Bernoulli **22**(2), 2016, 711–731
DOI: 10.3150/14-BEJ673

Infinite-dimensional statistical manifolds based on a balanced chart

NIGEL J. NEWTON[1]

[1]*School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK. E-mail: njn@essex.ac.uk.*

We develop a family of infinite-dimensional Banach manifolds of measures on an abstract measurable space, employing charts that are "balanced" between the density and log-density functions. The manifolds, $(\tilde{M}_\lambda, \lambda \in [2, \infty))$, retain many of the features of finite-dimensional information geometry; in particular, the α-divergences are of class $C^{\lceil \lambda \rceil - 1}$, enabling the definition of the Fisher metric and α-derivatives of particular classes of vector fields. Manifolds of *probability* measures, $(M_\lambda, \lambda \in [2, \infty))$, based on centred versions of the charts are shown to be $C^{\lceil \lambda \rceil - 1}$-embedded submanifolds of the $\tilde{M}_\lambda$. The Fisher metric is a pseudo-Riemannian metric on $\tilde{M}_\lambda$. However, when restricted to finite-dimensional embedded submanifolds it becomes a Riemannian metric, allowing the full development of the geometry of α-covariant derivatives. $\tilde{M}_\lambda$ and M_λ provide natural settings for the study and comparison of approximations to posterior distributions in problems of Bayesian estimation.

Keywords: Banach manifold; Bayesian estimation; Fisher metric; information geometry; non-parametric statistics

1. Introduction

This paper develops a family of infinite-dimensional manifolds of measures, each containing a smoothly embedded submanifold of *probability* measures. It was motivated by problems of Bayesian estimation, in which posterior distributions have to be computed from a variety of partial observations. This can rarely be done exactly owing to issues of dimension and nonlinearity, and the study of approximations is contingent on the development of appropriate measures of error. The manifolds we construct have metrics suited to such problems.

Suppose, for example, that $X : \Omega \to \mathbb{X}$ and $Y : \Omega \to \mathbb{Y}$ are random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in metric spaces $\mathbb{X}$ and $\mathbb{Y}$, respectively. X is the *estimand* whose posterior distribution we seek, and Y is the observable. Let $\mathcal{P}$ be the set of probability measures on the Borel subsets of $\mathbb{X}$. Under mild conditions (see, e.g., [14]) an abstract Bayes formula defines a regular conditional distribution for X given Y, $\Pi : \mathbb{Y} \to \mathcal{P}$. (For any Borel set $B \subseteq X$, $\Pi(\cdot)(B) : \mathbb{Y} \to [0, 1]$ is measurable, and $\mathbb{P}(X \in B \mid Y) = \Pi(Y)(B)$.) In the applications we have in mind, $\Pi(\mathbb{Y})$ is typically of infinite dimension and so we need to construct approximations of the form $\hat{\Pi} : \mathbb{Y} \to \mathcal{Q} \subset \mathcal{P}$, where $\mathcal{Q}$ is of finite dimension.

Single estimation objectives, such as minimum mean-square error in the approximation of a real-valued random variable $f(X)$, induce their own specific measures of error on $\mathcal{P}$, but these may not be easy to use. On the other hand, if f is sufficiently regular, then a more generic measure of error such as the L^2 metric on densities may be useful. If $\mu \in \mathcal{P}$ is a reference measure with

respect to which $\Pi(y)$ and $\hat{\Pi}(y)$ have densities $\pi(y)$ and $\hat{\pi}(y)$, then the difference between the minimum mean-square error estimate of $f(X)$ and the mean of f under $\hat{\Pi}(y)$ can be bounded by means of the Cauchy–Schwarz inequality:

$$(\mathbf{E}_{\Pi(y)} f - \mathbf{E}_{\hat{\Pi}(y)} f)^2 \leq \mathbf{E}_\mu f^2 \mathbf{E}_\mu \big(\pi(y) - \hat{\pi}(y)\big)^2. \tag{1}$$

Although, in this context, the L^2 metric on densities induces an appropriate topology on $\mathcal{P}$, it may still be poor in practice. This is so, for example, if f is the indicator function of a rare, but important, event. Moreover, we often need generic measures of error that are suitable for a *variety* of objectives. This is especially important if the underlying estimation problem is inherently multi-objective, as is the case, for example, when tracking the movement of many objects.

