Ann Inst Stat Math (2010) 62:873–896
DOI 10.1007/s10463-008-0191-3

Dual connections in nonparametric classical information geometry

M. R. Grasselli

Received: 8 September 2006 / Revised: 29 May 2008 / Published online: 15 August 2008

Abstract We construct an infinite-dimensional information manifold based on exponential Orlicz spaces without using the notion of exponential convergence. We then show that convex mixtures of probability densities lie on the same connected component of this manifold, and characterize the class of densities for which this mixture can be extended to an open segment containing the extreme points. For this class, we define an infinite-dimensional analogue of the mixture parallel transport and prove that it is dual to the exponential parallel transport with respect to the Fisher information. We also define α-derivatives and prove that they are convex mixtures of the extremal (± 1)-derivatives.

Keywords Information geometry · Statistical manifiold · Fisher metric · Orlicz spaces · Amari–Nagaoka duality

1 Introduction

Information geometry is the branch of probability theory dedicated to provide families of probability distributions with differential geometrical structures. One then uses the tools of differential geometry in order to have a clear and intuitive picture, as well as rigor, in a variety of practical applications ranging from neural networks to statistical estimation, from mathematical finance to nonequilibrium statistical mechanics (see Sollich et al. 2001).

M. R. Grasselli (✉)
Department of Mathematics and Statistics, McMaster University,
1280 Main Street West, Hamilton, ON L8S 4K1, Canada
e-mail: grasselli@math.mcmaster.ca

It was just over half a century ago that the Fisher information

$$g_{ij} = \int \frac{\partial \log p(x,\theta)}{\partial \theta^i} \frac{\partial \log p(x,\theta)}{\partial \theta^j} p(x,\theta)\mathrm{d}x \tag{1}$$

was independently suggested by Rao (1945) and Jeffreys (1946) as a Riemannian metric for a parametric statistical model $\{p(x,\theta), \theta = (\theta^1, \ldots, \theta^n)\}$. The Riemannian geometry of statistical models was then studied as a mathematical curiosity for some years, with an emphasis in the geodesic distances associated with the Levi–Civita connection for this metric. A greater amount of attention was devoted to the subject after Efron (1975) introduced the concept of statistical curvature, pointing out its importance to statistical inference, as well as implicitly using a new affine connection, which would be known as the exponential connection. This exponential connection, together with another connection, later to be called the mixture connection, were further investigated by Dawid (1975). The work of several years on the geometric aspects of parametric statistical models culminated with the masterful account in Amari (1985), where the whole finite dimensional differential-geometric machinery is employed, including a one-parameter family of α-connections, the essential concept of duality and the notions of statistical divergence, projections and minimization procedures. Among the successes of the research at these early stages one could single out the rigidity of the geometric structures, such as the result concerning the uniqueness of the Fisher metric with respect to monotonicity in Čencov (1982) and Amari's result concerning the uniqueness of the α-connections introduced by invariant statistical divergences. The ideas were then extensively used in statistics, in particular higher order asymptotic inference and curved exponential models (see Kass and Vos 1997).

A different line of investigation in Information Geometry took off in the nineties: the search for a fully fledge infinite dimensional manifold of probability measures. As for motivations for this quest, one had, on the practical side, the need to deal with nonparametric models in statistics, where the shape of the underlying distribution is not assumed to be known. On a more fundamental level, there was the desire of having parametric statistical manifolds defined simply as finite dimensional submanifolds of a well defined manifold of all probability measures on a sample space. The motivating idea was already in Dawid (1975) and was also addressed by Amari (1985). The first sound mathematical construction, however, is due to Pistone and Sempi (1995). Given a probability space $(\Omega, \mathcal{F}, \mu)$, they showed how to construct a Banach manifold $\mathcal{M}$ of all probability measures equivalent to μ. The Banach space used as generalized coordinates was the Orlicz space L^{Φ_1}, where Φ_1 is an exponential Young function. In a subsequent work, Pistone and Rogantin (1999) analyzed further properties of this manifold, in particular the concepts of orthogonality and submanifolds. In Sect. 3, we review their construction and present an alternative proof of the main result in Pistone and Sempi (1995), namely that the collection of covering neighborhoods $\mathcal{U}_p$ and charts e_p^{-1} form an affine C^∞-atlas for $\mathcal{M}$. The crux is Proposition 1, where we show that the image of overlapping neighborhoods under any chart e_p^{-1} is open in the topology of the target space L^{Φ_1}.

The next step in this development was the Gibilisco and Pistone (1998) definition of the exponential connection as the natural connection induced by the use of L^{Φ_1}. These authors then propose a mixture connection acting on the pretangent bundle ${}^*T\mathcal{M}$ and prove that it is dual to the exponential connection, in the sense of duality for Banach spaces. They further define the α-connections through generalized α-embeddings and show that the formal relation between the exponential, mixture and α-connections are the same as in the parametric case, that is

$$\nabla^{(\alpha)} = \frac{1+\alpha}{2}\nabla^{(e)} + \frac{1-\alpha}{2}\nabla^{(m)}. \tag{2}$$

We argue, however, that neither of these two results (duality for the exponential and mixture connection and α-connections as convex mixture of them) is a proper generalization of the corresponding parametric ones, the reason being twofold. First, Banach space duality is not Amari–Nagaoka duality. The latter refers to a metric being preserved by the joint action of two parallel transports, which are then said to be dual (see (25)). Secondly, all the α-connections in the parametric case act on the tangent bundle, whereas in Gibilisco and Pistone (1998) each of them acts on its own bundle-connection pair, making a formula like (2) at least difficult to interpret.

In order to address these problems, we define in Sect. 4 an isomorphism $\tau^{(-1)}$ of tangent spaces, which satisfy the Amari–Nagaoka duality relation with respect to the Fisher metric when paired with the exponential parallel transport $\tau^{(1)}$. However, it turns out that our map $\tau^{(-1)}$ can only be rigorously defined between points q_1 and q_2 in $\mathcal{M}$ whose ratio is a bounded random variable. Proposition 3 then characterizes the extended convex mixtures between such points.

In Sect. 5, we rearrange the definitions of Gibilisco and Pistone (1998) in order to have α-derivatives all acting on the same tangent bundle, but defined only for a restricted class of tangent vectors. We then show that the desired relation (2) holds for our definitions. We then finalize the paper by showing that the α-auto-parallel curves between two points whose ratio is a bounded function belong to the connected component $\mathcal{E}(p)$.

2 Orlicz spaces

We present here the aspects of the theory of Orlicz spaces that will be relevant for the construction of the information manifold. For more comprehensive accounts, as well as for the proofs of all statements in this section, the reader is referred to the monographs of Rao and Ren (1991) and Krasnosel′skiĭ and Rutickiĭ (1961).

The general theory of Orlicz spaces is developed around the concept of a *Young* function, that is, a convex function $\Phi : \mathbb{R} \mapsto \overline{\mathbb{R}}^+$ satisfying

(i) $\Phi(x) = \Phi(-x), \quad x \in \mathbb{R}$,
(ii) $\Phi(0) = 0$,
(iii) $\lim_{x \mapsto \infty} \Phi(x) = +\infty$.

For applications in information geometry, it is enough to consider Young functions of the form

$$\Phi(x) = \int_0^{|x|} \phi(t)\mathrm{d}t, \quad x \geq 0, \tag{3}$$

where $\phi : [0, \infty) \mapsto [0, \infty)$ is nondecreasing, continuous and such that $\phi(0) = 0$ and $\lim_{x\to\infty} \phi(x) = +\infty$. Young functions of this type include the monomials $|x|^r/r$, for $1 < r < \infty$, and the following examples arising in information geometry:

$$\Phi_1(x) = \cosh x - 1, \tag{4}$$

$$\Phi_2(x) = e^{|x|} - |x| - 1, \tag{5}$$

$$\Phi_3(x) = (1 + |x|)\log(1 + |x|) - |x| \tag{6}$$

(in the sequel, Φ_1,Φ_2 and Φ_3 will always refer to these three particular functions, with other symbols being used to denote generic Young functions).

