}^{(\alpha,\beta)}$ turns out to be the family of alpha divergence for $\alpha = \pm 1$ or for $\beta = 1$, and the family of "Jensen difference" (Rao, 1987) for $\beta = -1$. Thus, Kullback-Leibler divergence, the one-parameter family of α-divergence and of Jensen difference, and the two-parameter family of (α, β)-divergence form nested families of measure-invariant divergence function(al) associated with the same statistical manifold studied in classical information geometry.

2 Divergence on Finite-Dimensional Parameter Space

In this section, we consider the finite-dimensional vector space R^n, or a convex subset thereof, that defines the parameter of a probabilistic (e.g., neural network) model. The goal is to introduce, with the help of an arbitrary strictly convex function Φ: $\mathrm{R}^n \to \mathrm{R}$, a family of asymmetric, nonnegative measures between two points in such space, called divergence functions (see proposition 1) and, through which to induce a well-defined statistical manifold with a Riemannian metric and a pair of dual connections (see proposition 2). The procedure used for linking a divergence function(al) to the underlying statistical manifold is due to Eguchi (1983); our notion of referential duality is reflected in the construction of dual connections. The notion of representational duality is introduced through equation 2.15 and proposition 5, based on the convex conjugacy operation (see remark 2.3.1). Biduality is thus shown to be the fundamental property of a statistical manifold induced by the family of divergence functions based on a convex function Φ.

2.1 Convexity and Divergence Functions.

Consider the n-dimensional vector space (e.g., the parameter space in the case of parametric probability density functions or neural network models). A set $S \subseteq \mathrm{R}^n$ is called convex if for any two points $x = [x^1, \ldots, x^n] \in S$, $y = [y^1, \ldots, y^n] \in S$ and any real number $\alpha \in [-1, 1]$, the convex mixture

$$\frac{1-\alpha}{2} x + \frac{1+\alpha}{2} y \in S,$$

that is, the line segment connecting any two points x and y, belongs to the set S. A strictly convex function of several variables $\Phi(x)$ is a function defined on a nonempty convex set $S = \operatorname{int dom}(\Phi) \subseteq \mathrm{R}^n$ such that for any two points $x \in S$, $y \in S$ and any real number $\alpha \in (-1, 1)$, the following,

$$\Phi\left(\frac{1-\alpha}{2} x + \frac{1+\alpha}{2} y\right) \leq \frac{1-\alpha}{2} \Phi(x) + \frac{1+\alpha}{2} \Phi(y), \tag{2.1}$$

is valid, with equality holding only when $x = y$. Equation 2.1 will sometimes be referred to as the fundamental convex inequality below. Intuitively, the difference between the left-hand side and the right-hand side of (equation 2.1) depends on some kind of proximity between the two points x and y in question, as well as on the degree of convexity (loosely speaking) of the function Φ. For convenience, Φ is assumed to be differentiable up to third order, though this condition can be slightly relaxed to the class of so-called essentially smooth and essentially strictly convex functions or the convex function of the Legendre type (Rockafellar, 1970) without affecting much of the subsequent analysis. Note that for $\alpha = \pm 1$, the equality in equation 2.1 holds uniformly for all x, y; for $\alpha \neq \pm 1$, the equality will not hold uniformly unless $\Phi(x)$ is the linear function $\langle a, x\rangle + b$ with a a constant vector and b a scalar.

Proposition 1. *For any smooth, strictly convex function Φ: $R^n \rightarrow R, x \mapsto \Phi(x)$ and $\alpha \in R$, the function*

$$\begin{aligned} \mathcal{D}_\Phi^{(\alpha)}(x, y) = \frac{4}{1-\alpha^2} \Bigg(&\frac{1-\alpha}{2} \Phi(x) + \frac{1+\alpha}{2} \Phi(y) \\ &- \Phi\left(\frac{1-\alpha}{2} x + \frac{1+\alpha}{2} y\right)\Bigg) \end{aligned} \tag{2.2}$$

with

$$\mathcal{D}_\Phi^{(\pm 1)}(x, y) = \lim_{\alpha \to \pm 1} \mathcal{D}_\Phi^{(\alpha)}(x, y) \tag{2.3}$$

is a parametric family of nonnegative functions of x, y that equal zero if and only if $x = y$. Here, the points x, y, and $z = \frac{1-\alpha}{2} x + \frac{1+\alpha}{2} y$ are all assumed to belong to $S = \mathit{int\,dom}(\Phi)$.

Proof. Clearly, for any $\alpha \in (-1, 1)$, $1 - \alpha^2 > 0$, so from equation 2.1, the functions $\mathcal{D}_\Phi^{(\alpha)}(x, y) \geq 0$ for all $x, y \in S$, with equality holding if and only if $x = y$. When $\alpha > 1$, we rewrite $y = \frac{2}{\alpha+1} z + \frac{\alpha-1}{\alpha+1} x$ as a convex mixture of z and x (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

$$\frac{2}{\alpha+1}\, \Phi(z) + \frac{\alpha-1}{\alpha+1}\, \Phi(x) \geq \Phi(y)$$

or explicitly

$$\frac{2}{1+\alpha} \left(\frac{1-\alpha}{2}\, \Phi(x) + \frac{1+\alpha}{2}\, \Phi(y) - \Phi\left(\frac{1-\alpha}{2}\, x + \frac{1+\alpha}{2}\, y \right) \right) \leq 0.$$

This, along with $1 - \alpha^2 < 0$, proves the nonnegativity of $\mathcal{D}_\Phi^{(\alpha)}(x, y) \geq 0$ for $\alpha > 1$, with equality holding if and only if $z = x$, that is, $x = y$. The case of $\alpha < -1$ is similarly proved by applying equation 2.1 to the three points y, z, and their convex mixture $x = \frac{2}{1-\alpha} z + \frac{-1-\alpha}{1-\alpha} x$. Finally, continuity of $\mathcal{D}_\Phi^{(\alpha)}(x, y)$ with respect to α guarantees that the above claim is also valid in the case of $\alpha = \pm 1$. $\diamond$

Remark 2.1.1. These functions are asymmetric, $\mathcal{D}_\Phi^{(\alpha)}(x, y) \neq \mathcal{D}_\Phi^{(\alpha)}(y, x)$ in general, but satisfy the dual relation

$$\mathcal{D}_\Phi^{(\alpha)}(x, y) = \mathcal{D}_\Phi^{(-\alpha)}(y, x). \tag{2.4}$$

Therefore, $\mathcal{D}_\Phi^{(\alpha)}$ as parameterized by $\alpha \in$ R, for a fixed Φ, properly form a family of divergence functions (also known as deviations or contrast functions) in the sense of Eguchi (1983, 1992), Amari (1982, 1985), Kaas & Vos (1997), and Amari & Nagaoka (2000).

Example 2.1.2. Take the negative of the entropy function $\Phi(x) = \sum_i x^i \log x^i$, which is easily seen to be convex. Then equation 2.2 becomes

$$\frac{4}{1-\alpha^2} \sum_i \left(\frac{1-\alpha}{2} x^i \log \frac{x^i}{\frac{1-\alpha}{2} x^i + \frac{1+\alpha}{2} y^i} \right. \\ \left. + \frac{1+\alpha}{2} y^i \log \frac{y^i}{\frac{1-\alpha}{2} x^i + \frac{1+\alpha}{2} y^i} \right) \equiv E^{(\alpha)}(x, y), \tag{2.5}$$

a family of divergence measure called Jensen difference (Rao, 1987), apart from the $\frac{4}{1-\alpha^2}$ factor. The Kullback-Leibler divergence, equation 1.1, is recovered by letting $\alpha \to \pm 1$ in $E^{(\alpha)}(x, y)$.

Example 2.1.3. Take $\Phi(x) = \sum_i e^{x^i}$ while denoting $p_i = \log x^i,\ q_i = \log y^i$. Then $D_\Phi^{(\alpha)}(x, y)$ becomes the α-divergence $\mathcal{A}^{(\alpha)}(p, q)$ in its discrete version.

2.2 Statistical Manifold Induced by $\mathcal{D}_\Phi^{(\alpha)}$. The divergence function $\mathcal{D}_\Phi^{(\alpha)}(x, y)$ provides a quantitative measure of the asymmetric (directed) distance between a comparison point y as measured from a fixed reference point x. Although this function is globally defined for x, y at large, information geometry provides a standard technique, due to Eguchi (1983), to investigate the differential geometric structure induced on S from any divergence function, through taking $\lim_{x\to x_0,\ y\to x_0} \mathcal{D}_\Phi^{(\alpha)}(x, y)$. The most important geometric objects on a differential manifold are the Riemannian metric g and the affine connection Γ. The metric tensor fixes the inner product operation on the manifold, whereas the affine connection establishes the affine correspondence among neighboring tangent spaces and defines covariant differentiation.

Proposition 2. *The statistical manifold* $\{S, g, \Gamma^{(\alpha)}, \Gamma^{*(\alpha)}\}$ *associated with* $\mathcal{D}_\Phi^{(\alpha)}$ *is given (in component forms) by*

$$g_{ij} = \Phi_{ij} \tag{2.6}$$

and

$$\Gamma_{ij,k}^{(\alpha)} = \frac{1-\alpha}{2}\,\Phi_{ijk}, \qquad \Gamma_{ij,k}^{*(\alpha)} = \frac{1+\alpha}{2}\,\Phi_{ijk}. \tag{2.7}$$

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

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

Proof. Assuming Fréchet differentiability of Φ, we calculate the Taylor expansion of $\mathcal{D}_\Phi^{(\alpha)}(x, y)$ around the reference point x_0 in the direction ξ for the first variable (i.e., $x = x_0 + \xi$) and in the direction of η for the second variable (i.e., $y = x_0 + \eta$), while renaming x_0 as x for clarity:[4]

$$\mathcal{D}_\Phi^{(\alpha)}(x + \xi, x + \eta) \simeq \frac{1}{2} \sum_{i,j} \Phi_{ij}\, (\xi^i - \eta^i)\, (\xi^j - \eta^j)$$

[4] We try to follow the conventions of tensor algebra for upper and lower indices, but do not invoke Einstein summation convention since many of the equalities are not in coordinate-invariant or tensorial form.

$$+\frac{1}{6}\sum_{i,j,k}\Phi_{ijk}\left(\frac{3-\alpha}{2}\xi^i\xi^j\xi^k+\frac{3+\alpha}{2}\eta^i\eta^j\eta^k\right.$$
$$\left.-\frac{3+3\alpha}{2}\eta^i\eta^j\xi^k-\frac{3-3\alpha}{2}\xi^i\xi^j\eta^k\right)+o(\xi^m\eta^l),$$

where higher orders in the expansion (i.e., $m+l \geq 4$) have been collected into $o(\cdot)$. Following Eguchi (1983), the metric tensor of the Riemannian geometry induced by $\mathcal{D}_\Phi^{(\alpha)}$ is

$$g_{ij}(x) = -\frac{\partial^2}{\partial\xi^i\partial\eta^j}\,\mathcal{D}_\Phi^{(\alpha)}(x+\xi, x+\eta)\Big|_{\eta=\xi=0}, \tag{2.8}$$

whereas the pair of dual affine connections Γ and Γ^* is

$$\Gamma_{ij,k}(x) = -\frac{\partial^3}{\partial\xi^i\partial\xi^j\partial\eta^k}\,\mathcal{D}_\Phi^{(\alpha)}(x+\xi, x+\eta)\Big|_{\eta=\xi=0}, \tag{2.9}$$

$$\Gamma^*_{ij,k}(x) = -\frac{\partial^3}{\partial\eta^i\partial\eta^j\partial\xi^k}\,\mathcal{D}_\Phi^{(\alpha)}(x+\xi, x+\eta)\Big|_{\eta=\xi=0}. \tag{2.10}$$