The mean-square error of $\mathbf{E}_{\hat{\Pi}(Y)} f$ admits the orthogonal decomposition:

$$\mathbb{E}\big(f(X) - \mathbf{E}_{\hat{\Pi}(Y)} f\big)^2 = \mathbb{E}\mathbf{E}_{\Pi(Y)}(f - \mathbf{E}_{\Pi(Y)} f)^2 + \mathbb{E}(\mathbf{E}_{\Pi(Y)} f - \mathbf{E}_{\hat{\Pi}(Y)} f)^2.$$

The first term on the right-hand side is the *estimation error* arising from the limitations of the observation Y; the second term is the *approximation error* arising from the use of $\hat{\Pi}$ instead of Π. When comparing errors for more than one random variable, it is natural to normalise the approximation errors by their associated estimation errors—there is no point in approximating the conditional mean, $\mathbf{E}_{\Pi(y)} f$, with high precision if it is itself a poor estimate of $f(X)$. With this in mind, we might propose the following extreme, multi-objective, mean-square measure of error on $\mathcal{P}$:

$$\begin{aligned} \mathcal{D}(Q \mid P) &:= \sup_{f \in \mathcal{L}^2(P)} \frac{(\mathbf{E}_P f - \mathbf{E}_Q f)^2}{\mathbf{E}_P (f - \mathbf{E}_P f)^2} \\ &= \sup_{f \in F} \big(\mathbf{E}_P f (1 - \mathrm{d}Q/\mathrm{d}P)\big)^2 \\ &= \mathbf{E}_P (1 - \mathrm{d}Q/\mathrm{d}P)^2, \end{aligned} \tag{2}$$

where $\mathcal{L}^2(P) = \{f : \mathbb{X} \to \mathbb{R} : \mathbf{E}_P f^2 < \infty\}$ and F is the subset of such functions having zero mean and unit variance. This is the χ^2-divergence. Although extreme, it illustrates a feature of many multi-objective measures of error: they ensure that probabilities of events that are small are approximated with greater absolute accuracy than those that are large. The L^p metrics on densities fail in this respect. (A related disadvantage is that spaces of probability densities have boundaries, which can create problems with numerical methods.) A commonly used, less extreme, multi-objective measure of error is the *Kullback–Leibler* divergence. This is widely used in variational Bayesian estimation. (See, e.g., [12,23].)

The regularity of the Kullback–Leibler divergence was central to the design of the manifolds in this paper. Each manifold is covered by a single chart, which places its elements in one-to-one correspondence with those of a Banach space. Because of this, the manifolds are also *metric spaces* of measures, with metrics tailored (at least locally) to problems of Bayesian estimation. The manifolds are large enough to include exact posterior distributions in many problems. They

also include, as smoothly embedded submanifolds, a large variety of finite-dimensional families of probability measures, on which approximations can be based.

The study and approximation of *nonlinear filters* (an application pursued elsewhere by the author) was a particular motivation. A nonlinear filter computes the posterior distributions of a Markov *signal process* from randomly-perturbed observations that become progressively available in time. For a modern perspective on the theory and application of nonlinear filtering, the reader is referred to [8]. Approximations based on information geometric projections onto finite-dimensional exponential families were studied in [5].

The equations of nonlinear filtering are often expressed in terms of the "un-normalised" version of the posterior distribution obtained when the marginal density of the observation is omitted from the denominator in Bayes' formula. This satisfies the so-called *Zakai equation*, which has a particularly simple (bilinear) form. A manifold of finite measures with a suitable metric is a natural space for such un-normalised posteriors. We develop a family of such manifolds in Section 3, not only because of this application, but also because many of the properties of the *statistical* manifolds of Section 4 are best understood in the context of their embedding in these larger manifolds. The manifolds are also natural settings in which to study and compare finite-dimensional statistical manifolds that admit the full geometry of α-covariant derivatives.