When a Young function Φ is given in the form (3) we can define its complementary (conjugate) function as the Young function Ψ given by

$$\Psi(y) = \int_0^{|y|} \psi(t)\mathrm{d}t, \quad y \geq 0, \tag{7}$$

where ψ is the inverse of ϕ. One can verify that (Φ_2, Φ_3) and $(|x|^r/r, |x|^s/s)$, with $r^{-1} + s^{-1} = 1$, are examples of complementary pairs. For a general Young function Φ, the complementary function Ψ is given less constructively by

$$\Psi(y) = \sup\{x \geq 0 : x|y| - \Phi(x)\}. \tag{8}$$

There are many different ways of introducing a partial order on the class of Young functions. A particularly straightforward one is to say that a Young function Ψ_2 is *stronger* than another Young function Ψ_1, denoted by $\Psi_1 \prec \Psi_2$, if there exist a constant $a > 0$ such that

$$\Psi_1(x) \leq \Psi_2(ax), \quad x \geq x_0, \tag{9}$$

for some $x_0 \geq 0$ (depending on a). For example, one can verify that

$$|x| \prec \Phi_3 \prec \frac{|x|^r}{r} \prec \frac{|x|^s}{s} \prec \Phi_2 \tag{10}$$

whenever $1 < r \leq s < \infty$. Two Young functions Ψ_1 and Ψ_2 are said to be *equivalent* if $\Psi_1 \prec \Psi_2$ and $\Psi_2 \prec \Psi_1$, that is, if there exist real numbers $0 < c_1 \leq c_2 < \infty$ and $x_0 \geq 0$ such that

$$\Psi_1(c_1 x) \leq \Psi_2(x) \leq \Psi_1(c_2 x), \quad x \geq x_0. \tag{11}$$

For example, the functions Φ_1 and Φ_2 are equivalent, both being of exponential type.

Now let (Ω, Σ, P) be a probability space. The *Orlicz class* associated with a Young function Φ is defined as

$$\tilde{L}^{\Phi}(P) = \left\{ f : \Omega \mapsto \overline{\mathbb{R}}, \text{measurable} : \int_{\Omega} \Phi(f) \mathrm{d}P < \infty \right\}. \tag{12}$$

Since P is a finite measure, the Banach space $L^{\infty}(\Omega, \Sigma, P)$ of essentially bounded random variables is easily seen to be a subset of $\tilde{L}^{\Phi}(P)$ for any Young function Φ. It is easy to see that $\tilde{L}^{\Phi}(P)$ is a convex set and that $h \in \tilde{L}^{\Phi}(P)$ and $|f| \leq |h|$ imply that $f \in \tilde{L}^{\Phi}(P)$. However, in general, $\tilde{L}^{\Phi}(P)$ is *not* a vector space, which leads to the definition of the *Orlicz space* associated with a Young function Φ as

$$L^{\Phi}(P) = \left\{ f : \Omega \mapsto \overline{\mathbb{R}}, \text{measurable} : \int_{\Omega} \Phi(\alpha f) \mathrm{d}P < \infty, \text{ for some } \alpha > 0 \right\}, \tag{13}$$

furnished with the *Luxembourg* norm (see Rao and Ren 1991, p. 67)

$$N_{\Phi}(f) = \inf \left\{ k > 0 : \int_{\Omega} \Phi \left(\frac{f}{k} \right) \mathrm{d}P \leq 1 \right\}. \tag{14}$$

or with the equivalent *Orlicz* norm (see Rao and Ren 1991, p. 61)

$$\|f\|_{\Phi} = \sup \left\{ \int_{\Omega} |fg| \mathrm{d}P : g \in L^{\Psi}(P), \int_{\Omega} \Psi(g) \mathrm{d}P \leq 1 \right\}, \tag{15}$$

where Ψ is the complementary Young function to Φ. We observe for later use that $\int_{\Omega} \Phi(f) \mathrm{d}P \leq 1$ iff $N_{\Phi}(f) \leq 1$ (see Rao and Ren 1991, p. 54).

A key ingredient in the analysis of Orlicz spaces is the generalized Hölder inequality (see Rao and Ren 1991, p. 58). If Φ and Ψ are complementary Young functions, $f \in L^{\Phi}(P)$, $g \in L^{\Psi}(P)$, then

$$\int_{\Omega} |fg| \mathrm{d}P \leq 2N_{\Phi}(f)N_{\Psi}(g). \tag{16}$$

It follows that each element $f \in L^{\Phi}(P)$ defines a continuous linear functional on $L^{\Psi}(P)$, so that if we denote its topological dual by $\left(L^{\Psi}\right)^*$ we obtain the continuous injection $L^{\Phi} \subset \left(L^{\Psi}\right)^*$ for any pair of complementary Young functions.

If $\Psi_2 \prec \Psi_1$ then there exist a constant k such that $N_{\Psi_2}(\cdot) \leq kN_{\Psi_1}(\cdot)$ and therefore $L^{\Psi_1}(P) \subset L^{\Psi_2}(P)$ (see Rao and Ren 1991, p. 155). For instance, due to (10) we obtain that for $1 < r \leq s < \infty$

$$L^{\Phi_2} \subset L^s \subset L^r \subset L^{\Phi_3} \subset L^1, \tag{17}$$

where L^r, $r \geq 1$ denote the usual Lebesgue spaces on (Ω, Σ, P), which coincide with the Orlicz space defined by the Young functions $|x|^r/r$, $r \geq 1$. If two Young functions are equivalent, then the Orlicz spaces associated with them are isomorphic, that is, they coincide as sets and have equivalent norms. For example, we have that $L^{\Phi_1}(P) = L^{\Phi_2}(P)$.

3 The Pistone–Sempi information manifold

We start by reviewing the construction of an infinite dimensional information manifold along the lines of Pistone and Sempi (1995); Pistone and Rogantin (1999); Gibilisco and Pistone (1998). Consider the set $\mathcal{M}$ of all densities of probability measures equivalent to a reference measure μ, that is,

$$\mathcal{M} \equiv \mathcal{M}(\Omega, \Sigma, \mu) = \{f : \Omega \mapsto \mathbb{R}, \text{measurable} : f > 0 \text{ a.e. and } \int_\Omega f \mathrm{d}\mu = 1\}.$$

For each point $p \in \mathcal{M}$, let $L^{\Phi_1}(p)$ be the exponential Orlicz space with norm $N_p^{\Phi_1}(\cdot)$ over the probability space $(\Omega, \Sigma, p\mathrm{d}\mu)$ and consider its closed subspace of p-centred random variables

$$B_p = \{u \in L^{\Phi_1}(p) : \int_\Omega up\mathrm{d}\mu = 0\} \tag{18}$$

as the coordinate Banach space.

In probabilistic terms, the set $L^{\Phi_1}(p)$ corresponds to random variables whose moment generating function with respect to the probability $p\mathrm{d}\mu$ is finite on a neighborhood of the origin (see Pistone and Sempi 1995, proposition 2.3). In statistics this are exactly the random variables used to define the one dimensional exponential model $p(t)$ associated with a point $p \in \mathcal{M}$ and a random variable u:

$$p(t) = \frac{e^{tu}}{Z_p(tu)} p, \qquad t \in (-\varepsilon, \varepsilon). \tag{19}$$

In particular, if we denote by $\mathcal{V}_p$ the unit ball in B_p, then it follows that the moment generating functional $Z_p(u) = \int_\Omega e^u p\mathrm{d}\mu$ is finite on $\mathcal{V}_p$ (see Pistone and Sempi 1995, proposition 2.4). The underlying idea for the Pistone–Sempi manifold is to parametrize the neighborhoods around points $p \in \mathcal{M}$ by all possible one dimensional exponential models passing through p. As a preliminary result, we mention that if two densities p and q are connected by a one dimensional exponential model, then $L^{\Phi_1}(p) = L^{\Phi_1}(q)$ (see Pistone and Rogantin 1999, proposition 5).

Pistone and Sempi define the inverse of a local chart around $p \in \mathcal{M}$ as

$$\begin{aligned} e_p : \mathcal{V}_p &\to \mathcal{M} \\ u &\mapsto \frac{e^u}{Z_p(u)} p. \end{aligned} \tag{20}$$

Denote by $\mathcal{U}_p$ the image of $\mathcal{V}_p$ under e_p. We verify that e_p is a bijection from $\mathcal{V}_p$ to $\mathcal{U}_p$, since

$$\frac{e^u}{Z_p(u)}p = \frac{e^v}{Z_p(v)}p$$

implies that $(u - v)$ is a constant random variable, which must vanish, since both u, v have zero p-expectation. Then let e_p^{-1} be the inverse of e_p on $\mathcal{U}_p$. One can check that

$$\begin{aligned} e_p^{-1} : \mathcal{U}_p &\to B_p \\ q &\mapsto \log\left(\frac{q}{p}\right) - \int_\Omega \log\left(\frac{q}{p}\right) p\mathrm{d}\mu. \end{aligned} \tag{21}$$

and also that, for any $p_1, p_2 \in \mathcal{M}$, the transition functions are given by

$$\begin{aligned} e_{p_2}^{-1}e_{p_1} : e_{p_1}^{-1}(\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2}) &\to e_{p_2}^{-1}(\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2}) \\ u &\mapsto u + \log\left(\frac{p_1}{p_2}\right) - \int_\Omega \left(u + \log\frac{p_1}{p_2}\right) p_2\mathrm{d}\mu. \end{aligned} \tag{22}$$

The main result of Pistone and Sempi (1995) is to show that the charts defined above lead to a well-defined infinite dimensional manifold. The crucial part of the proof is to show that, for any two points $p_1, p_2 \in \mathcal{M}$, the image of the overlapping neighborhoods $\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2}$ under $e_{p_1}^{-1}$ is open in the topology of the model space B_{p_1}. To do so they introduce a topology induced by the notion of exponential convergence, with respect to which the sets $\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2}$ are open, and then show that $e_{p_1}^{-1}$ is sequentially continuous from exponential convergence to L^{Φ_1}-convergence. In what follows, we bypass the use of exponential convergence and present a direct proof that the Pistone and Sempi construction yields a Banach manifold. We first need to establish the following proposition.