Carrying out differentiation yields equations 2.6 and 2.7. ◇

Remark 2.2.1. Clearly, the metric tensor g_{ij}, which is symmetric and positive semidefinite due to the strict convexity of Φ, is actually independent of α, whereas the α-dependent affine connections are torsion free (since $\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(\alpha)}_{ji,k}$) and satisfy the duality

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

in accordance with equation 2.4, the duality in the selection of reference versus comparison point in $\mathcal{D}_\Phi^{(\alpha)}$. Dual α-connections in the form of equation 2.7 were formally introduced and investigated in Lauritzen (1987). Here, the family of $\mathcal{D}_\Phi^{(\alpha)}$-divergence is shown to induce these α-connections. Clearly, when $\alpha = 0$, the connection $\Gamma^{(0)} = \Gamma^{*(0)} \equiv \Gamma^{LC}$ is the self-dual, metric (Levi-Civita) connection, as through direct verification it satisfies

$$\Gamma^{LC}_{ij,k} = \frac{1}{2}\left(\frac{\partial g_{ik}(x)}{\partial x^j}+\frac{\partial g_{kj}(x)}{\partial x^i}-\frac{\partial g_{ij}(x)}{\partial x^k}\right).$$

The Levi-Civita connection and other members in the α-connection family are related through

$$\Gamma^{LC}_{ij,k} = \frac{1}{2}\left(\Gamma^{(\alpha)}_{ij,k}+\Gamma^{*(\alpha)}_{ij,k}\right).$$

Note the covariant form of the affine connection, $\Gamma_{ij,k}$, is related to its contravariant form Γ_{ij}^k through g_{ij}:

$$\sum g_{kl}\Gamma_{ij}^k = \Gamma_{ij,l}$$

(actually, Γ_{ij}^k is the more primitive quantity since it does not involve the metric). The curvature or flatness of a connection is described by the Riemann-Christoffel curvature tensor,

$$R^i_{j\mu\nu}(x) = \frac{\partial\Gamma^i_{\nu j}(x)}{\partial x^\mu} - \frac{\partial\Gamma^i_{\mu j}(x)}{\partial x^\nu} + \sum_k \Gamma^i_{\mu k}(x)\Gamma^k_{\nu j}(x) - \sum_k \Gamma^i_{\nu k}(x)\Gamma^k_{\mu j}(x),$$

or equivalently by

$$R_{ij\mu\nu} = \sum_l g_{il}\, R^l_{j\mu\nu}.$$

It is antisymmetric when $i \leftrightarrow j$ or when $\mu \leftrightarrow \nu$ and symmetric when $(i, j) \leftrightarrow (\mu, \nu)$. Since the curvature $R^*_{ij\mu\nu}$ of the dual connection Γ^* equals $R_{ij\mu\nu}$ (Lauritzen, 1987),

$$R^{(\alpha)}_{ij\mu\nu} = R^{(-\alpha)}_{ij\mu\nu} = R^{*(\alpha)}_{ij\mu\nu}.$$

Proposition 3. *The Riemann-Christoffel curvature tensor for the α-connection $\Gamma^{(\alpha)}_{ij,k}$ induced by $\mathcal{D}^{(\alpha)}_\Phi$ is*

$$R^{(\alpha)}_{ij\mu\nu} = \frac{1-\alpha^2}{4}\sum_{l,k}(\Phi_{il\nu}\Phi_{jk\mu} - \Phi_{il\mu}\Phi_{jk\nu})\Phi^{lk}, \tag{2.11}$$

where $\Phi^{ij} = g^{ij}$ is the matrix inverse of Φ_{ij}.

Proof. First, from its definition, $R_{ij\mu\nu}$ can be recast into

$$R_{ij\mu\nu} = \frac{\partial\Gamma_{\nu j,i}}{\partial x^\mu} - \frac{\partial\Gamma_{\mu j,i}}{\partial x^\nu} + \sum_k \left(\Gamma^k_{\nu j}\left(\Gamma_{\mu k,i} - \frac{\partial g_{ik}}{\partial x^\mu}\right) - \Gamma^k_{\mu j}\left(\Gamma_{\nu k,i} - \frac{\partial g_{ik}}{\partial x^\nu}\right)\right). \tag{2.12}$$

Substituting in the expression of α-connections (equation 2.7), the first two terms cancel and the terms under $\sum_k$ give rise to equation 2.11. ◇

Remark 2.2.2. The metric (see equation 2.6), dual α-connections (see equation 2.7), and the curvature (see equation 2.11) in such forms first appeared in Amari (1985) where $\Phi(x)$ is the cumulant generating function of an exponential family. Here, the statistical manifold $\{S, g, \Gamma^{(\alpha)}, \Gamma^{*(\alpha)}\}$ is induced by a divergence function via the Eguchi relation, and $\Phi(x)$ can be any (smooth and strictly) convex function. The purpose of this proposition is to remind readers that for any convex Φ in general, the curvature is determined by both an α-dependent factor $\frac{4}{1-\alpha^2}$ and a Φ-related component, the latter depending on Φ_{ij} plus its derivatives and inverse. This leads to the following conformal property:

Corollary 1. *If two smooth, strictly convex functions $\Phi(x)$ and $\hat{\Phi}(x)$ are conformally related, that is, if there exists some positive function $\sigma(x) > 0$ such that $\hat{\Phi}_{ij} = \sigma \Phi_{ij}$, then the curvatures of their respective α-connection satisfy*

$$\hat{R}^{(\alpha)}_{ij\mu\nu} = \sigma R^{(\alpha)}_{ij\mu\nu}. \tag{2.13}$$

Proof. Observe that

$$\hat{\Phi}_{il\nu} = \sigma \Phi_{il\nu} + \sigma_i \Phi_{l\nu},$$

where σ_i denotes $\partial\sigma/\partial x^i$. Permutating the index set (i, l, ν) to (i, l, μ), to (j, k, μ), and to (j, k, ν) yield three other similar identities. Noting $\hat{\Phi}^{ij} = (\sigma)^{-1}\Phi^{ij}$, direct substitution of these relations into the expression of $R^{(\alpha)}_{ij\mu\nu}$ in equation 2.11 yields equation 2.13. ◇

2.3 Dually Flat Statistical Manifold ($\alpha = \pm 1$).

When $\alpha = \pm 1$, all components of the curvature tensor vanish, that is, $R^{(\pm 1)}_{ij\mu\nu} = 0$. In this case, there exists a coordinate system under which either $\Gamma^{*(-1)}_{ij,k} = 0$ or $\Gamma^{(1)}_{ij,k} = 0$. This is the well-studied dually flat statistical manifold (Amari, 1982, 1985; Amari & Nagaoka, 2000), under which a pair of global biorthogonal coordinates, related to each other through the Legendre transform with respect to Φ, can be found to cast the divergence function into its canonical form. The Riemannian manifold with metric tensor given by equation 2.6, along with the dually flat $\Gamma^{(1)}$ and $\Gamma^{*(-1)}$, is known as the Hessian manifold (Shima, 1978; Shima & Yagi, 1997).

Proposition 4. *When $\alpha \to \pm 1$, $\mathcal{D}^{(\alpha)}_{\Phi}$ reduces to the Bregman divergence (see equation 1.7)*

$$\begin{aligned} \mathcal{D}^{(-1)}_{\Phi}(x, y) &= \mathcal{D}^{(1)}_{\Phi}(y, x) = B_{\Phi}(x, y), \\ \mathcal{D}^{(1)}_{\Phi}(x, y) &= \mathcal{D}^{(-1)}_{\Phi}(y, x) = B_{\Phi}(y, x). \end{aligned}$$

Proof. Assuming that the Gâteaux derivative of Φ,

$$\lim_{\lambda\to 0}\frac{\Phi(x+\lambda\xi)-\Phi(x)}{\lambda},$$

exists and equals $\langle \xi, \nabla\Phi(x)\rangle$, where ∇ is the gradient (subdifferential) operator and $\langle\cdot,\cdot\rangle$ denotes the standard inner product. Similarly,

$$\lim_{\lambda\to 1}\frac{\Phi(y+(1-\lambda)\eta)-\Phi(y)}{1-\lambda}=\langle \eta, \nabla\Phi(y)\rangle.$$

Taking $\xi = y - x$, $\eta = x - y$, and $\lambda = \frac{1\pm\alpha}{2}$, and substituting these into equation 2.3 immediately yields the results. ◇

Remark 2.3.1. Introducing the convex conjugate of Φ through the Legendre-Fenchel transform (see, e.g., Rockafellar, 1970),

$$\Phi^*(u) = \langle u, (\nabla\Phi)^{-1}(u)\rangle - \Phi((\nabla\Phi)^{-1}(u)), \tag{2.14}$$

it can be shown that the function Φ^* is also convex (on a convex region $S' \ni u$ where $u = \nabla\Phi(x)$) and has Φ as its conjugate,

$$(\Phi^*)^* = \Phi.$$

Since $\nabla\Phi$ and $\nabla\Phi^*$ are inverse functions of each other, as verified by differentiating equation 2.14, it is convenient to denote this one-to-one correspondence between $x \in S$ and $u \in S'$ by

$$x = \nabla\Phi^*(u) = (\nabla\Phi)^{-1}(u), \qquad u = \nabla\Phi(x) = (\nabla\Phi^*)^{-1}(x). \tag{2.15}$$

With these, it can be shown that the Bregman divergence $\mathcal{D}_\Phi^{(\pm 1)}$ is actually the canonical divergence (Amari & Nagaoka, 2000) in disguise.

Corollary 2. *The $\mathcal{D}_\Phi^{(\pm 1)}$-divergence is the canonical divergence of a dually flat statistical manifold:*

$$\begin{aligned}\mathcal{D}_\Phi^{(1)}(x, (\nabla\Phi)^{-1}(v)) &= A_\Phi(x, v) \equiv \Phi(x) + \Phi^*(v) - \langle x, v\rangle,\\ \mathcal{D}_\Phi^{(-1)}((\nabla\Phi)^{-1}(u), y) &= A_{\Phi^*}(u, y) \equiv \Phi(y) + \Phi^*(u) - \langle u, y\rangle.\end{aligned} \tag{2.16}$$

Proof. Using the convex conjugate Φ^*, we have

$$\begin{aligned}\mathcal{D}_\Phi^{(1)}(x, y) &= \Phi(x) - \langle x, \nabla\Phi(y)\rangle + \Phi^*(\nabla\Phi(y)),\\ \mathcal{D}_\Phi^{(-1)}(x, y) &= \Phi(y) - \langle y, \nabla\Phi(x)\rangle + \Phi^*(\nabla\Phi(x)).\end{aligned}$$

Substituting $u = \nabla\Phi(x)$, $v = \nabla\Phi(y)$ yields equation 2.16. So $\mathcal{D}_{\Phi}^{(1)}(x, (\nabla\Phi)^{-1}(v))$, when viewed as a function of x, v, is the canonical divergence. So is $\mathcal{D}_{\Phi}^{(-1)}$. ⋄

Remark 2.3.2. The canonical divergence $A_{\Phi}(x, v)$ based on the Legendre-Fenchel inequality is 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, equation 2.16, is "canonical" because it is uniquely specified in a dually flat space where the pair of canonical coordinates (see equation 2.15) induced by the dual potential functions is biorthogonal,

$$\frac{\partial u_i}{\partial x^j} = g_{ij}(x), \qquad \frac{\partial x_i}{\partial u^j} = \tilde{g}^{ij}(u), \tag{2.17}$$

where $\tilde{g}^{ij}(u(x)) = g^{ij}(x)$ is the matrix inverse of $g_{ij}(x)$ given by equation 2.6.