The paper is structured as follows. Section 2 provides a brief introduction to information geometry. Section 3 introduces the one-parameter family of manifolds of finite measures, $((\tilde{M}_\lambda, \tilde{\phi}_\lambda), \lambda \in [2, \infty))$, and studies on them the properties of Amari's α-embedding maps. Section 4 develops the family of manifolds of *probability* measures $((M_\lambda, \phi_\lambda), \lambda \in [2, \infty))$ in which the chart ϕ_λ is a "centred" version of $\tilde{\phi}_\lambda$. Section 5 studies the properties of the α-divergences on $\tilde{M}_\lambda$ and M_λ, defining the Fisher metric, and a limited notion of α-parallel transport on the tangent bundle. Some examples of finite-dimensional embedded submanifolds of $\tilde{M}_\lambda$ and M_λ are outlined in Section 6. A sketch of some of the results of Sections 4 and 5.1 was given, without proofs, in [18].

2. Information geometry

We begin by reviewing a classical finite-dimensional example: the exponential statistical manifold. (See, e.g., [2].) Let $(\mathbb{X}, \mathcal{X}, \mu)$ be a probability space supporting real-valued random variables $(\eta_i; i = 1, \ldots, d)$ with the following properties: (i) the variables $(1, \eta_1, \eta_2, \ldots, \eta_d)$ are linearly independent elements of $L^0(\mu)$, that is, $\mu(\alpha + \sum_i y^i \eta_i = 0) = 1$ if and only if $\alpha = 0$ and $\mathbb{R}^d \ni y = 0$; (ii) $\mathbf{E}_\mu \exp(\sum_i y^i \eta_i) < \infty$ for all y in a non-empty open subset $B \subseteq \mathbb{R}^d$. For each $y \in B$, let P_y be the probability measure on $\mathcal{X}$ with density

$$\frac{\mathrm{d}P_y}{\mathrm{d}\mu} = \exp\left(\sum_i y^i \eta_i - c(y)\right), \tag{3}$$

where $c(y) = \log \mathbf{E}_\mu \exp(\sum_i y^i \eta_i)$, and let $N := \{P_y\colon\ y \in B\}$. It follows from (i) that the map $B \ni y \mapsto P_y \in N$ is a bijection. Let $\theta : N \to B$ be its inverse; then (N, B, θ) is an *exponential statistical manifold*, with an atlas comprising the single chart θ. We can think of a *tangent vector* at $P \in N$, U, as being an equivalence class of differentiable curves passing through P: two curves

(expressed in coordinates), $(\mathbf{y}(t) \in B, t \in (-\varepsilon, \varepsilon))$ and $(\mathbf{z}(t) \in B, t \in (-\varepsilon, \varepsilon))$, being equivalent at P if $\mathbf{y}(0) = \mathbf{z}(0) = \theta(P)$ and $\dot{\mathbf{y}}(0) = \dot{\mathbf{z}}(0)$. The *tangent space* at P, $T_P N$, is the linear space of all such tangent vectors, and is spanned by the vectors $(\partial_i; i = 1, \ldots, d)$, where ∂_i is the equivalence class containing the curve $(\mathbf{y}_i(t) := \theta(P) + t\mathbf{e}_i, t \in (-\varepsilon, \varepsilon))$, and $\mathbf{e}_i^j$ is equal to the Krönecker delta. The *tangent bundle* is the disjoint union $TN := \bigcup_{P \in N}(P, T_P N)$, and admits the global chart $\Theta : TN \to B \times \mathbb{R}^d$, where $\Theta^{-1}(y, u) = (\theta^{-1}(y), u^i \partial_i)$. If a function $f : N \to \mathbb{R}^n$ is differentiable, and $U \in T_P N$, then we write

$$Uf = u^i \partial_i f := u^i \frac{\mathrm{d}}{\mathrm{d}t}\big(f \circ \theta^{-1}\big)\big(\mathbf{y}_i(t)\big)\bigg|_{t=0} = u^i \frac{\partial (f \circ \theta^{-1})}{\partial y^i}(y), \tag{4}$$

where $(y, u) = \Theta(P, U) = (\theta(P), U\theta)$, and we have used the Einstein summation convention, that indices appearing once as a superscript and once as a subscript are summed out.