Proposition 1 *For any $p_1, p_2 \in \mathcal{M}$, the set $e_{p_1}^{-1}(\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2})$ is open in the topology of B_{p_1}.*

Proof Suppose that $q \in \mathcal{U}_{p_1} \cap \mathcal{U}_{p_2}$ for some $p_1, p_2 \in \mathcal{M}$. Then we can write it as

$$q = \frac{e^u}{Z_{p_1}(u)}p_1,$$

for some $u \in \mathcal{V}_{p_1}$. Using (22), we find

$$e_{p_2}^{-1}(q) = u + \log\left(\frac{p_1}{p_2}\right) - \int_\Omega \left(u + \log\frac{p_1}{p_2}\right) p_2\mathrm{d}\mu.$$

Since $e_{p_2}^{-1}(q) \in \mathcal{V}_{p_2}$, we have that

$$N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(q)\right) = N_{p_2}^{\Phi_1}\left(u + \log\left(\frac{p_1}{p_2}\right) - \int_\Omega \left(u + \log\frac{p_1}{p_2}\right) p_2\mathrm{d}\mu\right) < 1.$$

Consider an open ball of radius r around $u = e_{p_1}^{-1}(q) \in e_{p_1}^{-1}(\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2})$ in the topology of B_{p_1}, that is, consider the set

$$A_r = \{v \in B_{p_1} : N_{p_1}^{\Phi_1}(v-u) < r\}$$

and let r be small enough so that $A_r \subset \mathcal{V}_{p_1}$. Then the image in $\mathcal{M}$ of each point $v \in A_r$ under e_{p_1} is

$$\tilde{q} = e_{p_1}(v) = \frac{e^v}{Z_{p_1}(v)} p_1.$$

We claim that $\tilde{q} \in \mathcal{U}_{p_1} \cap \mathcal{U}_{p_2}$ if r is sufficiently small. Indeed, applying $e_{p_2}^{-1}$ to it we find

$$e_{p_2}^{-1}(\tilde{q}) = v + \log\left(\frac{p_1}{p_2}\right) - \int_\Omega \left(v + \log\frac{p_1}{p_2}\right) p_2 \mathrm{d}\mu,$$

so

$$\begin{aligned} N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(\tilde{q})\right) &\leq N_{p_2}^{\Phi_1}(v-u) + N_{p_2}^{\Phi_1}\left(u + \log\left(\frac{p_1}{p_2}\right) - \int_\Omega \left(u + \log\frac{p_1}{p_2}\right) p_2 \mathrm{d}\mu\right) \\ &\quad + N_{p_2}^{\Phi_1}\left(\int_\Omega (v-u) p_2 \mathrm{d}\mu\right) \\ &\leq N_{p_2}^{\Phi_1}(v-u) + N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(q)\right) + N_{p_2}^{\Phi_1}(1) \int_\Omega |v-u| p_2 \mathrm{d}\mu \\ &= N_{p_2}^{\Phi_1}(v-u) + N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(q)\right) + \|v-u\|_{1,p_2} K, \end{aligned}$$

where $K = N_{p_2}^{\Phi_1}(1)$ and we use the notation $\|\cdot\|_{1,p_2}$ for the $L^1(p_2)$-norm. As we have seen in the previous section, it follows from the growth properties of Φ_1 that there exists $c_1 > 0$ such that $\|f\|_{1,p_2} \leq c_1 N_{p_2}^{\Phi_1}(f)$. Moreover, since $L^{\Phi_1}(p_1) = L^{\Phi_1}(p_2)$ (since both p_1 and p_2 are connected to q by one dimensional exponential models) it follows that there exists a constant $c_2 > 0$ such that $N_{p_2}^{\Phi_1}(f) \leq c_2 N_{p_1}^{\Phi_1}(f)$. Therefore, the previous inequality becomes

$$\begin{aligned} N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(\tilde{q})\right) &\leq c_2 N_{p_1}^{\Phi_1}(v-u) + N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(q)\right) + c_1 c_2 K N_{p_1}^{\Phi_1}(v-u) \\ &= c_2(1 + c_1 K) N_{p_1}^{\Phi_1}(v-u) + N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(q)\right). \end{aligned}$$

Thus, if we choose

$$r < \frac{1 - N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(q)\right)}{c_2(1 + c_1 K)},$$

we will have that

$$N_{p_2}^{\Phi_1}\left(e_{p_2}^{-1}(\tilde{q})\right) < 1$$

which proves the claim. What we have just proved is that $e_{p_1}^{-1}(\mathcal{U}_{p_1} \cap \mathcal{U}_{p_2})$ consists entirely of interior points in the topology of B_{p_1}, and is therefore open in B_{p_1}.

We then have that the collection $\{(\mathcal{U}_p, e_p^{-1}),\ p \in \mathcal{M}\}$ satisfies the three axioms for being a C^∞-atlas for $\mathcal{M}$ (see Lang 1995, p. 20). Moreover, since for each connect component all the spaces B_p are isomorphic as topological vector spaces, we can say that $\mathcal{M}$ is a C^∞-manifold modeled on B_p.

As usual, the tangent space at each point $p \in \mathcal{M}$ can be abstractly identified with B_p. A concrete realization has been given in Pistone and Rogantin (1999, proposition 21), namely each curve through $p \in \mathcal{M}$ is tangent to a one-dimensional exponential model $\frac{e^{tu}}{Z_p(tu)}p$, so we take u as the tangent vector representing the equivalence class of such a curve.

Finally, given a point $p \in \mathcal{M}$, the connected component of $\mathcal{M}$ containing p coincides with the *maximal exponential model* obtained from p (see Pistone and Sempi 1995, Theorem 4.1):

$$\mathcal{E}(p) = \left\{ \frac{e^u}{Z_p(u)} p, u \in B_p \cap \mathcal{Z}_p \right\}, \tag{23}$$

where $\mathcal{Z}_p = \{f : Z_p(f) < \infty\}^0$. □

4 The Fisher information and dual connections

In the parametric version of information geometry, Amari and Nagaoka have introduced the concept of dual connections with respect to a Riemannian metric (see Amari and Nagaoka 2000 and the references given therein to their earlier work). For finite dimensional manifolds, any continuous assignment of a positive definite symmetric bilinear form to each tangent space determines a Riemannian metric. In infinite dimensions, we need to impose that the tangent space be self-dual and that the bilinear form be bounded. Since our tangent spaces B_p are not even reflexive, let alone self-dual, we abandon the idea of having a Riemannian structure on $\mathcal{M}$ and propose a weaker version of duality, the duality with respect to a continuous scalar product. When restricted to finite dimensional submanifolds, the scalar product becomes a Riemannian metric and the original definition of duality is recovered.

Let $\langle \cdot, \cdot \rangle_p$ be a continuous positive definite symmetric bilinear form assigned continuously to each $B_p \simeq T_p\mathcal{M}$. A pair of connections (∇, ∇^*) are said to be dual with respect to $\langle \cdot, \cdot \rangle_p$ if

$$\langle \tau u, \tau^* v \rangle_q = \langle u, v \rangle_p \tag{24}$$

for all $u, v \in T_p\mathcal{M}$ and all smooth curves $\gamma : [0,1] \to \mathcal{M}$ such that $\gamma(0) = p$, $\gamma(1) = q$, where τ and τ^* denote the parallel transports associated with ∇ and ∇^*, respectively. Equivalently, (∇, ∇^*) are dual with respect to $\langle\cdot,\cdot\rangle_p$ if

$$v\left(\langle s_1, s_2\rangle_p\right) = \langle \nabla_v s_1, s_2\rangle_p + \langle s_1, \nabla^*_v s_2\rangle_p \tag{25}$$

for all $v \in T_p\mathcal{M}$ and all smooth vector fields s_1 and s_2.