Remark 2.3.3. We point out that there are two kinds of duality associated with the divergence (directed distance) defined on a dually flat statistical manifold: one between $\mathcal{D}_{\Phi}^{(-1)} \leftrightarrow \mathcal{D}_{\Phi}^{(1)}$ and $\mathcal{D}_{\Phi^*}^{(-1)} \leftrightarrow \mathcal{D}_{\Phi^*}^{(1)}$, the other between $\mathcal{D}_{\Phi}^{(-1)} \leftrightarrow \mathcal{D}_{\Phi^*}^{(-1)}$ and $\mathcal{D}_{\Phi}^{(1)} \leftrightarrow \mathcal{D}_{\Phi^*}^{(1)}$. The first kind is related to the duality in the choice of the reference and the comparison status for the two points (x versus y) for computing the value of the divergence, and hence is 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 (x versus u) in the expression of the divergence function, and hence is called representational duality. More concretely,

$$\begin{aligned}\mathcal{D}_{\Phi}^{(-1)}(x, y) &= \mathcal{D}_{\Phi^*}^{(-1)}(\nabla\Phi(y), \nabla\Phi(x)) \\ &= \mathcal{D}_{\Phi^*}^{(1)}(\nabla\Phi(x), \nabla\Phi(y)) = \mathcal{D}_{\Phi}^{(1)}(y, x).\end{aligned}$$

The biduality is compactly reflected in the canonical divergence as

$$A_{\Phi}(x, v) = A_{\Phi^*}(v, x).$$

2.4 Biduality of Statistical Manifold for General α. A natural question to ask is whether biduality is a general property of the divergence $\mathcal{D}_{\Phi}^{(\alpha)}$ and hence a property of any statistical manifold admitting a metric and a pair of dual (but not necessarily flat) connections. Proposition 5 provides a positive answer to this question after considering the geometry generated by the pair of conjugate divergence functions, $\mathcal{D}_{\Phi}^{(\alpha)}$ and $\mathcal{D}_{\Phi^*}^{(\alpha)}$, for each $\alpha \in$ R.

Proposition 5. *For the statistical manifold $\{S', \tilde{g}, \tilde{\Gamma}^{(\alpha)}, \tilde{\Gamma}^{*(\alpha)}\}$ induced by $\mathcal{D}^{(\alpha)}_{\Phi^*}(u, v)$ defined on $u, v \in S'$, denote the Riemannian metric as $\tilde{g}^{mn}(u)$, the pair of dual connections as $\tilde{\Gamma}^{(\alpha)mn,l}(u)$, $\tilde{\Gamma}^{*(\alpha)mn,l}(u)$, and the Riemann-Christoffel curvature tensor as $\tilde{R}^{(\alpha)klmn}(u)$. They are related to those (in lower scripts and without the tilde) induced by $\mathcal{D}^{(\alpha)}_{\Phi}(x, y)$ via*

$$\sum_l g_{il}(x)\tilde{g}^{ln}(u) = \delta_i^n,$$

and

$$\begin{aligned}
\tilde{\Gamma}^{(\alpha)mn,l}(u) &= -\sum_{i,j,k} \tilde{g}^{im}(u)\tilde{g}^{jn}(u)\tilde{g}^{kl}(u)\Gamma^{(\alpha)}_{ij,k}(x),\\
\tilde{\Gamma}^{*(\alpha)mn,l}(u) &= -\sum_{i,j,k} \tilde{g}^{im}(u)\tilde{g}^{jn}(u)\tilde{g}^{kl}(u)\Gamma^{(-\alpha)}_{ij,k}(x),\\
\tilde{R}^{(\alpha)klmn}(u) &= \sum_{i,j,\mu,\nu} \tilde{g}^{ik}(u)\tilde{g}^{jl}(u)\tilde{g}^{\mu m}(u)\tilde{g}^{\nu n}(u)R^{(\alpha)}_{ij\mu\nu}(x)
\end{aligned}$$

where the dual correspondence (see equation 2.15) is invoked.

Proof. Following the proof of proposition 2 (i.e., using the Eguchi relation), the metric and dual connections induced on S' are simply

$$\tilde{g}^{mn} = (\Phi^*)^{mn}$$

and

$$\tilde{\Gamma}^{(\alpha)mn,l} = \frac{1-\alpha}{2}(\Phi^*)^{mnl}, \qquad \tilde{\Gamma}^{*(\alpha)mn,l} = \frac{1+\alpha}{2}(\Phi^*)^{mnl},$$

with the corresponding Riemann-Christoffel curvature of $\tilde{\Gamma}^{(\alpha)mn,l}$ given as

$$\tilde{R}^{(\alpha)klmn} = \frac{1-\alpha^2}{4}\sum_{\rho,\tau}((\Phi^*)^{k\rho n}(\Phi^*)^{l\tau m} - (\Phi^*)^{k\rho m}(\Phi^*)^{l\tau n})(\Phi^*)_{\rho\tau},$$

where

$$(\Phi^*)^{mn} = \frac{\partial^2\Phi^*(u)}{\partial u_m \partial u_n}, \qquad (\Phi^*)^{mnl} = \frac{\partial^3\Phi^*(u)}{\partial u_m \partial u_n \partial u_l},$$

and $(\Phi^*)_{\rho\tau}$ is the matrix inverse of $(\Phi^*)^{mn}$. That $\sum_l g_{il}(x)\tilde{g}^{ln}(u(x)) = \sum_l g_{il}(x(u))\tilde{g}^{ln}(u) = \delta_i^n$ is due to equations 2.15 and 2.17. Differentiating this identity with respect to x^k yields

$$\sum_m \frac{\partial g_{im}(x)}{\partial x^k}\,\tilde{g}^{mn}(u) = -\sum_m g_{im}(x)\,\frac{\partial \tilde{g}^{mn}(u)}{\partial x^k}$$
$$= -\sum_m g_{im}(x)\left(\sum_l \frac{\partial (\Phi^*)^{mn}(u)}{\partial u_l}\frac{\partial u_l}{\partial x^k}\right)$$

or

$$\sum_m \Phi_{imk}(x)\,\tilde{g}^{mn}(u) = -\sum_{m,l} g_{im}(x)\,g_{kl}(x)\,(\Phi^*)^{mnl}(u),$$

which immediately gives rise to the desired relations between the α-connections. Simple substitution yields the relation between $\tilde{R}^{(\alpha)klmn}$ and $R^{(\alpha)}_{ij\mu\nu}$. $\diamond$

Remark 2.4.1. The biorthogonal coordinates x and u were originally defined on the manifold S and its dual S', respectively. Because of the bijectivity of the mapping between x and u, we may identify points in S with points in S' and simply view $x \leftrightarrow u$ as coordinate transformations on the same underlying manifold. The relations between the metric g, dual connections $\Gamma^{(\pm\alpha)}$, or the curvature $R^{(\alpha)}$ written in superscripts with tilde and those written in subscripts without tilde are merely expressions of the same geometric entities using different coordinate systems. The dualistic geometric structure $\Gamma^{(\alpha)} \leftrightarrow \Gamma^{*(\alpha)}$ under g, which reflects referential duality, is preserved under the mapping $x \leftrightarrow u$, which reflects representational duality. When the manifold is dually flat ($\alpha = \pm 1$), x and u enjoy the additional property of being geodesic coordinates.

Remark 2.4.2. Matumoto (1993) proved that a divergence function always exists for a statistical manifold equipped with an arbitrary metric tensor and a pair of dual connections. Given a convex function Φ, along with its unique conjugate Φ^*, are there other families of divergence functions $\mathcal{D}^{(\alpha)}_{\Phi}(x, y)$ and $\mathcal{D}^{(\alpha)}_{\Phi^*}(u, v)$ that induce the same bidualistic statistical manifolds $\{S, g, \Gamma^{(\alpha)}, \Gamma^{*(\alpha)}\}$? The answer is positive. Consider the family of divergence functions,

$$D^{(\alpha)}_{\Phi}(x, y) = \frac{1-\alpha}{2}\,\mathcal{D}^{(-1)}_{\Phi}(x, y) + \frac{1+\alpha}{2}\,\mathcal{D}^{(1)}_{\Phi}(x, y),$$

and their conjugate (replacing Φ with Φ^*). Recall from proposition 2 that the metric tensor induced by $\mathcal{D}^{(-1)}_{\Phi}(x, y)$ and $\mathcal{D}^{(1)}_{\Phi}(x, y)$ is the same g_{ij}, while the induced connections satisfy $\Gamma^{(-1)}_{ij,k} = \Gamma^{*(1)}_{ij,k} = \Phi_{ijk}$, $\Gamma^{(1)}_{ij,k} = \Gamma^{*(-1)}_{ij,k} = 0$.

Since the Eguchi relations, equations 2.8 to 2.10, are linear with respect to inducing functions, the family of divergence functions $D^{(\alpha)}_{\Phi}(x, y)$, as convex mixture of $\mathcal{D}^{(-1)}_{\Phi}(x, y)$ and $\mathcal{D}^{(1)}_{\Phi}(x, y)$, will necessarily induce the metric

$$\frac{1-\alpha}{2} g_{ij} + \frac{1+\alpha}{2} g_{ij} = g_{ij},$$

and dual connections

$$\begin{aligned} \frac{1-\alpha}{2} \Gamma^{(-1)}_{ij,k} + \frac{1+\alpha}{2} \Gamma^{(1)}_{ij,k} &= \Gamma^{(\alpha)}_{ij,k}, \\ \frac{1-\alpha}{2} \Gamma^{*(-1)}_{ij,k} + \frac{1+\alpha}{2} \Gamma^{*(1)}_{ij,k} &= \Gamma^{*(\alpha)}_{ij,k}. \end{aligned}$$

Similar arguments apply to $D^{(\alpha)}_{\Phi^*}(u, v)$. This is, $D^{(\alpha)}_{\Phi}(x, y)$ and $D^{(\alpha)}_{\Phi^*}(u, v)$ form another pair of families of divergence functions that induce the same statistical manifold $\{S, g, \Gamma^{(\alpha)}, \Gamma^{*(\alpha)}\}$. The two pairs, $(\mathcal{D}^{(\alpha)}_{\Phi}(x, y), \mathcal{D}^{(\alpha)}_{\Phi^*}(u, v))$ pair, and $(D^{(\alpha)}_{\Phi}(x, y), D^{(\alpha)}_{\Phi^*}(u, v))$ pair, agree on $\alpha = \pm 1$, the dually flat case when there is a single form of canonical divergence. They differ for any other values of α, including the self-dual element ($\alpha = 0$). The two pairs $(\mathcal{D}^{(\alpha)}_{\Phi}, \mathcal{D}^{(\alpha)}_{\Phi^*})$ versus $(D^{(\alpha)}_{\Phi}, D^{(\alpha)}_{\Phi^*})$ coincide with each other up to the third order when Taylor expanding $(1 \pm \alpha)(y - x)$. That is why they induce an identical statistical manifold. They differ on the fourth-order terms in their expansions.