For each $\alpha \in [-1, 1]$, let $\mathcal{D}_\alpha : N \times N \to [0, \infty)$ be the α-divergence

$$\mathcal{D}_\alpha(P \mid Q) := \begin{cases} \mathbf{E}_Q \dfrac{\mathrm{d}P}{\mathrm{d}Q} \log \dfrac{\mathrm{d}P}{\mathrm{d}Q}, & \text{if } \alpha = -1, \\ \dfrac{4}{1-\alpha^2}\left(1 - \mathbf{E}_Q\left(\dfrac{\mathrm{d}P}{\mathrm{d}Q}\right)^{(1-\alpha)/2}\right), & \text{if } \alpha \in (-1, 1), \\ \mathbf{E}_Q \log \dfrac{\mathrm{d}Q}{\mathrm{d}P}, & \text{if } \alpha = 1. \end{cases} \tag{5}$$

(The Kullback–Leibler divergence corresponds to the case $\alpha = -1$.) These are of class C^∞; their mixed second derivatives define the *Fisher metric* as a Riemannian metric on N: for any $P \in N$, any $U, V \in T_P N$, and any $\alpha \in [-1, 1]$,

$$\langle U, V \rangle_P := -UV\mathcal{D}_\alpha = u^i g(P)_{i,j} v^j, \tag{6}$$

where U and V act on the first and second argument of $\mathcal{D}_\alpha$, respectively, and

$$g(P)_{i,j} := \langle \partial_i, \partial_j \rangle_P = \mathbf{E}_P(\eta_i - \mathbf{E}_P \eta_i)(\eta_j - \mathbf{E}_P \eta_j). \tag{7}$$

The mixed third derivatives of the α-divergences define a family of *covariant derivatives* on N. If $\mathbf{U}, \mathbf{V} : N \to TN$ are sufficiently smooth vector fields of N then the Chentsov–Amari α-covariant derivative is defined as follows:

$$\nabla^\alpha_{\mathbf{U}} \mathbf{V}(P) = \mathbf{U}\mathbf{v}^k(P)\partial_k + \Gamma_\alpha(P)^k_{i,j}\mathbf{u}(P)^i \mathbf{v}(P)^j \partial_k. \tag{8}$$

Here $\mathbf{u}(P) = \mathbf{U}(P)\theta$, $\mathbf{v}(P) = \mathbf{V}(P)\theta$, and the *Christoffel symbols* are as follows

$$\begin{aligned} \Gamma_\alpha(P)^k_{i,j} &= -g(P)^{k,l}\partial_i \partial_j \partial_l \mathcal{D}_\alpha \\ &= \frac{1-\alpha}{2} g(P)^{k,l} \mathbf{E}_P(\eta_i - \mathbf{E}_P \eta_i)(\eta_j - \mathbf{E}_P \eta_j)(\eta_l - \mathbf{E}_P \eta_l), \end{aligned} \tag{9}$$

where $g(P)^{k,l}$ is the (k, l) element of the inverse of the matrix $g(P)$, ∂_i and ∂_j act on the first argument of $\mathcal{D}_\alpha$, and ∂_l acts on the second argument.

The covariant derivatives ∇^α and $\nabla^{-\alpha}$ are *dual* in the sense that, for appropriately smooth vector fields $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$,

$$\mathbf{U}\langle \mathbf{V}, \mathbf{W}\rangle_P = \left\langle \nabla^\alpha_{\mathbf{U}}\mathbf{V}, \mathbf{W}\right\rangle_P + \left\langle \mathbf{V}, \nabla^{-\alpha}_{\mathbf{U}}\mathbf{W}\right\rangle_P. \tag{10}$$

Each α-covariant derivative defines a notion of *parallel transport* on the tangent bundle. Equation (10) shows that, if two tangent vectors at base point P are parallel transported along a differentiable curve, one according to ∇^α and the other according to $\nabla^{-\alpha}$, then their Fisher dot product remains constant. The α-covariant derivatives thus generalise the Levi–Civita covariant derivative of Riemannian geometry, which corresponds to the special case $\alpha = 0$.

Information geometry is the study of such structures, and has a history going back to the work of Rao [22]. It derives its importance from the fundamental role played by the Fisher information in estimation theory. An example application in asymptotic statistics is to decompose the error of a *second-order efficient estimator* into a term arising from the choice of the estimator and terms arising from the curvature of the parametric model from which the estimate is chosen. (See Chapter 4 in [2].) For more applications, from a variety of fields, the reader is referred to [19].