We stress that this *is not* the kind of duality obtained when a connection ∇ on a bundle $\mathcal{F}$ is used to construct another connection ∇' on the dual bundle $\mathcal{F}^*$ as defined, for instance, in Gibilisco and Pistone (1998, Definiton 6). The latter is a construction that does not involve any metric or scalar product and the two connections act on different bundles, while Amari–Nagaoka duality is a duality with respect to a specific scalar product (or metric, in the finite dimensional case) and the dual connections act on the same bundle, the tangent bundle.

The infinite dimensional generalization of the Fisher information is given by

$$\langle u, v\rangle_p = \int_\Omega (uv)p\mathrm{d}\mu, \quad \forall u, v \in B_p. \tag{26}$$

This is clearly bilinear, symmetric and positive definite. Moreover, continuity follows from that fact that, since $L^{\Phi_1}(p) = L^{\Phi_2}(p) \subset L^{\Phi_3}(p)$, the generalized Hölder inequality gives

$$|\langle u, v\rangle_p| \leq K N_p^{\Phi_1}(u) N_p^{\Phi_1}(v), \quad \forall u, v \in B_p. \tag{27}$$

The use of exponential Orlicz space to model the manifold naturally induces a globally flat affine connection on the tangent bundle $T\mathcal{M}$, called the *exponential* connection and denoted by $\nabla^{(1)}$. It is defined on each connected component of the manifold $\mathcal{M}$, which is equivalent to saying that its parallel transport is defined between points connected by an exponential model. If q_1 and q_2 are two such points, then the exponential parallel transport is given by

$$\begin{aligned}\tau^{(1)}_{q_1 q_2} : T_{q_1}\mathcal{M} &\to T_{q_2}\mathcal{M} \\ u &\mapsto u - \int_\Omega u q_2 d\mu.\end{aligned} \tag{28}$$

It is a well-defined isomorphism, since $T_{q_1}\mathcal{M} = B_{q_1}$ and $T_{q_2}\mathcal{M} = B_{q_2}$ are subsets of the same Orlicz space $L^{\Phi_1}(q_1) = L^{\Phi_1}(q_2)$, so the exponential parallel transport just subtracts a constant from u to make it centred around the right point.

We now want to obtain the dual parallel transport to $\tau^{(1)}$ with respect to the Fisher information, which in the parametric version of information geometry is called the mixture parallel transport since it is derived from the convex mixture of two densities. We therefore start with a result regarding such mixtures.

Proposition 2 *If q_1 and q_2 are two points in $\mathcal{U}_p$ for some $p \in \mathcal{M}$, then*

$$q(t) = tq_1 + (1-t)q_2$$

belongs to $\mathcal{E}(p)$ for all $t \in [0,1]$.

Proof We begin by writing

$$q_1 = \frac{e^{u_1}}{Z_p(u_1)}p \quad \text{and} \quad q_2 = \frac{e^{u_2}}{Z_p(u_2)}p,$$

for some $u_1, u_2 \in \mathcal{V}_p \subset L^{\Phi_1}(p)$. Therefore, there exist constants $\beta_1 > 1$ and $\beta_2 > 1$ such that $\int_\Omega \Phi_1(\beta_1 u_1) p \mathrm{d}\mu < \infty$ and $\int_\Omega \Phi_1(\beta_2 u_2) p \mathrm{d}\mu < \infty$. To simplify the notation, let us define

$$\tilde{u}_1 = u_1 - \log Z_p(u_1) \quad \text{and} \quad \tilde{u}_2 = u_2 - \log Z_p(u_2).$$

We want to show that, if we write

$$e^{\tilde{u}} p = q(t) = te^{\tilde{u}_1} p + (1-t)e^{\tilde{u}_2} p,$$

then $\tilde{u}$ is an element of $\mathcal{Z}_p$, so that

$$u = \tilde{u} - \int_\Omega \tilde{u} p \mathrm{d}\mu \in B_p \cap \mathcal{Z}_p$$

and

$$q(t) = \frac{e^u}{Z_p(u)} p \in \mathcal{E}(p).$$

For this, let $\beta = \min(\beta_1, \beta_2) > 1$ and observe that, on account of the inequality $|a+b|^\beta \leq 2^{\beta-1}(|a|^\beta + |b|^\beta)$, we have that

$$\begin{aligned} e^{\beta\tilde{u}} &= \left| te^{\tilde{u}_1} + (1-t)e^{\tilde{u}_2} \right|^\beta \\ &\leq 2^{\beta-1}\left(|t|^\beta e^{\beta\tilde{u}_1} + |1-t|^\beta e^{\beta\tilde{u}_2} \right). \end{aligned}$$

Thus

$$\int_\Omega e^{\beta\tilde{u}} p \mathrm{d}\mu \leq 2^{\beta-1}|t|^\beta \int_\Omega e^{\beta\tilde{u}_1} p \mathrm{d}\mu + 2^{\beta-1}|1-t|^\beta \int_\Omega e^{\beta\tilde{u}_2} p \mathrm{d}\mu < \infty \tag{29}$$

since both $\beta\tilde{u}_1$ and $\beta\tilde{u}_2$ are in $\tilde{L}^{\Phi_1}(p)$. On the other hand, we observe that

$$e^{-\beta\tilde{u}} = \frac{1}{\left(te^{\tilde{u}_1} + (1-t)e^{\tilde{u}_2}\right)^\beta} \leq \frac{1}{t^\beta e^{\beta\tilde{u}_1}}.$$

Therefore

$$\int_\Omega e^{-\beta\tilde{u}} p \mathrm{d}\mu \leq t^{-\beta} \int_\Omega e^{-\beta\tilde{u}_1} p \mathrm{d}\mu < \infty, \tag{30}$$

since $\beta\tilde{u}_1 \in \tilde{L}^{\Phi_1}(p)$. But this completes the proof, since (29) and (30) together imply that $\tilde{u} \in \mathcal{Z}_p$. □

We now explore the possibility of extending the convex mixture between q_1 and q_2 beyond these extreme points while maintaining positivity of $q(t)$. This depends on the relative sizes of q_1 and q_2, as shown in the next proposition:

Proposition 3 *Let $q_1 = \frac{e^{u_1}}{Z_p(u_1)} p$ and $q_2 = \frac{e^{u_2}}{Z_p(u_2)} p$ be two points in $\mathcal{U}_p$. Then there exist constants $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that $q(t) = [tq_1 + (1-t)q_2] \in \mathcal{E}_p$ for all $t \in (-2\varepsilon_1, 1+2\varepsilon_2)$ if and only if $(u_1 - u_2) \in L^\infty$. Moreover, if $(u_1 - u_2) \in L^\infty$, then $L^{\Phi_3}(q_1) = L^{\Phi_3}(q_2)$.*

Proof Suppose that $q(t) \in \mathcal{E}_p$ for all $t \in (-2\varepsilon_1, 1+2\varepsilon_2)$. Then since $q(-\varepsilon_1) \geq 0$ we have that

$$-\varepsilon_1 q_1 + (1+\varepsilon_1)q_2 \geq 0 \quad \Rightarrow \quad \frac{q_1}{q_2} \leq \frac{1+\varepsilon_1}{\varepsilon_1}.$$

Similarly, since $q(1+\varepsilon_2) \geq 0$ we have that

$$(1+\varepsilon_2)q_1 - \varepsilon_2 q_2 \geq 0 \quad \Rightarrow \quad \frac{q_1}{q_2} \geq \frac{\varepsilon_2}{1+\varepsilon_2}.$$

Therefore, the random variable

$$u_1 - u_2 = \log\left(\frac{q_1}{q_2}\right) + \int_\Omega \log\left(\frac{q_1}{p}\right) p \mathrm{d}\mu - \int_\Omega \log\left(\frac{q_2}{p}\right) p \mathrm{d}\mu$$

is uniformly bounded from above and below.