3 Divergence on Probability and Positive Measures

The previous sections have discussed divergence functions defined between points in R^n or in its dual space, or both. In particular, they apply to probability measures of finite support, or the finite-dimensional parameter space, which defines parametric probability models. Traditionally, divergence functionals are also defined with respect to infinite-dimensional probability densities (or positive measures in general if normalization constraint is removed). To the extent that a probability density function can be embedded into a finite-dimensional parameter space, a divergence measure on density functions will implicitly induce a divergence on the parameter space (technically, via pullback). In fact, this is the original setup in Amari (1985), where each α-divergence ($\alpha \in \mathrm{R}$) is seen as the canonical divergence arising from the α-embedding of the probability density function into an affine submanifold (the condition of α-affinity). The approach outlined below avoids such a flatness assumption while still achieving conjugate representations (embeddings) of probability densities, and therefore extends the notion of biduality to the infinite-dimensional case. It will be proved (in proposition 9) that if the embedding manifold is flat, then the induced

α-connections reduce to the ones introduced in the previous section, with the natural and expectation parameters arising out of these conjugate representations forming biorthogonal coordinates just like the ones induced by dual potential functions in the finite-dimensional case.

To follow the procedures of section 2.1 and construct divergence functionals, a smooth and strictly convex function defined on the real line f: R → R is introduced. Recall that any such f can be written as an integral of a strictly monotone-increasing function g and vice versa: $f(\gamma) = \int_c^\gamma g(t)dt + f(c)$, with $g'(t) > 0$. The convex conjugate f^*: R → R is given by $f^*(\delta) = \int_{g(c)}^\delta g^{-1}(t)dt + f^*(g(c))$, with g^{-1} also strictly monotonic and $\gamma, \delta \in$ R. Here, the monotonicity condition replaces the requirement of a positive semidefinite Hessian in the case of a convex function of several variables. The Legendre-Fenchel inequality $f(\gamma) + f^*(\delta) \geq \gamma\delta$ can be rewritten as Young's inequality,

$$\int_c^\gamma g(t)\,dt + \int_{g(c)}^\delta g^{-1}(t)\,dt + cg(c) \geq \gamma\,\delta,$$

with equality holding if and only if $\delta = g(\gamma)$. The use of a pair of strictly monotonic functions $f' = g$ and $(f^*)' = g^{-1}$, which we call ρ- and τ-representations below, provides a means to define conjugate embeddings (representations) of density functions and therefore a method to extend the analysis in the previous sections to the infinite-dimensional manifold of positive measures (after integrating over the sample space).

3.1 Divergence Functional Based on Convex Function on the Real Line. Recall that the fundamental inequality (see equation 2.1) of a strictly convex function Φ, now for f: R → R, can be used to define a nonnegative quantity (for any $\alpha \in$ R),

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

Note that here, γ and δ are numbers instead of finite-dimensional vectors. In particular, they can be the values of two probability density functions $\gamma = p(\zeta)$ and $\delta = q(\zeta)$ where $\zeta \in \mathcal{X}$ is a point in the sample space $\mathcal{X}$. The nonnegativity of the above expression for each value of ζ allows us to define a global divergence measure, called a divergence functional, over the space of a (denormalized) probability density function after integrating over the sample space (with appropriate measure $\mu(d\zeta) = d\mu$),

$$\begin{aligned} d_f^{(\alpha)}(p, q) &= \int d_f^{(\alpha)}(p(\zeta), q(\zeta))\,d\mu \\ &= \frac{4}{1-\alpha^2}\left\{\frac{1-\alpha}{2}\left(\int f(p)\,d\mu\right) + \frac{1+\alpha}{2}\left(\int f(q)\,d\mu\right)\right. \\ &\qquad \left. - \int f\left(\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q\right)d\mu\right\}, \end{aligned}$$

with

$$d_f^{(-1)}(p,q) = d_f^{(1)}(q,p) = \int (f(q) - f(p) - (q-p)f'(p))\, d\mu \tag{3.1}$$

$$= \int (f(q) + f^*(f'(p)) - qf'(p)) d\mu \equiv A_f(q, f'(p)) \tag{3.2}$$

where f^*: R → R, defined by

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

is the convex conjugate to f, with $(f^*)^* = f$ and $(f^*)' = (f')^{-1}$.

In this way, $d_f^{(\alpha)}(p,q)$ over the infinite-dimensional functional space is defined in much the same way as $\mathcal{D}_\Phi^{(\alpha)}(x,y)$ defined on the finite-dimensional vector space. The integration $\int f(p)d\mu = \int f(p(\zeta))d\mu$, which is a nonlinear functional of p, assumes the role of Φ of the finite-dimensional case discussed in section 2; this is most transparent if we consider, for the latter, the special class of "separable" convex functions $\Phi(x) = \sum_{i=1}^n f(x^i)$, $x \in \mathrm{R}^n$ such that $\nabla\Phi(x)$ is simply $[f'(x^1), \ldots, f'(x^n)]$. With $\sum_i \leftrightarrow \int d\mu$, the expressions of the divergence function and the divergence functional look similar. However, one should not conclude that divergence functions are special cases of divergence functionals or vice versa. There is a subtle but important difference: in the former, the inducing function $\Phi(x)$ is strictly convex in x; in the latter, $f(p)$ is strictly convex in p, but its pullback on $\mathcal{X}$, $f(p(\zeta))$ is not assumed to be convex at all. So even when the sample space may be finite, $(f \circ p)(\zeta)$ is generally not a convex function of ζ.

Example 3.1.1. Take $f(t) = t \log t - t\,(t > 0)$, with conjugate function $f^*(u) = e^u$. The divergence

$$A_f(p,u) = \int \big((p \log p - p) + e^u - p\,u\big)\, d\mu$$
$$= \int \left(p \log \frac{p}{e^u} - p + e^u\right)\, d\mu$$

is the Kullback-Leibler divergence $K(p, e^u)$ between $p(\zeta)$ and $q(\zeta) = e^{u(\zeta)}$.

Example 3.1.2. Take $f(t) = \frac{t^r}{r}$ $(t > 0)$ with the conjugate function $f^*(t) = \frac{t^{r'}}{r'}$, where the pair of conjugated real exponents $r > 1, r' > 1$ satisfies

$$\frac{1}{r} + \frac{1}{r'} = 1.$$

The divergence A_f is a nonnegative expression based on Hölder's inequality for two functions $u(\zeta)$, $v(\zeta)$,

$$A_f(u, v) = \int \left(\frac{u^r}{r} + \frac{v^{r'}}{r'} - u\,v \right) d\mu \geq 0,$$

with equality holding if and only if $u^r(\zeta) = v^{r'}(\zeta)$ for all $\zeta \in \mathcal{X}$. Denote $r = \frac{2}{1-\alpha}$ and $r' = \frac{2}{1+\alpha}$, with $\alpha \in (-1, 1)$. The above divergence is just $\mathcal{A}^\alpha(p, q)$ between $p(\zeta) = (u(\zeta))^{\frac{2}{1-\alpha}}$ and $q(\zeta) = (v(\zeta))^{\frac{2}{1+\alpha}}$, apart from a factor $\frac{4}{1-\alpha^2}$.

3.2 Conjugate Representations and Induced Statistical Manifold. We introduce the notion of ρ-representation of a (not necessarily normalized) probability density by defining a mapping ρ: $\mathrm{R}_+ \to \mathrm{R}$, $p \mapsto \rho(p)$ where ρ is a strictly monotone increasing function. This is a generalization of the α-representation (Amari, 1985; Amari & Nagaoka, 2000) where $\rho(p) = l^{(\alpha)}(p)$, as given by equation 1.8. For a smooth and strictly convex function f: $\mathrm{R} \to \mathrm{R}$, the τ-representation of the density function $p \mapsto \tau(p)$ is said to be conjugate to the ρ-representation with respect to f if

$$\begin{aligned} &\tau(p) = f'(\rho(p)) = ((f^*)')^{-1}(\rho(p)) \longleftrightarrow \\ &\rho(p) = (f')^{-1}(\tau(p)) = (f^*)'(\tau(p)). \end{aligned} \tag{3.3}$$

Just like the construction in section 3.1 of divergence functionals for two densities p, q, one may construct divergence functionals for two densities under ρ-representations $\mathcal{D}^{(\alpha)}_{f,\rho}(p, q) \equiv d^{(\alpha)}_f(\rho(p), \rho(q))$ or under τ-representations $\mathcal{D}^{(\alpha)}_{f^*,\tau}(p, q) \equiv d^{(\alpha)}_{f^*}(\tau(p), \tau(q))$.

Proposition 6. *For $\alpha \in R$, $\mathcal{D}^{(\alpha)}_{f,\rho}(p, q)$, $\mathcal{D}^{(\alpha)}_{f,\tau}(p, q)$, $\mathcal{D}^{(\alpha)}_{f^*,\rho}(p, q)$, and $\mathcal{D}^{(\alpha)}_{f^*,\tau}(p, q)$, each forms a one-parameter family of divergence functionals, with the (± 1)-divergence functional,*

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

taking the following canonical form,

$$\begin{aligned} A_f(\rho(p), \tau(q)) &\equiv \int \big(f(\rho(p)) + f^*(\tau(q)) - \rho(p)\,\tau(q) \big)\, d\mu \\ &= A_{f^*}(\tau(q), \rho(p)) \ . \end{aligned} \tag{3.4}$$

Proof. The proof for nonnegativity of these functionals for all $\alpha \in \mathrm{R}$ follows that in proposition 2. Taking $\lim_{\alpha\to\pm 1} \mathcal{D}^{(\alpha)}_{f,\rho}(p,q)$ and noting equation 3.3 immediately leads to the expressions of (±1)-divergence functional. ◇

Example 3.2.1. Amari's α-embedding where $\rho(p) = l^{(\alpha)}(p),\ \tau(p) = l^{(-\alpha)}(p)$ corresponds to (assuming $\alpha \neq \pm 1$)

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

Writing out $A_f(\rho(p), \tau(q))$ explicitly yields the α-divergence in the form of equation 1.2. For $\alpha = \pm 1$, see example 3.1.1.

Now we restrict attention to a finite-dimensional submanifold of probability densities whose ρ-representations are parameterized using $\theta = [\theta^1, \ldots, \theta^n] \in \mathcal{M}_\theta$. Under such a statistical model, the divergence functional of any two densities p and q, assumed to be specified by θ_p and θ_q, respectively, becomes an implicit function of $\theta_p, \theta_q \in \mathrm{R}^n$. In other words, through introducing parametric models (i.e., a finite-dimensional submanifold) of the infinite-dimensional manifold of probability densities, we again arrive at divergence functions over the vector space. We denote the ρ-representation of a parameterized probability density as $\rho(p(\zeta;\theta))$, or sometimes simply $\rho(\theta_p)$, while suppressing the sample space variable ζ, and denote the corresponding divergence function by

$$\mathcal{D}^{(\alpha)}_{f,\rho}(\theta_p,\theta_q) = \frac{4}{1-\alpha^2}\,\mathrm{E}_\mu\left\{\frac{1-\alpha}{2}f(\rho(\theta_p)) + \frac{1+\alpha}{2}f(\rho(\theta_q)) - f\left(\frac{1-\alpha}{2}\rho(\theta_p) + \frac{1+\alpha}{2}\rho(\theta_q)\right)\right\}, \tag{3.5}$$

where $\mathrm{E}_\mu\{\cdot\}$ denotes $\int\{\cdot\}\,d\mu$. We will also use $\mathrm{E}_p\{\cdot\}$ to denote $\int\{\cdot\}\,p\,d\mu$ later. Similarly, the parametrically embedded probability density under τ-representation is denoted $\tau(p(\zeta;\theta))$ or simply $\tau(\theta_p)$.