Conversely, if we have that $u_1, u_2 \in \mathcal{V}_p$ with $(u_1 - u_2) \in L^\infty$ and $K = \|u_1 - u_2\|_\infty$, then

$$\frac{Z_p(u_2)}{Z_p(u_1)} e^{-K} \leq \frac{q_1}{q_2} \leq \frac{Z_p(u_2)}{Z_p(u_1)} e^{K}.$$

We can then conclude that there exist constants $0 < \xi_1 < 1$ and $\xi_2 > 1$ such that

$$\xi_1 \leq \frac{q_1}{q_2} \leq \xi_2.$$

Then observe that for $t \leq 0$ we have

$$q(t) = \left(t\frac{q_1}{q_2} + (1-t)\right) q_2 \geq (t\xi_2 + (1-t))q_2.$$

Therefore, provided $\frac{1}{1-\xi_2} < t \leq 0$, the inequality above ensures that $q(t)$ is strictly positive. Using the same notation as in the proof of Proposition 2, the same inequality gives

$$\begin{aligned}\int_\Omega e^{-\beta\tilde{u}} p\mathrm{d}\mu &= \int_\Omega \left(\left(t\frac{q_1}{q_2} + (1-t)\right)\frac{q_2}{p}\right)^{-\beta} p\mathrm{d}\mu \\ &\leq \frac{(Z_p(u_2))^\beta}{(t\xi_2 + (1-t))^\beta}\int_\Omega e^{-\beta u_2} p\mathrm{d}\mu < \infty. \qquad (31)\end{aligned}$$

Similarly, for $t \geq 0$ we have

$$q(t) = \left(t\frac{q_1}{q_2} + (1-t)\right) q_2 \geq (t\xi_1 + (1-t))q_2.$$

Therefore, provided $0 \leq t < \frac{1}{1-\xi_1}$, this inequality shows that $q(t)$ is strictly positive and that

$$\begin{aligned}\int_\Omega e^{-\beta\tilde{u}} p\mathrm{d}\mu &= \int_\Omega \left(\left(t\frac{q_1}{q_2} + (1-t)\right)\frac{q_2}{p}\right)^{-\beta} p\mathrm{d}\mu \\ &\leq \frac{(Z_p(u_2))^\beta}{(t\xi_1 + (1-t))^\beta}\int_\Omega e^{-\beta u_2} p\mathrm{d}\mu < \infty. \qquad (32)\end{aligned}$$

Moreover, since the first part of the proof of Proposition 2 holds provided $q(t)$ is positive, we have that $q(t) = tq_1 + (1-t)q_2 \in \mathcal{E}_p$ for all $t \in \left(\frac{1}{1-\xi_2}, \frac{1}{1-\xi_1}\right)$, which completes the proof for the first statement by setting $\varepsilon_1 = \frac{1}{2(\xi_2-1)}$ and $\varepsilon_2 = \frac{\xi_1}{2(1-\xi_1)}$.

For the second statement in the Proposition, observe that

$$q_1 = \frac{1}{1+\varepsilon_2} q(1+\varepsilon_2) + \frac{\varepsilon_2}{1+\varepsilon_2} q_2 \qquad (33)$$

$$q_2 = \frac{1}{1+\varepsilon_1} q(-\varepsilon_1) + \frac{\varepsilon_1}{1+\varepsilon_1} q_1, \qquad (34)$$

for positive densities $q(1+\varepsilon_2)$ and $q(-\varepsilon_1)$. Therefore

$$\int_\Omega \Phi_3(\beta v) q_1 \mathrm{d}\mu < \infty \Longleftrightarrow \int_\Omega \Phi_3(\beta v) q_2 \mathrm{d}\mu < \infty,$$

which implies that $L^{\Phi_3}(q_1) = L^{\Phi_3}(q_2)$. Moreover, equations (33) and (34) can be used to show that the norms $N_{\Phi_3,q_1}(\cdot)$ and $N_{\Phi_3,q_2}(\cdot)$ are equivalent. To see this, let $u \in L^{\Phi_3}(q_1)$ and consider $v = \frac{u}{N_{\Phi_3,q_1}(u)}$, so that $N_{\Phi_3,q_1}(v) = 1$ and consequently

$$\int_\Omega \Phi_3(v) q_1 \mathrm{d}\mu \leq 1.$$

Using (33) we see that

$$\frac{1}{1+\varepsilon_2}\int_\Omega \Phi_3(v)q(1+\varepsilon_2)\mathrm{d}\mu+\frac{\varepsilon_2}{1+\varepsilon_2}\int_\Omega \Phi_3(v)q_2\mathrm{d}\mu\le 1,$$

which implies that

$$\frac{\varepsilon_2}{1+\varepsilon_2}\int_\Omega \Phi_3(v)q_2\mathrm{d}\mu\le 1. \tag{35}$$

On the other hand, it follows from convexity of Φ_3 that

$$\Phi_3\left(\frac{\varepsilon_2}{1+\varepsilon_2}v\right)\le\frac{\varepsilon_2}{1+\varepsilon_2}\Phi_3\left(v\right).$$

Inserting this into (35) and denoting $K=\frac{1+\varepsilon_2}{\varepsilon_2}$ gives

$$\int_\Omega \Phi_3\left(\frac{v}{K}\right)q_2\mathrm{d}\mu\le 1,$$

which means that $N_{\Phi_3,q_2}(v)\le K$. Consequently we have that

$$N_{\Phi_3,q_2}(u)=N_{\Phi_3,q_2}(v)N_{\Phi_3,q_1}(u)\le KN_{\Phi_3,q_1}(u).$$

Similarly, let $f\in L^{\Phi_3}(q_2)$ and consider $g=\frac{f}{N_{\Phi_3,q_2}(f)}$, so that $N_{\Phi_3,q_2}(g)=1$ and consequently

$$\int_\Omega \Phi_3(g)q_2\mathrm{d}\mu\le 1.$$

Using (34) we see that

$$\frac{1}{1+\varepsilon_1}\int_\Omega \Phi_3(g)q(-\varepsilon_1)\mathrm{d}\mu+\frac{\varepsilon_1}{1+\varepsilon_1}\int_\Omega \Phi_3(g)q_1\mathrm{d}\mu\le 1,$$

which implies that

$$\frac{\varepsilon_1}{1+\varepsilon_1}\int_\Omega \Phi_3(g)q_1\mathrm{d}\mu\le 1. \tag{36}$$

Again, it follows from convexity of Φ_3 that

$$\Phi_3\left(\frac{\varepsilon_1}{1+\varepsilon_1}g\right)\le\frac{\varepsilon_1}{1+\varepsilon_1}\Phi_3\left(g\right).$$

Inserting this into (36) and denoting $k = \frac{1+\varepsilon_1}{\varepsilon_1}$ gives

$$\int_\Omega \Phi_3\left(\frac{g}{k}\right) q_1 \mathrm{d}\mu \leq 1,$$

which means that $N_{\Phi_3,q_1}(g) \leq k$. Consequently we have that

$$N_{\Phi_3,q_1}(f) = N_{\Phi_3,q_1}(g) N_{\Phi_3,q_2}(f) \leq k N_{\Phi_3,q_2}(f). \tag{37}$$

□

Proposition 4 *Let $q_1 = \frac{e^{u_1}}{Z_p(u_1)} p$ and $q_2 = \frac{e^{u_2}}{Z_p(u_2)} p$ be two points in $\mathcal{U}_p$ such that $(u_1 - u_2) \in L^\infty$. Then the map*

$$\begin{aligned} \tau^{(-1)}_{q_1 q_2} : T_{q_1}\mathcal{M} &\to T_{q_2}\mathcal{M} \\ u &\mapsto \frac{q_1}{q_2} u, \end{aligned} \tag{38}$$

is an isomorphism of Banach spaces.

Proof In view of Proposition 3, we have that $L^{\Phi_3}(q_1) = L^{\Phi_3}(q_2)$ and that the norms $N_{\Phi,q_1}(\cdot)$ and $N_{\Phi,q_2}(\cdot)$ are equivalent, from which it follows that, for $u \in B_{q_1}$,

$$\begin{aligned} \left\| \frac{q_1}{q_2} u \right\|_{\Phi_1,q_2} &= \sup\left\{ \int_\Omega \left| \frac{q_1}{q_2} uv \right| q_2 \mathrm{d}\mu : v \in L^{\Phi_3}(q_2), \int_\Omega \Phi_3(v) q_2 \mathrm{d}\mu \leq 1 \right\} \\ &= \sup\left\{ \int_\Omega \left| \frac{q_1}{q_2} uv \right| q_2 \mathrm{d}\mu : v \in L^{\Phi_3}(q_2), N_{\Phi_3,q_2}(v) \leq 1 \right\} \\ &= k \sup\left\{ \int_\Omega \left| \frac{q_1}{q_2} u \left(\frac{v}{k}\right) \right| q_2 \mathrm{d}\mu : v \in L^{\Phi_3}(q_2), N_{\Phi_3,q_2}(v) \leq 1 \right\} \\ &\leq k \sup\left\{ \int_\Omega |uf| q_1 \mathrm{d}\mu : f \in L^{\Phi_3}(q_1), N_{\Phi_3,q_1}(f) \leq 1 \right\} \\ &= k \|u\|_{\Phi_1,q_1} < \infty, \end{aligned}$$

where k is the constant appearing in (37). Thus, $\frac{q_1}{q_2} u \in L^{\Phi_1}(q_2)$, and since $\frac{q_1}{q_2} u$ is centred around q_2 we have that $\frac{q_1}{q_2} u \in B_{q_2}$. Therefore $u \mapsto \frac{q_1}{q_2} u$ is a well-defined continuous bijection from B_{q_1} to B_{q_2} whose inverse map $v \mapsto \frac{q_2}{q_1} v$ is well-defined by the same arguments.

We denoted the map in the previous proposition by $\tau^{(-1)}$ since it satisfies the duality relation

$$\begin{aligned}\langle \tau^{(1)}u, \tau^{(-1)}v\rangle_{q_2} &= \left\langle u - \int_\Omega u q_2 \mathrm{d}\mu, \frac{q_1}{q_2}v \right\rangle_{q_2}\\ &= \int_\Omega u\frac{q_1}{q_2}vq_2\mathrm{d}\mu - \left(\int_\Omega uq_2\mathrm{d}\mu\right)\int_\Omega \frac{q_1}{q_2}vq_2\mathrm{d}\mu\\ &= \int_\Omega uvq_1\mathrm{d}\mu\\ &= \langle u, v\rangle_{q_1}, \quad \forall u, v \in B_{q_1},\end{aligned}$$

where the third equality follows from the fact that v is centred around q_1. □

Let us now reflect on the collective results of Propositions 2 to 4. Proposition 2 tells us that the convex mixture of two probability densities in the same $\mathcal{U}_p$ remains in the connected component $\mathcal{E}_p$ of $\mathcal{M}$, but not necessarily in the same neighbourhood. Proposition 3 then characterizes exactly those pairs q_1 and q_2 for which the convex mixture can be extended beyond the extreme points while remaining in the same connected component $\mathcal{E}_p$. For such pairs, Proposition 4 gives an isomorphism of tangent spaces $\tau^{(-1)}$ which satisfies the duality relation (24) with respect to the Fisher information. However, we refrain from calling $\tau^{(-1)}$ a parallel transport, since it might fail to be well-defined when the points q_1 and q_2 do not satisfy the conditions of Proposition 4. Nevertheless, we can still compute the derivative of $\tau^{(-1)}$ along curves that satisfy these conditions, as is done in the next proposition.

Proposition 5 *Let $v \in T_p\mathcal{M}$ be a tangent vector at $p \in \mathcal{M}$ and $s \in S(T\mathcal{M})$ be a differentiable vector field. If there exist a differentiable curve $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ such that $p = \gamma(0)$, $v = \gamma'(0)$, and whose image consists entirely of points satisfying the hypotheses of Proposition 4, then*

$$(D_v^{(-1)}s)(p) := (d_v s)(p) + s(p)\ell'(0) \in B_p, \tag{39}$$

where $(d_v s)(p)$ denotes the directional derivative of s in the direction of v in the Banach space $L^{\Phi_1}(p)$, and $\ell(t) = \log(\gamma(t))$.

Proof For h sufficiently small, it follows from Proposition 4 that $\tau^{(-1)}_{\gamma(h)\gamma(0)}s(\gamma(h)) \in B_{\gamma(0)}$. The result then follows from the following calculation:

$$\begin{aligned}\lim_{h\to 0}\frac{1}{h}\left[\tau^{(-1)}_{\gamma(h)\gamma(0)}s(\gamma(h)) - s(\gamma(0))\right] &= \lim_{h\to 0}\frac{1}{h}\left[\frac{\gamma(h)}{\gamma(0)}s(\gamma(h)) - s(\gamma(0))\right]\\ &= \lim_{h\to 0}\frac{1}{h}\left[s(\gamma(h)) - s(\gamma(0))\right]\\ &\quad + \lim_{h\to 0}\frac{1}{h}\left[\frac{\gamma(h)-\gamma(0)}{\gamma(0)}s(\gamma(h))\right]\\ &= (d_v s)(p) + s(p)\ell'(0).\end{aligned}$$

Despite satisfying all the usual properties of a covariant derivative, such as linearity and Leibniz rule, the differential operator $D_v^{(-1)}$ might fail to be well-defined when no curve satisfying the conditions of Proposition 5 exists. For this reason, we simply call it the (-1)-derivative in the direction of those v for which it is well-defined. In the next section, we will see that $D_v^{(-1)}$ is part of a one-parameter family of derivatives defined for exactly this class of tangent vectors. □

5 α-Derivatives

In this section, we define an infinite-dimensional analogue of the α-connections introduced in the parametric case independently in Čencov (1982) and Amari (1985). We use the same technique proposed in Gibilisco and Pistone (1998), namely exploring the geometry of spheres in the Lebesgue spaces L^r, but modified in such a way that the resulting derivatives all act on sections of the tangent bundle $T\mathcal{M}$. The price we pay is that our derivatives are not defined for all tangent vectors, but only those satisfying the conditions of Proposition 5.

We begin with the Amari–Nagaoka α-embeddings

$$\begin{aligned}\ell_\alpha : \mathcal{M} &\to L^r(\mu)\\ p &\mapsto \frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}}, \quad \alpha \in [-1, 1),\end{aligned} \tag{40}$$

where $r = \frac{2}{1-\alpha}$.

Observe that

$$\|\ell_\alpha(p)\|_r = \left[\int_\Omega \ell_\alpha(p)^r \mathrm{d}\mu\right]^{1/r} = \left[\int_\Omega \left(\frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}}\right)^r \mathrm{d}\mu\right]^{1/r} = r,$$

so $\ell_\alpha(p) \in S^r(\mu)$, the sphere of radius r in $L^r(\mu)$ (we warn the reader that, throughout this paper, the r in S^r refers to the fact that this is a sphere of radius r, while the fact that it is a subset of L^r is judiciously omitted from the notation).

According to Gibilisco and Pistone (1998), the tangent space to $S^r(\mu)$ at a point f is

$$T_f S^r(\mu) = \left\{ g \in L^r(\mu) : \int_\Omega g f^* \mathrm{d}\mu = 0 \right\}, \tag{41}$$

where $f^* = \mathrm{sgn}(f)|f|^{r-1}$. In our case,

$$f = \ell_\alpha(p) = rp^{1/r} \tag{42}$$

so that

$$f^* = \left(rp^{1/r}\right)^{r-1} = r^{r-1} p^{1-1/r}. \tag{43}$$

Therefore, the tangent space to $S^r(\mu)$ at $rp^{1/r}$ is

$$T_{rp^{1/r}}S^r(\mu) = \left\{ g \in L^r(\mu) : \int_\Omega gp^{1-1/r} \mathrm{d}\mu = 0 \right\}. \tag{44}$$

We now look for a concrete realization of the push-forward of the map ℓ_α when the tangent space $T_p\mathcal{M}$ is identified with B_p as in the previous sections. Since

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{2}{1-\alpha}p^{\frac{1-\alpha}{2}}\right) = p^{\frac{1-\alpha}{2}}\frac{\mathrm{d}\log p}{\mathrm{d}t},$$

the α-push-forward can be formally implemented as

$$\begin{aligned} (\ell_\alpha)_{*(p)} : T_p\mathcal{M} = B_p &\to T_{rp^{1/r}}S^r(\mu) \\ u &\mapsto p^{\frac{1-\alpha}{2}}u. \end{aligned} \tag{45}$$

For this to be well defined, we need to check that $p^{\frac{1-\alpha}{2}}u$ is an element of $T_{rp^{1/r}}S^r(\mu)$. Indeed, since $L^{\Phi_1}(p) \subset L^s(p)$ for all $s \geq 1$, we have that

$$\int_\Omega \left|p^{1/r}u\right|^r \mathrm{d}\mu = \int_\Omega |u|rf^r p\mathrm{d}\mu < \infty,$$

so $p^{\frac{1-\alpha}{2}}u \in L^r(\mu)$. Moreover

$$\int_\Omega p^{1/r}up^{1-1/r}\mathrm{d}\mu = \int_\Omega up\mathrm{d}\mu = 0,$$

which verifies that $p^{1/r}u \in T_{rp^{1/r}}S^r(\mu)$.