Proposition 7. *The family of divergence functions $\mathcal{D}^{(\alpha)}_{f,\rho}(\theta_p,\theta_q)$ induces a dually affine Riemannian manifold $\{\mathcal{M}_\theta, g, \Gamma^{(\alpha)}, \Gamma^{*(\alpha)}\}$ for each $\alpha \in R$, with the metric tensor*

$$g_{ij}(\theta) = E_\mu\left\{f''(\rho(\theta))\,\frac{\partial\rho}{\partial\theta^i}\,\frac{\partial\rho}{\partial\theta^j}\right\} \tag{3.6}$$

and the dual affine connections

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\mu\left\{\frac{1-\alpha}{2}\,f'''(\rho)\,A_{ijk} + f''(\rho)\,B_{ijk}\right\}, \tag{3.7}$$

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = E_\mu \left\{ \frac{1+\alpha}{2} f'''(\rho)\, A_{ijk} + f''(\rho)\, B_{ijk} \right\}. \tag{3.8}$$

Here, ρ and all its partial derivatives (with respect to θ) are functions of θ and ζ, while A_{ijk}, B_{ijk} denote

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

Proof. We follow the same technique of section 2.2 to expand the value of divergence measure $\mathcal{D}^{(\alpha)}_{f,\rho}(\theta_p, \theta_q)$ around $\theta_p = \theta + \xi$, $\theta_q = \theta + \eta$ for small $\xi, \eta \in \mathrm{R}^n$. Considering the order of expansion $o(\xi^m \eta^l)$ with nonnegative integers m, l, the terms with $m + l \le 1$ vanish uniformly. The terms with $m + l = 2$ are

$$\mathrm{E}_\mu \left\{ \frac{1}{2} \sum_{i,j} f''(\rho) \frac{\partial\rho}{\partial\theta^i}\frac{\partial\rho}{\partial\theta^j} (\xi^i - \eta^i)(\xi^j - \eta^j) \right\},$$

which is α-independent. The terms with $m + l = 3$ are (after lengthy calculation)

$$\mathrm{E}_\mu \left\{ \frac{1}{6} \sum_{i,j,k} f'''(\rho)\, A_{ijk} \left(\frac{3-\alpha}{2}\xi^i\xi^j\xi^k + \frac{3+\alpha}{2}\eta^i\eta^j\eta^k - \frac{3-3\alpha}{2}\xi^i\xi^j\eta^k - \frac{3+3\alpha}{2}\eta^i\eta^j\xi^k \right) + \frac{1}{2}\sum_{i,j,k} f''(\rho)\, B_{ijk} (\xi^i\xi^j - \eta^i\eta^j)(\xi^k - \eta^k) \right\}.$$

Applying Eguchi relations 2.8 to 2.10 and carrying out differentiation yields the desired results. ◇

Remark 3.2.2. Strict convexity of f requires that $f'' > 0$; thereby, the positive semidefiniteness of g_{ij} is guaranteed. Clearly, the α-connections form dual pairs $\Gamma^{*(\alpha)}_{ijk} = \Gamma^{(-\alpha)}_{ijk}$. The induced statistical manifold will be shown to demonstrate biduality just as in section 2. It is important to realize that while f is strictly convex in p, $f(p(\theta))$ is not at all convex in θ. Therefore, propositions 2 and 7 do not imply one another necessarily!

Example 3.2.3. For $f(t) = e^t$, and $\rho(p) = \log p$, that is, $\tau(p) = p$, the identity function, the above expressions reduce to the Fisher information and

α-connections of the exponential family. When $\rho(p) = l^{(\beta)}(p)$ is the parametric alpha representation, then $g_{ij}(\theta)$ reduces to $\int(\partial\rho^{(\beta)}/\partial\theta^i)\,(\partial\rho^{(-\beta)}/\partial\theta^j)d\mu$ as given in Amari (1985) and Amari and Nagaoka (2000). The formula for α-connections, however, differs from theirs, since with $\rho = l^{(\beta)}$ parametrically, we will get a two-parameter family of connections with α and β as parameters. (See section 3.5.)

3.3 Biduality of Statistical Manifold

Proposition 8. *Under conjugate ρ- and τ-representations, the metric tensor induced by $\mathcal{D}^{(\alpha)}_{f,\rho}(\theta_p, \theta_q)$ can be expressed as*

$$g_{ij}(\theta) = E_\mu\left\{\frac{\partial\rho}{\partial\theta^i}\frac{\partial\tau}{\partial\theta^j}\right\} = E_\mu\left\{\frac{\partial\tau}{\partial\theta^i}\frac{\partial\rho}{\partial\theta^j}\right\}, \tag{3.9}$$

while the induced dual connections are

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\mu\left\{\frac{1-\alpha}{2}\frac{\partial^2\tau}{\partial\theta^i\partial\theta^j}\frac{\partial\rho}{\partial\theta^k} + \frac{1+\alpha}{2}\frac{\partial^2\rho}{\partial\theta^i\partial\theta^j}\frac{\partial\tau}{\partial\theta^k}\right\}, \tag{3.10}$$

$$\Gamma^{*(\alpha)}_{ij,k}(\theta) = E_\mu\left\{\frac{1+\alpha}{2}\frac{\partial^2\tau}{\partial\theta^i\partial\theta^j}\frac{\partial\rho}{\partial\theta^k} + \frac{1-\alpha}{2}\frac{\partial^2\rho}{\partial\theta^i\partial\theta^j}\frac{\partial\tau}{\partial\theta^k}\right\}. \tag{3.11}$$

Proof. First,

$$\mathrm{E}_\mu\left\{f''(\rho)\frac{\partial\rho}{\partial\theta^i}\frac{\partial\rho}{\partial\theta^j}\right\} = \mathrm{E}_\mu\left\{\frac{\partial f'(\rho)}{\partial\theta^i}\frac{\partial\rho}{\partial\theta^j}\right\}.$$

Realizing $f'(\rho(p)) = \tau(p)$ proves equation 3.9. Second, differentiating the above with respect to θ^k yields

$$\mathrm{E}_\mu\left\{f'''(\rho)\frac{\partial\rho}{\partial\theta^i}\frac{\partial\rho}{\partial\theta^j}\frac{\partial\rho}{\partial\theta^k} + f''(\rho)\frac{\partial^2\rho}{\partial\theta^i\partial\theta^k}\frac{\partial\rho}{\partial\theta^j}\right\} = \mathrm{E}_\mu\left\{\frac{\partial^2 f'(\rho)}{\partial\theta^i\partial\theta^k}\frac{\partial\rho}{\partial\theta^j}\right\}.$$

Rearranging,

$$\mathrm{E}_\mu\{f'''(\rho)A_{ijk}\} = T_{ikj}, \tag{3.12}$$

where

$$\begin{aligned} T_{ijk} &\equiv \mathrm{E}_\mu\left\{\frac{\partial^2\tau}{\partial\theta^i\partial\theta^j}\frac{\partial\rho}{\partial\theta^k} - \frac{\partial^2\rho}{\partial\theta^i\partial\theta^j}\frac{\partial\tau}{\partial\theta^k}\right\} \\ &= \mathrm{E}_\mu\left\{-\left(\frac{\partial^2\rho}{\partial\theta^i\partial\theta^k}\frac{\partial\tau}{\partial\theta^j} - \frac{\partial^2\tau}{\partial\theta^i\partial\theta^k}\frac{\partial\rho}{\partial\theta^j}\right)\right\} = T_{ikj} \end{aligned}$$

is a totally symmetric tensor. Substituting into the expression for $\Gamma_{ij,k}^{(\alpha)}$,

$$\begin{aligned}\Gamma_{ij,k}^{(\alpha)} &= \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} \left(\frac{\partial^2 f'(\rho)}{\partial\theta^i \partial\theta^j} \frac{\partial\rho}{\partial\theta^k} - \frac{\partial^2 \rho}{\partial\theta^i \partial^j} \frac{\partial f'(\rho)}{\partial\theta^k} \right) + \frac{\partial^2 f'(\rho)}{\partial\theta^i \partial\theta^j} \frac{\partial\rho}{\partial\theta^k} \right\} \\ &= \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial^2 f'(\rho)}{\partial\theta^i \partial\theta^j} \frac{\partial\rho}{\partial\theta^k} + \frac{1+\alpha}{2} \frac{\partial^2 \rho}{\partial\theta^i \partial\theta^j} \frac{\partial f'(\rho)}{\partial\theta^k} \right\}.\end{aligned}$$

This proves equation 3.10, the expression of $\Gamma_{ij,k}^{(\alpha)}$ under conjugate ρ- and τ-representations. The proof concerning $\Gamma_{ij,k}^{*(\alpha)}$ is similar. ◇

Remark 3.3.1. Amari's alpha representation (index by β here to avoid confusion) corresponds to ρ-representation ($\rho(p) = l^{(\beta)}(p),\ \tau(p) = l^{(-\beta)}(p)$) with $\alpha = 1$, or τ-representation ($\rho(p) = l^{(-\beta)}(p),\ \tau(p) = l^{(\beta)}(p)$) with $\alpha = -1$. Note also that $\alpha = 0$ corresponds to the metric-compatible Levi-Civita connection $\Gamma_{ij,k}^{(0)}$ and that

$$\Gamma_{ij,k}^{(\alpha)} = \Gamma_{ij,k}^{(0)} - \frac{\alpha}{2} T_{ijk}$$

conforms to the formal definition of α-connection (Lauritzen, 1987).

Corollary 3. *The metric $\tilde{g}_{ij}$ and the dual connections $\tilde{\Gamma}_{ij,k}^{(\alpha)}$, $\tilde{\Gamma}_{ij,k}^{*(\alpha)}$ of the statistical manifold induced by the conjugate divergence function $\mathcal{D}_{f^*,\tau}^{(\alpha)}(\theta_p, \theta_q)$ are related to those induced by $\mathcal{D}_{f,\rho}^{(\alpha)}(\theta_p, \theta_q)$ via*

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

with

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

Proof. Observe that $\rho = (f^*)'(\tau)$, from proposition 8,

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

which is just $\tilde{g}_{ij}$. To prove the second part, we follow the steps in the proof of proposition 8 to show

$$\mathrm{E}_\mu \left\{ \frac{\partial^2 \rho}{\partial\theta^i \partial^j} \frac{\partial f'(\rho)}{\partial\theta^k} - \frac{\partial^2 f'(\rho)}{\partial\theta^i \partial^j} \frac{\partial\rho}{\partial\theta^k} \right\} = \mathrm{E}_\mu \left\{ (f^*)'''(\tau) \frac{\partial\tau}{\partial\theta^i} \frac{\partial\tau}{\partial\theta^i} \frac{\partial\tau}{\partial\theta^i} \right\},$$

which is analogous to equation 3.12, and then show

$$\Gamma_{ij,k}^{(\alpha)} = \mathrm{E}_\mu \left\{ \frac{1+\alpha}{2} \left((f^*)'''(\tau) \frac{\partial \tau}{\partial \theta^i} \frac{\partial \tau}{\partial \theta^j} \frac{\partial \tau}{\partial \theta^k} \right) + \frac{\partial^2 \tau}{\partial \theta^i \partial^j} \frac{\partial \tau}{\partial \theta^k} (f^*)''(\tau) \right\}$$
$$\equiv \tilde{\Gamma}_{ij,k}^{(-\alpha)},$$

which is, by definition, the connection induced by $\mathcal{D}_{f^*,\tau}^{(-\alpha)}(\theta_p, \theta_q)$. ◇

Remark 3.3.2. Note that the statistical manifold associated with $\mathcal{D}_{f^*,\tau}^{(\alpha)}$ is the same finite-dimensional θ-manifold $\mathcal{M}_\theta$ as induced by $\mathcal{D}_{f,\rho}^{(\alpha)}$. The difference between $\mathcal{D}_{f,\rho}^{(\alpha)}(\theta_p, \theta_q)$ and $\mathcal{D}_{f^*,\tau}^{(\alpha)}(\theta_p, \theta_q)$ is due to the difference in the ρ- and τ-representations of densities, which are conjugate to each other. Corollary 3 says that the duality $\Gamma \leftrightarrow \Gamma^*$, in addition to reflecting $\theta_p \leftrightarrow \theta_q$, reflects the conjugacy between $\rho(p)$ and $\tau(p)$, that is, the dual representations of probability density function, so that

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

3.4 Natural and Expectation Parameters of Statistical Manifold. Assume now that θ, as appeared in $\rho(\theta)$ and $\tau(\theta)$, is the natural parameter of a parametric statistical model. For an exponential family, it is well known that one might as well parameterize density functions by the expectation parameter η, which is dual to the natural parameter. We generalize this duality (biorthogonality) between the natural parameter and the expectation parameter for arbitrary ρ- and τ-embedding by considering the pullback of $\mathcal{D}_{f,\rho}^{(\alpha)}$ and $\mathcal{D}_{f^*,\tau}^{(\alpha)}$ in both $\mathcal{M}_\theta$ and $\mathcal{M}_\eta$.

We now introduce the notion of ρ-affinity (a generalization of α-affinity). A family of the (denormalized) probability densities is said to be *ρ-affine* if the ρ-representation of these densities can be embedded into a finite-dimensional affine space, that is, if there exists a set of linearly independent functions $\lambda_i(\zeta)$ over the sample space $\mathcal{X} \ni \zeta$ such that

$$\rho(p(\zeta)) = \sum_i \theta^i \lambda_i(\zeta) \ . \tag{3.13}$$

Here, $\theta = [\theta^1, \ldots, \theta^n] \in \mathcal{M}_\theta$ is called the natural parameter of this ρ-affine family. For any density $p(\zeta)$, the projection of its τ-representation onto the functions $\lambda_i(\zeta)$

$$\eta_i = \int \tau(p(\zeta))\, \lambda_i(\zeta)\, d\mu, \tag{3.14}$$

forms a vector $\eta = [\eta_1, \ldots, \eta_n] \in \mathcal{M}_\eta$; η is called the expectation parameter of $p(\zeta)$.

Proposition 9. *For a ρ-affine family of densities,*

i. The functions

$$\Phi(\theta) = \int f(\rho(\theta))\, d\mu, \qquad \Phi^*(\eta) = \int f^*(\tau(\eta))\, d\mu$$

are a pair of conjugated, strictly convex functions.

ii. The natural parameter θ and the expectation parameter η form biorthogonal coordinates

$$\frac{\partial \Phi}{\partial \theta^i} = \eta_i, \qquad \frac{\partial \Phi^*}{\partial \eta_i} = \theta^i,$$

with

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

where the metric $g_{ij}(\theta)$ is positive-definite with $\tilde{g}^{ij}(\eta)$ as its matrix inverse.

iii. The dual affine connections become

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

iv. The divergence functional $\mathcal{D}^{(\alpha)}_{f,\rho}(p,q)$ becomes the divergence function $D^{(\alpha)}_{\Phi}(\theta_p, \theta_q)$.

v. The canonical divergence functional

$$A_f(\rho(p), \tau(q)) = \Phi(\theta_p) + \Phi^*(\eta_q) - \sum_i \theta_p^i \eta_{qi} \equiv \mathcal{A}_\Phi(\theta_p, \eta_q), \quad (3.15)$$

where $\mathcal{A}_\Phi(\theta, \eta)$ is the canonical divergence function as given by corollary 2.

Proof. The assumption 3.13 implies that $\frac{\partial \rho}{\partial \theta^i} = \lambda_i(\zeta)$, so from equation 3.6,

$$g_{ij} = \int f''(\rho)\, \lambda_i(\zeta)\, \lambda_j(\zeta)\, d\mu.$$

That g_{ij} is positive definite is seen by observing

$$\sum_{ij} g_{ij}\xi^i\xi^j = \int f''(\rho) \left(\sum_i \lambda_i(\zeta)\xi^i \right)^2 d\mu > 0$$

for any $\xi = [\xi^1, \ldots, \xi^n] \in \mathrm{R}^n$, due to linear independence of the λ_i's and the strict convexity of f. Now

$$\frac{\partial \Phi}{\partial \theta^i} = \int \frac{\partial f(\rho)}{\partial \theta^i}\, d\mu = \int f'(\rho)\, \frac{\partial \rho}{\partial \theta^i}\, d\mu = \int \tau(p(\zeta))\, \lambda_i(\zeta)\, d\mu = \eta_i$$

by definition 3.14. We can verify straightforwardly that

$$\frac{\partial^2 \Phi}{\partial \theta^i \partial \theta^j} = \partial \eta_i / \partial \theta^j = \int f''(\rho(\zeta))\, \frac{\partial \rho}{\partial \theta_j}\, \lambda_i(\zeta)\, d\mu = g_{ij}(\theta)$$

is positive definite, so $\Phi(\theta)$ must be a strictly convex function. Parts iv and then iii follow proposition 2 once strict convexity of $\Phi(\theta)$ is established. Differentiating both sides of equation 3.14 with respect to η_j yields

$$\delta_i^j = \int \frac{\partial \tau}{\partial \eta_j}\, \lambda_i(\zeta)\, d\mu.$$

Thus,

$$\begin{aligned}
\frac{\partial \Phi^*}{\partial \eta_i} &= \int \frac{\partial f^*(\tau)}{\partial \eta_i}\, d\mu = \int (f^*)'(\tau)\, \frac{\partial \tau}{\partial \eta_i}\, d\mu = \int \rho(p(\zeta))\, \frac{\partial \tau}{\partial \eta_i}\, d\mu \\
&= \int \left(\sum_j \theta^j \lambda_j(\zeta) \right) \frac{\partial \tau}{\partial \eta_i}\, d\mu = \sum_j \theta^j \left(\int \frac{\partial \tau}{\partial \eta_i}\, \lambda_j(\zeta)\, d\mu \right) \\
&= \sum_j \theta^j \delta_i^j = \theta^i.
\end{aligned}$$

Part ii, namely, biorthogonality of θ and η, is thus established. Evaluating

$$\begin{aligned}
\sum_i \theta^i \frac{\partial \Phi}{\partial \theta^i} - \Phi(\theta) &= \int \tau(p(\zeta;\theta)) \left(\sum_i \theta^i \lambda_i(\zeta) \right) d\mu - \int f(\rho(p(\theta))\, d\mu \\
&= \int \tau(p(\zeta;\theta))\, \rho(p(\zeta;\theta))\, d\mu - \int f(\rho(p(\zeta;\theta))\, d\mu \\
&= \int f^*(\tau(p(\zeta;\eta))\, d\mu = \Phi^*(\eta)
\end{aligned}$$

establishes the conjugacy between Φ and Φ^*, and hence strict convexity of Φ^*, as claimed in part i. Finally, substituting these expressions into equation 3.4 establishes part v. Therefore, we have proved all the relations stated in this proposition. ◇

Remark 3.4.1. This is a generalization of the results about α-affine manifolds (Amari, 1985; Amari & Nagaoka, 2000), where ρ- and τ-representations are just α- and $(-\alpha)$-representations, respectively. Proposition 9 says that when $\lambda_i(\zeta)$'s are used as the basis functions of the sample space, θ is the natural (contravariant) coordinate to express $\rho(p)$, while η is the expectation (covariant) coordinate to express $\tau(p)$. They are biorthogonal

$$\int \frac{\partial \rho}{\partial \theta^i} \frac{\partial \tau}{\partial \eta_j} \, d\mu = \delta_i^j,$$

when the ρ- (or τ)-representation of the density function is embeddable into the finite-dimensional affine space. The natural and expectation parameters are related to the τ- and to the ρ-representation, respectively, via

$$\tau(p(\zeta)) = f'\left(\sum_i \theta^i \lambda_i(\zeta)\right), \qquad \eta_i = \int f'(\rho(p))\lambda_i(\zeta)\, d\mu.$$

With the expectation parameter η, one may express the divergence functional $\mathcal{D}^{(\alpha)}_{f,\rho}(\eta_p, \eta_q)$ and obtain the corresponding metric and dual connection pair. The properties of the statistical manifold on $\mathcal{M}_\eta$ are shown by the next proposition.

Proposition 10. *The metric tensor $\hat{g}^{ij}$ and the dual connections $\hat{\Gamma}^{(\alpha)ij,k}$, $\hat{\Gamma}^{*(\alpha)ij,k}$ induced by $\mathcal{D}^{(\alpha)}_{f,\rho}(\eta_p, \eta_q)$ are related to those (expressed in lower induces) induced by $\mathcal{D}^{(\alpha)}_{f,\rho}(\theta_p, \theta_q)$ via*

$$\sum_l g_{il}(\theta)\hat{g}^{lm}(\eta) = \delta_i^m, \tag{3.16}$$

and

$$\hat{\Gamma}^{(\alpha)ij,k}(\eta) = -\sum_{l,m,n} \hat{g}^{im}(\eta)\hat{g}^{jn}(\eta)\hat{g}^{kl}(\eta)\Gamma^{(-\alpha)}_{ml,n}(\theta), \tag{3.17}$$

$$\hat{\Gamma}^{*(\alpha)ij,k}(\eta) = -\sum_{l,m,n} \hat{g}^{im}(\eta)\hat{g}^{jn}(\eta)\hat{g}^{kl}(\eta)\Gamma^{(\alpha)}_{ml,n}(\theta), \tag{3.18}$$

where η and θ are biorthogonal.