The sphere $S^r(\mu)$ inherits a natural connection obtained by projecting the trivial connection on $L^r(\mu)$ (the one where parallel transport is just the identity map) onto its tangent space at each point. For each $f \in S^r(\mu)$, a canonical projection from the tangent space $T_f L^r(\mu)$ onto the tangent space $T_f S^r(\mu)$ can be uniquely defined, since the spaces $L^r(\mu)$ are uniformly convex (see Gibilisco and Isola 1999), and is given by

$$\begin{aligned} \Pi_f : T_f L^r(\mu) &\to T_f S^r(\mu) \\ g &\mapsto g - \left(r^{-r}\int_\Omega gf^*\mathrm{d}\mu\right)f. \end{aligned} \tag{46}$$

When $f = rp^{1/r}$ and $f^* = r^{r-1}p^{1-1/r}$, the formula above gives

$$\begin{aligned} \Pi_{rp^{1/r}} : T_{rp^{1/r}}L^r(\mu) &\to T_{rp^{1/r}}S^r(\mu) \\ g &\mapsto g - \left(\int_\Omega gp^{1-1/r}\mathrm{d}\mu\right)p^{1/r}. \end{aligned} \tag{47}$$

We are now ready to introduce the α-derivative. Suppose that $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ is a smooth curve whose image consists entirely of points satisfying the conditions of Proposition 4. Then the α-push-forward of an arbitrary vector field $s \in S(T\mathcal{M})$ along γ is

$$(\ell_\alpha)_{*(\gamma(t))}s = \gamma(t)^{1/r}s(\gamma(t)),$$

while the α-push-forward of the tangent vector $v = \dot{\gamma}(0) \in T\mathcal{M}_p$ is

$$(\ell_\alpha)_{*(p)}v = p^{1/r}v.$$

Therefore, the covariant derivative of $(\ell_\alpha)_{*(\gamma(t))}s(\gamma(t))$ in the direction of $(\ell_\alpha)_{*(p)}v$ with respect to the trivial connection $\widetilde{\nabla}$ on $L^r(\mu)$ is given by

$$\begin{aligned}\widetilde{\nabla}_{(\ell_\alpha)_{*(p)}v}(\ell_\alpha)_{*(\gamma(t))}s &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\gamma(t)^{1/r}s(\gamma(t))\right)\Big|_{t=0} \\ &= \frac{1}{r}p^{1/r}\left.\frac{\mathrm{d}\log(\gamma(t))}{\mathrm{d}t}\right|_{t=0}s(p) + p^{1/r}\left.\frac{\mathrm{d}s(\gamma(t))}{\mathrm{d}t}\right|_{t=0} \\ &= \frac{1}{r}p^{1/r}\ell'(0)s(p) + p^{1/r}\left(d_v s\right)(p).\end{aligned}$$

Using the projection $\Pi_{rp^{1/r}}$ to obtain a tangent vector in $T_{rp^{1/r}}S^r(\mu)$ we get

$$\begin{aligned}\Pi_{rp^{1/r}}\widetilde{\nabla}_{(\ell_\alpha)_{*(p)}v}(\ell_\alpha)_{*(\gamma(t))}s = p^{1/r}\Bigg[&\frac{1}{r}\ell'(0)s(p) + (d_v s)(p) \\ &- \int_\Omega\left(\frac{1}{r}\ell'(0)s(p) + (d_v s)(p)\right)p\mathrm{d}\mu\Bigg].\end{aligned}$$

It then follows from Proposition 5 that $\frac{1}{r}\ell'(0)s(p) + (d_v s)(p) \in L^{\Phi_1}(p)$, which implies that the expression above belongs to the image of $(\ell_\alpha)_{*(p)}$. Therefore, we can pull it back to $T_p\mathcal{M}$ using $(\ell_\alpha)^{-1}_{*(p)}$, from which we obtain

$$\begin{aligned}(\ell_\alpha)^{-1}_{*(p)}\left[\Pi_{rp^{1/r}}\widetilde{\nabla}_{(\ell_\alpha)_{*(p)}v}(\ell_\alpha)_{*(\gamma(t))}s\right] &= \frac{1}{r}\ell'(0)s(p) + (d_v s)(p) \\ &\quad - \int_\Omega\left(\frac{1}{r}\ell'(0)s(p) + (d_v s)(p)\right)p\mathrm{d}\mu. \qquad (48)\end{aligned}$$

As this construction shows, the α-derivatives can be rigorously defined on the tangent bundle $T\mathcal{M}$ as follows:

Definition 1 *For $\alpha \in [-1, 1)$, let $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ be a smooth curve such that $p = \gamma(0)$ and whose image consists entirely of points satisfying the conditions of Proposition 4. The α-derivative of a differentiable vector field $s \in S(T\mathcal{M})$ in the direction of $v = \gamma'(0)$ is given by*

$$\left(D^\alpha_v s\right)(p) = (\ell_\alpha)^{-1}_{*(p)}\left[\Pi_{rp^{1/r}}\widetilde{\nabla}_{(\ell_\alpha)_{*(p)}v}(\ell_\alpha)_{*(\gamma(t))}s\right]. \qquad (49)$$

Before we proceed, observe that since $s(\gamma(t)) \in B_{\gamma(t)}$ for each $t \in (-\varepsilon, \varepsilon)$, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega s(\gamma(t))\gamma(t)\mathrm{d}\mu = 0 \\
&\int_\Omega \frac{\mathrm{d}s(\gamma(t))}{\mathrm{d}t}\gamma(t)\mathrm{d}\mu = -\int_\Omega s(\gamma(t))\frac{\mathrm{d}\gamma(t)}{\mathrm{d}t}\mathrm{d}\mu \\
&\int_\Omega \frac{\mathrm{d}s(\gamma(t))}{\mathrm{d}t}\gamma(t)\mathrm{d}\mu = -\int_\Omega s(\gamma(t))\frac{\mathrm{d}\log(\gamma(t))}{\mathrm{d}t}\gamma(t)\mathrm{d}\mu.
\end{aligned}$$

In particular, for $t = 0$, we get

$$\int_\Omega (d_v s)\,(p)\,p\mathrm{d}\mu = -\int_\Omega s(p)\dot{\ell}(0)p\mathrm{d}\mu \tag{50}$$

Inserting this relation into (48) with $\alpha = -1$, corresponding to $r = 1$, leads to

$$\left(D_v^{(-1)}s\right)(p) = (d_v s)(p) + s(p)\ell'(0), \tag{51}$$

which coincides with (39).

Recall that the covariant derivative associated with the exponential parallel transport (28) was computed in Gibilisco and Pistone (1998, proposition 25) as

$$\left(\nabla_v^{(1)}s\right)(p) = (d_v s)(p) - \int_\Omega (d_v s)(p)p\mathrm{d}\mu. \tag{52}$$

The next proposition shows that the relation between the exponential connection and the α-derivatives just defined is the same as in the parametric case. Its proof resembles the calculation in the last pages of Gibilisco and Pistone (1998), except that all our derivatives act on the same bundle, whereas in Gibilisco and Pistone (1998) each one is defined on its own bundle-connection pair.

Proposition 6 *The exponential connection and the α-derivatives on $T\mathcal{M}$ satisfy*

$$D^\alpha = \frac{1+\alpha}{2}\nabla^{(1)} + \frac{1-\alpha}{2}D^{(-1)}. \tag{53}$$

Proof Let $\ell(t) = \log(\gamma(t))$ with γ, s, p and v as in definition 1. Inserting (50) into (49) gives

$$\begin{aligned}
\left(D_v^\alpha s\right)(p) &= \frac{1}{r}\ell'(0)s(p) + (d_v s)(p) + \left(\frac{1}{r} - 1\right)\int_\Omega (d_v s)(p)p\mathrm{d}\mu \\
&= \left(\frac{1+\alpha}{2}\right)\left[(d_v s)(p) - \int_\Omega (d_v s)(p)\right]
\end{aligned}$$

$$+\left(\frac{1-\alpha}{2}\right)[(d_v s)(p)+s(p)\ell'(0)]$$
$$=\frac{1+\alpha}{2}\left(\nabla_v^{(1)}s\right)(p)+\frac{1-\alpha}{2}\left(D_v^{(-1)}s\right)(p).$$
□

6 Auto-parallel curves

We now investigate some of the auto-parallel curves associated with the derivatives introduced in the previous sections. First observe that a one-dimensional exponential model of the form

$$q(t)=\frac{e^{tu}}{Z_p(tu)}p, \qquad u\in B_p, \quad t\in(-\varepsilon,\varepsilon),$$

which obviously belong to the connected component $\mathcal{E}_p$, is an auto-parallel curve for $\nabla^{(1)}$, since its tangent vector field $s(t)=\frac{\mathrm{d}}{\mathrm{d}t}\left(\log\frac{q(t)}{p}\right)$ (according to Pistone and Rogantin 1999, proposition 21) satisfies

$$\left(\nabla_{\dot q(t)}^{(1)}s(t)\right)(q(t))=\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(tu-\log Z_p(tu)\right)-E_{q(t)}\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(tu-\log Z_p(tu)\right)\right)$$
$$=-\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log Z_p(tu)+E_{q(t)}\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log Z_p(tu)\right)=0.$$

Next observe that for q_1 and q_2 satisfying the conditions of Proposition 4, a mixture model of the form

$$q(t)=tq_1+(1-t)q_2, \qquad q_1,q_2\in\mathcal{U}_p, \quad t\in[0,1],$$

which belongs to the connected component $\mathcal{E}_p$ according to Proposition 2, is an auto-parallel curve for $D^{(-1)}$, since the tangent vector field $s(t)=\frac{\mathrm{d}}{\mathrm{d}t}\left(\log\frac{q(t)}{p}\right)$ satisfies

$$\left(D_{(e_p^{-1}\circ q)'(t)}^{(-1)}s(t)\right)(q(t))=\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left[\log\frac{tq_1+(1-t)q_2}{p}\right]$$
$$+\frac{\mathrm{d}}{\mathrm{d}t}\left[\log\frac{tq_1+(1-t)q_2}{p}\right]\frac{\mathrm{d}}{\mathrm{d}t}[\log tq_1+(1-t)q_2]$$
$$=\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{p}{tq_1+(1-t)q_2}\frac{(q_1-q_2)}{p}\right]$$
$$+\left(\frac{p}{tq_1+(1-t)q_2}\frac{q_1-q_2}{p}\right)\frac{(q_1-q_2)}{tq_1+(1-t)q_2}$$
$$=-\left(\frac{(q_1-q_2)}{tq_1+(1-t)q_2}\right)^2+\left(\frac{(q_1-q_2)}{tq_1+(1-t)q_2}\right)^2=0.$$

The next theorem establishes the corresponding result for the α-derivatives.