Proof. The relation 3.16 follows proposition 9. To prove equation 3.17, we write out $\hat{\Gamma}^{(\alpha)ij,k}$ following proposition 8 (note that upper- and lower-case here are pro forma):

$$\begin{aligned}
\hat{\Gamma}^{(\alpha)ij,k} &= \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial^2 \tau}{\partial \eta_i \partial \eta_j} \frac{\partial \rho}{\partial \eta_k} + \frac{1+\alpha}{2} \frac{\partial^2 \rho}{\partial \eta_i \partial \eta_j} \frac{\partial \tau}{\partial \eta_k} \right\} \\
&= \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} \left(\sum_l \frac{\partial \rho}{\partial \theta^l} \frac{\partial \theta^l}{\partial \eta_k} \right) \left(\sum_m \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial}{\partial \theta^m} \frac{\partial \tau}{\partial \eta_j} \right) \right. \\
&\qquad \left. + \frac{1+\alpha}{2} \left(\sum_l \frac{\partial \tau}{\partial \theta^l} \frac{\partial \theta^l}{\partial \eta_k} \right) \left(\sum_m \frac{\partial \theta^m}{\partial \eta_i} \frac{\partial}{\partial \theta^m} \frac{\partial \rho}{\partial \eta_j} \right) \right\} \\
&= \sum_{l,m} \frac{\partial \theta^l}{\partial \eta_k} \frac{\partial \theta^m}{\partial \eta_i} \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial \rho}{\partial \theta^l} \frac{\partial}{\partial \theta^m} \left(\sum_n \frac{\partial \tau}{\partial \theta^n} \frac{\partial \theta^n}{\partial \eta_j} \right) \right. \\
&\qquad \left. + \frac{1+\alpha}{2} \frac{\partial \tau}{\partial \theta^l} \frac{\partial}{\partial \theta^m} \left(\sum_n \frac{\partial \rho}{\partial \theta^n} \frac{\partial \theta^n}{\partial \eta_j} \right) \right\} \\
&= \sum_{l,m,n} \frac{\partial \theta^l}{\partial \eta_k} \frac{\partial \theta^m}{\partial \eta_i} \left(\frac{\partial \theta^n}{\partial \eta_j} \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial \rho}{\partial \theta^l} \frac{\partial^2 \tau}{\partial \theta^m \partial \theta^n} \right. \right. \\
&\qquad \left. + \frac{1+\alpha}{2} \frac{\partial \tau}{\partial \theta^l} \frac{\partial^2 \rho}{\partial \theta^m \partial \theta^n} \right\} \\
&\quad \left. + \left(\frac{\partial}{\partial \theta^m} \frac{\partial \theta^n}{\partial \eta_j} \right) \mathrm{E}_\mu \left\{ \frac{1-\alpha}{2} \frac{\partial \rho}{\partial \theta^l} \frac{\partial \tau}{\partial \theta^n} + \frac{1+\alpha}{2} \frac{\partial \tau}{\partial \theta^l} \frac{\partial \rho}{\partial \theta^n} \right\} \right) \\
&= \sum_{l,m,n} \frac{\partial \theta^l}{\partial \eta_k} \frac{\partial \theta^m}{\partial \eta_i} \left(\frac{\partial \theta^n}{\partial \eta_j} \Gamma^{(\alpha)}_{mn,l} + \frac{\partial}{\partial \theta^m} \frac{\partial \theta^n}{\partial \eta_j} g_{nl} \right).
\end{aligned}$$

Since

$$\begin{aligned}
\sum_n \left(\frac{\partial}{\partial \theta^m} \frac{\partial \theta^n}{\partial \eta_j} \right) g_{nl} &= \sum_n \frac{\partial g^{jn}}{\partial \theta^m} g_{nl} = -\sum_n \frac{\partial g_{nl}}{\partial \theta^m} g^{jn} \\
&= -\sum_n g^{jn} \left(\Gamma^{(\alpha)}_{mn,l} + \Gamma^{(-\alpha)}_{ml,n} \right),
\end{aligned}$$

where the last step is from

$$\frac{\partial g_{nl}}{\partial \theta^m} = \Gamma^{(\alpha)}_{mn,l} + \Gamma^{*(\alpha)}_{ml,n} = \Gamma^{(\alpha)}_{mn,l} + \Gamma^{(-\alpha)}_{ml,n},$$

assertion 3.17 is proved after direct substitution. Observing the duality $\hat{\Gamma}^{*(\alpha)ij,k} = \hat{\Gamma}^{(-\alpha)ij,k}$ leads to equation 3.18. ◇

Remark 3.4.2. The relation between g and Γ in their subscript and superscript forms is analogous to that stated by proposition 5. However, note the conjugacy of α in $\hat{\Gamma}^{(\alpha)ij,k} \leftrightarrow \Gamma^{(-\alpha)}_{ml,n}$ correspondence, due to the change between θ- and η-coordinates, both under the ρ-representation. On the other hand, similar to corollary 3, the metric $\bar{g}^{ij}$ and the dual affine connection $\bar{\Gamma}^{(\alpha)ij,k}$, $\bar{\Gamma}^{*(\alpha)ij,k}$ of the statistical manifold (denoted using bar) induced by the conjugate divergence functions $\mathcal{D}^{(\alpha)}_{f^*,\tau}(\eta_p, \eta_q)$ are related to those (denoted using hat) induced by $\mathcal{D}^{(\alpha)}_{f^*,\rho}(\eta_p, \eta_q)$ via

$$\bar{g}^{ij}(\eta) = \hat{g}^{ij}(\eta),$$

with

$$\bar{\Gamma}^{(\alpha)ij,k}(\eta) = \hat{\Gamma}^{(-\alpha)ij,k}(\eta), \qquad \bar{\Gamma}^{*(\alpha)ij,k}(\eta) = \hat{\Gamma}^{(\alpha)ij,k}(\eta).$$

3.5 Divergence Functional from Generalized Mean. When f is, in addition to being strictly convex, strictly monotone increasing, we may set $\rho = f^{-1}$, so that the divergence functional becomes

$$\mathcal{D}^{(\alpha)}_{\rho}(p, q) = \frac{4}{1-\alpha^2} \int \left(\frac{1-\alpha}{2} p + \frac{1+\alpha}{2} q \right. \\ \left. - \rho^{-1}\left(\frac{1-\alpha}{2} \rho(p) + \frac{1+\alpha}{2} \rho(q) \right) \right) d\mu. \tag{3.19}$$

Note that for $\alpha \in [-1, 1]$,

$$M^{(\alpha)}_{\rho}(p, q) \equiv \rho^{-1}\left(\frac{1-\alpha}{2} \rho(p) + \frac{1+\alpha}{2} \rho(q) \right)$$

defines a generalized mean ("quasi-linear mean" by Hardy, Littlewood, & Pólya, 1952) associated with a concave and monotone function ρ: $\mathrm{R}_+ \to \mathrm{R}$. Viewed in this way, the divergence is related to the departure of the linear (arithmetic) mean from a quasi-linear mean induced by a nonlinear function with nonzero concavity/convexity.

Example 3.5.1. Take $\rho(p) = \log p$, then $M^{(\alpha)}_{\rho}(p, q) = p^{\frac{1-\alpha}{2}} q^{\frac{1+\alpha}{2}}$, and $\mathcal{D}^{(\alpha)}_{\rho}(p, q)$ is the α-divergence (see equation 1.2). For a general concave ρ,

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

is an immediate generalization of the extended Kullback-Leibler divergence in equation 1.1.

To further explore the divergence functionals associated with the quasi-linear mean operator, we impose a homogeneity requirement, such that the divergence is invariant after scaling ($\kappa \in \mathrm{R}_+$):

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

Proposition 11. *The only measure-invariant divergence functional associated with quasi-linear mean operator $M^{(\alpha)}_{\rho}$ is a two-parameter family,*

$$\mathcal{D}^{(\alpha,\beta)}(p, q) \equiv \frac{4}{1-\alpha^2}\frac{2}{1+\beta}\int\left(\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q - \left(\frac{1-\alpha}{2}p^{\frac{1-\beta}{2}} + \frac{1+\alpha}{2}q^{\frac{1-\beta}{2}}\right)^{\frac{2}{1-\beta}}\right) d\mu, \tag{3.20}$$

which results from the alpha-representation (indexed by β here) $\rho(p) = l^{(\beta)}(p)$ as given by equation 1.8. Here $(\alpha, \beta) \in [-1, 1] \times [-1, 1]$, and the factor $2/(1 + \beta)$ is introduced to make $\mathcal{D}^{(\alpha,\beta)}(p, q)$ well defined for $\beta = -1$.

Proof. This homogeneity requirement implies that

$$\rho^{-1}\left(\frac{1-\alpha}{2}\rho(\kappa p) + \frac{1+\alpha}{2}\rho(\kappa q)\right) = \kappa\rho^{-1}\left(\frac{1-\alpha}{2}\rho(p) + \frac{1+\alpha}{2}\rho(q)\right).$$

By a lemma in Hardy et al. (1952, p. 68), the general solution to the above functional equation is

$$\rho(t) = \begin{cases} a\,t^s + b & s \neq 0 \\ a \log t + b & s = 0, \end{cases}$$

with corresponding

$$M^{(\alpha)}_s(p, q) = \left(\frac{1-\alpha}{2}p^s + \frac{1+\alpha}{2}q^s\right)^{\frac{1}{s}}, \quad M^{(\alpha)}_0(p, q) = p^{\frac{1-\alpha}{2}}\, q^{\frac{1+\alpha}{2}}.$$

Here a, b, s are all constants. Strict concavity of ρ requires $0 \leq s \leq 1$ and $a > 0$. Since it is easily verified $\mathcal{D}^{(\alpha)}_{\rho} = \mathcal{D}^{(\alpha)}_{a\rho+b}$, without loss of generality, we have $\rho(p) = l^{(\beta)}(p)$, $\beta \in [-1, 1]$ where $s = \frac{1-\beta}{2}$. This gives rise to equation 3.20. ◇

Proposition 12 (corollary to Proposition 7). *The two-parameter family of divergence functions $\mathcal{D}^{(\alpha,\beta)}(\theta_p, \theta_q)$ induces a statistical manifold with Fisher information as its metric and generic alpha-connections as its dual connection pair,*

$$g_{ij} = E_p\left\{\frac{\partial \log p}{\partial \theta^i}\frac{\partial \log p}{\partial \theta^j}\right\}$$

$$\Gamma_{ij,k}^{(\alpha,\beta)} = E_p \left\{ \frac{\partial^2 \log p}{\partial\theta^i \partial\theta^j} \frac{\partial \log p}{\partial\theta^k} + \frac{1-\alpha\beta}{2} \frac{\partial \log p}{\partial\theta^i} \frac{\partial \log p}{\partial\theta^j} \frac{\partial \log p}{\partial\theta^k} \right\},$$

$$\Gamma_{ij,k}^{*(\alpha,\beta)} = E_p \left\{ \frac{\partial^2 \log p}{\partial\theta^i \partial\theta^j} \frac{\partial \log p}{\partial\theta^k} + \frac{1+\alpha\beta}{2} \frac{\partial \log p}{\partial\theta^i} \frac{\partial \log p}{\partial\theta^j} \frac{\partial \log p}{\partial\theta^k} \right\}.$$

Proof. Applying formulas 3.6 to 3.8 to the measure-invariant divergence functional $\mathcal{D}_\rho^{(\alpha)}(p,q)$ with $\rho(p) = \log p$ and $f = \rho^{-1}$ gives rise to the desired result. ◇

Remark 3.5.2. This two-parameter family of affine connections $\Gamma_{ij,k}^{(\alpha,\beta)}$, indexed now by the numerical product $\alpha\beta \in [-1,1]$, is actually in the generic form of an alpha-connection,

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

with biduality compactly expressed as

$$\Gamma_{ij,k}^{*(\alpha,\beta)} = \Gamma_{ij,k}^{(-\alpha,\beta)} = \Gamma_{ij,k}^{(\alpha,-\beta)}. \tag{3.21}$$

The parameters $\alpha \in [-1,1]$ and $\beta \in [-1,1]$ reflect referential duality and representational duality, respectively. Among this two-parameter family, the Levi-Civita connection results when either α or β equals 0. When $\alpha = \pm 1$ or $\beta = \pm 1$, each case reduces to the one-parameter version of the generic alpha-connection. The family $\mathcal{D}^{(\alpha,\beta)}$ is then a generalization of Amari's alpha-divergence, equation 1.2, with

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

where the last equation is due to $\lim_{\beta \to 1} M_s^{(\alpha)} = M_0^{(\alpha)} = p^{\frac{1-\alpha}{2}} q^{\frac{1+\alpha}{2}}$. On the other hand, when $\beta \to -1$, we have the interesting asymptotic relation,

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

where $E^{(\alpha)}$ was the Jensen difference, equation 2.5, discussed by Rao (1987).

3.6 Parametric Family of Csiszár's f-Divergence. The fact (see Proposition 12) that our two-parameter family of divergence functions $\mathcal{D}^{(\alpha,\beta)}$ actually induces a one-dimensional family of alpha-connection is by no means surprising. This is because $\mathcal{D}^{(\alpha,\beta)}$ obviously falls within Csiszár's

f-divergence (see equation 1.3), the generic form for measure-invariant divergence, where

$$f^{(\alpha,\beta)}(t) = \frac{8}{(1-\alpha^2)(1+\beta)} \left(\frac{1-\alpha}{2} + \frac{1+\alpha}{2} t \right. \\ \left. - \left(\frac{1-\alpha}{2} + \frac{1+\alpha}{2} t^{\frac{1-\beta}{2}} \right)^{\frac{2}{1-\beta}} \right), \tag{3.22}$$

is now a two-parameter family with $f^{(\alpha,\beta)}(1) = 0$; $(f^{(\alpha,\beta)})'(1) = 0$; $(f^{(\alpha,\beta)})''(1) = 1$. That the alpha index is given by the product $\alpha\beta$ in this case follows explicitly from calculating $(f^{(\alpha,\beta)})'''(1)$ using equation 1.5. What is interesting in this regard is the distinct roles played by α (for reference duality) and by β (for representational duality). The parameters $(\alpha, \beta) \in [0,1] \times [0,1]$ form an interesting topological structure of a Moebius band in the space of divergence functions, all with identical Fisher information and the family of alpha-connections.

We may generalize Csiszár's f-divergence to construct a family of measure-invariant divergence functional in the following way. Given a smooth, strictly convex function $f(t)$, construct the family (for $\gamma \in$ R)

$$G_f^{(\gamma)}(t) = \frac{4}{1-\gamma^2} \left(\frac{1-\gamma}{2} f(1) + \frac{1+\gamma}{2} f(t) - f\left(\frac{1-\gamma}{2} + \frac{1+\gamma}{2} t \right) \right),$$

with $G_f^{(-1)}(t) = g(t)$ as given in equation 1.6. It is easy to verify that for an arbitrary γ, $G_f^{(\gamma)}$ is a proper Csiszár's function with $G_f^{(\gamma)}(1) = 0$, $(G_f^{(\gamma)})'(1) = 0$, and that

$$(G_f^{(\gamma)})''(1) = f''(1), \qquad (G_f^{(\gamma)})'''(1) = \frac{\gamma+3}{2} f'''(1),$$

so the statistical manifold generated by $G_f^{(\gamma)}$ has the same metric as that generated by f but a family of parameterized alpha-connections. If we take $f(t) = f^{(\alpha)}(t)$ as in equation 1.4, then $G^{(\gamma,\alpha)}$ will generate a two-parameter family of alpha-connections with the effective alpha value $3+(\alpha-3)(\gamma+3)/2$. We note in passing that repeating this process, by having $G^{(\gamma,\alpha)}$ now take the role of f, may lead to nested (e.g., two-, three- parameter) families of alpha-connections.

4 General Discussion

This article introduced several families of divergence functions and functionals all based on the fundamental inequality of an arbitrary smooth and strictly convex function. In the finite-dimensional case, the convex mixture parameter, α, which reflects reference duality, turns out to correspond to the α parameter in the one-parameter family of α-connection in the sense of Lauritzen (1987), which includes the flat connections ($\alpha = \pm 1$) induced by Bregman divergence. The biorthogonal coordinates related to the inducing convex function and its conjugate (Amari's dual potentials) reflect representational duality. In the infinite-dimensional cases, with the notion of conjugate (i.e., ρ- and τ-) embeddings of density functions, the form of the constructed divergence functionals generalizes the familiar ones (α-divergence and f-divergence). The resulting α-connections, equation 3.7, or equivalently, equation 3.10, have the most generalized yet explicit form found in the literature. When densities are ρ-affine, they specialize to α-connections in the finite-dimensional vector space mentioned above. When measure-invariance is imposed, they specialize to the family of alpha-connections proper, but now with two parameters—one reflecting reference duality and the other representational duality. These findings will enrich the theory of information geometry and make it applicable to finite-dimensional vector space (not necessarily of parameters of probability densities) as well as to infinite-dimensional functional space (not necessarily of normalized density functions).

In terms of neural computation, to the extent that alpha-divergence and alpha-connections generate deep analytic insights (e.g., Amari, Ikeda, & Shimokawa, 2001; Takeuchi & Amari, submitted), these theoretical results may help facilitate those analyses by clarifying the meaning of duality in projection-based algorithms. Previously, alpha-divergence, in its extended form (see equation 1.2), was shown (Amari & Nagaoka, 2000) to be the canonical divergence for the α-affine family of densities (densities that, under the α-representation $l^{(\alpha)}$, are spanned by an affine subspace). Therefore, for a given α value, there is only one such family that induces the flat (α-)connection with all components zero when expressed in suitable coordinates (as special cases, $\Gamma^{(1)} = 0$ for the exponential family and $\Gamma^{(-1)} = 0$ for the mixture family). This is slightly different from the view of Zhu and Rohwer (1995, 1997) who, in their Bayesian inference framework, simply treated α as a parameter in the entire class of (α-)divergence (between any two densities) which yields, through Eguchi relation, flat connections only when $\alpha = \pm 1$. These apparently disparate interpretations, despite being subtly so, have now been straightened out. The current framework points out two related but different senses of duality in information geometry: representational duality and reference duality. Further, it has been clarified how the same one-parameter family of dual alpha-connections actually may embody both kinds of dualities. Future research will illuminate how this notion of biduality in characterizing the asymmetric difference of two density functions or two parameters may have captured the very essence of computational algorithms of inference, optimization, and adaptation.

Acknowledgments

I thank Huaiyu Zhu for introducing the topic of divergence functions and information geometry. Part of this work was presented to the International Conference on Information Geometry and Its Applications, Pescara, 2002, where I benefited from direct feedback and extensive discussions with many conference participants, including S. Amari and S. Eguchi in particular. Discussions with Matt Jones at the University of Michigan have helped improve the presentation.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

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*. New York: Springer-Verlag.

Amari, S. (1991). Dualistic geometry of the manifold higher-order neurons. *Neural Networks*, *4*, 443–451.

Amari, S. (1995). Information geometry of EM and EM algorithms for neural networks. *Neural Networks*, *8*, 1379–1408.

Amari, S., Kurata, K., & Nagaoka, H. (1992). Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, *3*, 260–271.

Amari, S., Ikeda, S., & Shimokawa, H. (2001). Information geometry and mean field approximation: The α-projection approach. In M. Opper, & D. Saad (Eds), *Advanced mean field methods—Theory and practice*, (pp. 241–257). Cambridge, MA: MIT Press.

Amari, S., & Nagaoka, H. (2000). *Method of information geometry*. New York: Oxford University Press.

Bauschke, H. H., Borwein, J. M., & Combettes, P. L. (2002). Bregman monotone optimization algorithms. CECM Preprint 02:184. Available on-line: http://www.cecm.sfu.ca/preprints/2002pp.html.

Bauschke, H. H., & Combettes, P. L. (2002). Iterating Bregman retractions. CECM Preprint 02:186. Available on-line: http://www.cecm.sfu.ca/preprints/2002pp.html.

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.

Chentsov, N. N. (1982). *Statistical decision rules and optimal inference.* Providence, RI: AMS, 1982.

Csiszár, I. (1967). On topical properties of *f*-divergence. *Studia Mathematicarum Hungarica*, *2*, 329–339.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (2002). *Duality and auxiliary functions for Bregman distances* (Tech. Rep. No. CMU-CS-01-109). Pittsburgh, PA: School of Computer Science, Carnegie Mellon University.

Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. *Annals of Statistics*, *11*, 793–803.

Eguchi, S. (1992). Geometry of minimum contrast. *Hiroshima Mathematical Journal*, *22*, 631–647.

Eguchi, S. (2002). *U-boosting method for classification and information geometry*. Paper presented at the SRCCS International Statistical Workshop, Seoul National University, June.

Hardy, G., Littlewood, J. E., & Pólya, G. (1952). *Inequalities*. Cambridge: Cambridge University Press.

Ikeda, S., Amari, S., & Nakahara, H. (1999) Convergence of the wake-sleep algorithm. In M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems, 11* (pp. 239–245). Cambridge, MA: MIT Press.

Kaas, R. E., & Vos, P. W. (1997). *Geometric foundation of asymptotic inference*. New York: Wiley.

Kurose, T. (1994). On the divergences of 1-conformally flat statistical manifolds. *Töhoko Mathematical Journal*, *46*, 427–433.

Lafferty, J., Della Pietra, S., & Della Pietra, V. (1997). Statistical learning algorithms based on Bregman distances. In *Proceedings of 1997 Canadian Workshop on Information Theory*, pp. 77–80. Toronto, Canada: Fields Institute.

Lauritzen, S. (1987). Statistical manifolds. In S. Amari, O. Barndorff-Nielsen, R. Kass, S. Lauritzen, and C. R. Rao (Eds.), *Differential geometry in statistical inference* (pp. 163–216). Hayward, CA: Institute of Mathematical Statistics.

Lebanon, G., & Lafferty, J. (2002). Boosting and maximum likelihood for exponential models. In T. G. Dietterich, S. Becker, & Z. Ghahramani (eds.) *Advances in neural information processing systems, 14* (pp. 447–454). Cambridge, MA: MIT Press.

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.

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.

Mihoko, M., & Eguchi, S. (2002). Robust blink source separation by beta-divergence. *Neural Computation*, *14*, 1859–1886.

Rao, C. R. (1987). Differential metrics in probability spaces. In S. Amari, O. Barndorff-Nielsen, R. Kass, S. Lauritzen, & C. R. Rao (Eds.), *Differential geometry in statistical inference*. (pp. 217–240). Hayward, CA: Institute of Mathematical Statistics.

Rockafellar, R. T. (1970). *Convex analysis*. Princeton, NJ: Princeton University Press.

Shima, H. (1978). Compact locally Hessian manifolds. *Osaka Journal of Mathematics*, *15*, 509–513.

Shima, H., & Yagi, K. (1997). Geometry of Hessian manifolds. *Differential Geometry and Its Applications, 7*, 277–290.

Takeuchi, J., & Amari, S. (submitted). α-Parallel prior and its properties. *IEEE Transaction on Information Theory*. Manuscript under review.

Uohashi, K., Ohara, A., & Fujii, T. (2000). 1-Conformally flat statistical submanifolds. *Osaka Journal of Mathematics*, *37*, 501–507.

Zhu, H. Y., & Rohwer, R. (1995). Bayesian invariant measurements of generalization. *Neural Processing Letter*, *2*, 28–31.

Zhu, H. Y., & Rohwer, R. (1997) Measurements of generalisation based on information geometry. In S. W. Ellacott, J. C. Mason, & I. J. Anderson (Eds.), *Mathematics of neural networks: Model algorithms and applications* (pp 394–398). Norwell, MA: Kluwer.

Received October 29, 2002; accepted June 26, 2003.