Proposition 7 *For $\alpha \in (-1, 1)$, the α-auto-parallel curves between two of points q_1 and q_2 in $\mathcal{U}_p$ satisfying the conditions of Proposition 4, for some $p \in \mathcal{M}$, belongs to the connected component $\mathcal{E}_p$.*

Proof Using the same notation as in Proposition 2, we have that the α-auto-parallel curve connecting $q_1, q_2 \in \mathcal{E}(p)$ is the pull back of the arc of great circle connecting their images $f_1 = \ell_\alpha(q_1)$ and $f_2 = \ell_\alpha(q_2)$ on the sphere $S^r(\mu)$. Now if $tf_1+(1-t)f_2$ is the straight line connecting f_1 and f_2 in $L^r(\mu)$, then for each fixed $t \in [0, 1]$ the corresponding point on the sphere of radius r is

$$f(t) = \frac{r}{k(t)}[tf_1 + (1-t)f_2], \tag{54}$$

where $k(t) = \|tf_1 + (1-t)f_2\|_r$. Let us write its inverse image with respect to the α-embedding as

$$e^{\tilde{u}}p = (\ell_\alpha)^{-1}(f(t)) = \frac{1}{k(t)^r}[tf_1 + (1-t)f_2]^r,$$

for some random variable $\tilde{u}$. Following the argument in proposition 2, we see that

$$\begin{aligned} e^{\beta\tilde{u}} &= \frac{1}{p^\beta k(t)^{\beta r}}[tf_1 + (1-t)f_2]^{\beta r} \\ &\le \left[\frac{2r}{k(t)}\right]^{\beta r}[t^{\beta r}e^{\beta\tilde{u}_1} + (1-t)^{\beta r}e^{\beta\tilde{u}_2}], \end{aligned}$$

so that

$$\int_\Omega e^{\beta\tilde{u}}p\mathrm{d}\mu < \infty, \tag{55}$$

since both βu_1 and βu_2 are in $\tilde{L}^{\Phi_1}(p)$. Furthermore,

$$e^{-\beta\tilde{u}} = \frac{p^\beta k(t)^{\beta r}}{[tf_1 + (1-t)f_2]^{\beta r}} \le \left[\frac{k(t)}{tr}\right]^{\beta r} e^{-\beta\tilde{u}_1},$$

so that

$$\int_\Omega e^{-\beta\tilde{u}}p\mathrm{d}\mu < \infty, \tag{56}$$

which together with (55), imply that $\tilde{u} \in \mathcal{Z}_p$. To complete the proof we can define

$$u = \tilde{u} - \int_\Omega \tilde{u}p\mathrm{d}\mu, \tag{57}$$

to obtain that $u \in B_p$ and

$$(\ell_\alpha)^{-1}(f(t)) = \frac{e^u}{Z_p(u)} p \in \mathcal{E}(p). \tag{58}$$

□

7 Further developments

We have seen that using L^{Φ_1} as the coordinate space for the infinite-dimensional information manifold leads to a well-defined isomorphism $\tau^{(-1)}$ between the tangent spaces B_{q_1} and B_{q_2} whenever the difference of their log-likelihoods u_1 and u_2 is bounded. Moreover, this isomorphism is dual to the exponential parallel transport with respect to the generalized Fisher metric. The next step in our program is to show that the Kullback–Leibler relative entropy is the statistical divergence associated with the dualistic triple $(\tau^{(1)}, \tau^{(-1)}, g)$ (see Amari and Nagaoka 2000). In the same vein, since our interpolation family of α-derivatives satisfy the same convex mixture structure as in finite dimensions, we are led to the study of the infinite-dimensional analogues of the α-divergences. The completion of this circle of ideas would be an infinite-dimensional generalization of the projection theorems obtained by Amari in the finite dimensional case. Namely, one seeks to prove that, given a point $p \in \mathcal{M}$ and an α-flat submanifold $\mathcal{S}$, then the point $q \in \mathcal{S}$ with minimal α-divergence from p is obtained by projecting p orthogonally (with respect to the Fisher metric) onto $\mathcal{S}$ following a $-\alpha$-geodesic. An equally ambitious result to be pursued is the infinite-dimensional analogue of Centsov's theorem, which would characterize the generalized Fisher metric as the unique continuous scalar product on $\mathcal{M}$ which is reduced by Markov morphisms on the tangent space.

Acknowledgments I would like to acknowledge the financial support from CAPES, Brazil, and of an ORS scholarship from the British government received while I was a PhD student at King's College London, where this research was initiated. It is a pleasure to thank R.F. Streater for suggesting the theme and thoroughly discussing the results obtained. I am also grateful to P. Gibilisco and G. Pistone for estimulating discussions during the Information Geometry and Applications conference held in Pescara, July 2002. Finally, I am indebted to H. Nagaoka for pointing out a crucial mistake in the previous version of this paper, and to two anonymous referees for marked improvements leading to the current version.

References

Amari, S.-I. (1985). *Differential-geometrical methods in statistics*. New York: Springer.

Amari, S.-I., Nagaoka, H. (2000). *Methods of information Geometry*. Providence, RI: American Mathematical Society. Translated from the 1993 Japanese original by Daishi Harada.

Čencov, N.N. (1982). Statistical decision rules and optimal inference. Providence, RI: American Mathematical Society. Translation from the Russian edited by Lev J. Leifman.

Dawid, A.P. (1975). On the concepts of sufficiency and ancillarity in the presence of nuisance parameters. *Journal of the Royal Statistical Society B, 37*, 248–258.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics, 3*, 1189–1242. With a discussion by C. R. Rao, Don A. Pierce, D. R. Cox,

D. V. Lindley, Lucien LeCam, J. K. Ghosh, J. Pfanzagl, Neils Keiding, A. P. Dawid, Jim Reeds and with a reply by the author.

Gibilisco, P., Isola, T. (1999). Connections on statistical manifolds of density operators by geometry of noncommutative L^p-spaces. *Infinite Dimensional Analysis Quantum Probability and Related Topics, 2*, 169–178.

Gibilisco, P., Pistone, G. (1998). Connections on non-parametric statistical manifolds by Orlicz space geometry. *Infinite Dimensional Analysis Quantum Probability and Related Topics, 1*, 325–347.

Grasselli, M.R. (2001). Classical and quantum information geometry. Ph.D. thesis, King's College, London.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of Royal Society A, 186*, 453–461.

Kass, R.E., Vos, P.W. (1997). *Geometrical foundations of asymptotic inference*. New York: Wiley-Interscience.

Krasnosel′skiĭ, M.A., Rutickiĭ, J.B. (1961). *Convex functions and Orlicz spaces*. Groningen: P. Noordhoff.

Lang, S. (1995). *Differential and Riemannian manifolds* (3rd ed.). New York: Springer.

Murray, M.K., Rice, J.W. (1993). *Differential geometry and statistics*. London: Chapman & Hall.

Pistone, G. (2001). New ideas in nonparametric estimation. In P. Sollich, et al. (Eds.), *Disordered and complex systems*. American Institute of Physics. AIP Conference Proceedings 553.

Pistone, G., Rogantin, M.P. (1999). The exponential statistical manifold: Mean parameters, orthogonality and space transformations. *Bernoulli, 5*, 721–760.

Pistone, G., Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Annals of Statistics, 23*, 1543–1561.

Rao, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society, 37*, 81–91.

Rao, M.M., Ren, Z.D. (1991). *Theory of Orlicz spaces*. New York: Marcel Dekker.

Sollich, P., et al. (Eds.) (2001). *Disordered and complex systems*. American Institute of Physics. AIP Conference Proceedings 553.