The literature on information geometry is dominated by the study of finite-dimensional manifolds of probability measures (parametric models) such as (N, B, θ) above. See [1,2,4,7,15] and the references therein for further information. However, these are not always sufficiently inclusive for the Bayesian applications outlined in Section 1, and any extension of the ideas to the non-parametric case must be based on charts with respect to which the α-divergences are suitably smooth. As is clear from the first equation in (5), the smoothness properties of the Kullback–Leibler divergence are closely connected with those of the density, $\mathrm{d}P/\mathrm{d}Q$, and its log (considered as elements of dual function spaces). In the series of papers [6,10,20,21], G. Pistone and his coworkers developed an infinite-dimensional exponential statistical manifold on an abstract probability space $(\mathbb{X}, \mathcal{X}, \mu)$. (See, also, [11,25].) Probability measures in the manifold are mutually absolutely continuous with respect to the reference measure μ, and the manifold is covered by the charts $s_Q(P) = \log \mathrm{d}P/\mathrm{d}Q - \mathbf{E}_Q \log \mathrm{d}P/\mathrm{d}Q$ for different "patch-centric" probability measures Q. These readily give $\log \mathrm{d}P/\mathrm{d}Q$ the desired regularity, but require ranges that are subsets of exponential Orlicz spaces in order to do the same for $\mathrm{d}P/\mathrm{d}Q$. The exponential Orlicz manifold is the natural infinite-dimensional extension of the exponential manifold (N, B, θ) described above; it has a strong topology, under which the α-divergences are of class C^∞. Variants of the Chentsov–Amari covariant derivatives are defined on it in [10]. However, with the exception of the case $\alpha = 1$, they are not defined on the tangent bundle. If $\alpha \in (-1, 1)$, for example, the α-connection is defined on the vector bundle whose fibre at base point P is the Lebesgue space $L^{2/(1-\alpha)}(P)$.

This approach is highly inclusive, but is technically demanding and leads to manifolds that are larger than needed in many applications. The author's aim in [17] and the present paper was to construct simpler infinite-dimensional statistical manifolds appropriate to problems in Bayesian estimation. The manifolds we construct differ from one another in the numbers of derivatives that the α-divergences admit. A minimal requirement is a mixed second derivative since this is needed in the construction of the Fisher metric. It is achieved in a Hilbert setting in [17]. However, it is also useful for the α-divergences to admit higher derivatives so that notions of parallel transport can be developed. This is achieved here in the context of Banach manifolds.

3. The manifolds of finite measures

Let $(\mathbb{X}, \mathcal{X}, \mu)$ be a probability space. For some $\lambda \in [2, \infty)$, we consider the set, $\tilde{M}$ $(= \tilde{M}_\lambda)$, of finite measures on $\mathcal{X}$ satisfying the following conditions:

(M1) P is mutually absolutely continuous with respect to μ;
(M2) $\mathbf{E}_\mu p^\lambda < \infty$;
(M3) $\mathbf{E}_\mu |\log p|^\lambda < \infty$.

(We denote measures in $\tilde{M}$ by the upper-case letters P, Q, ..., and their densities with respect to μ by the corresponding lower case letters, p, q, ...) In order to control both the density p and its log, we employ a "balanced" chart involving their sum. Let $\tilde{G}$ $(= \tilde{G}_\lambda := L^\lambda(\mu))$ be the Lebesgue space of (equivalence classes of) random variables $\tilde{a} : \mathbb{X} \to \mathbb{R}$ for which $\mathbf{E}_\mu |\tilde{a}|^\lambda < \infty$, and let $\tilde{\phi} : \tilde{M} \to \tilde{G}$ be defined as follows:

$$\tilde{\phi}(P) = p - 1 + \log p. \tag{11}$$

Proposition 3.1. *$\tilde{\phi}$ is a bijection onto $\tilde{G}$.*

Proof. For $y \in (0, \infty)$ let $\theta(y) = y - 1 + \log y$; then $\inf_y \theta(y) = -\infty$, $\sup_y \theta(y) = +\infty$, and θ is of class C^∞ with first derivative $\theta^{(1)}(y) = 1 + y^{-1} > 0$. So, according to the inverse function theorem, $\theta : (0, \infty) \to \mathbb{R}$ is a diffeomorphism. Let $\psi : \mathbb{R} \to (0, \infty)$ be its inverse; we have

$$\begin{aligned} \psi(z) &= \theta^{-1}(z) = W \circ \exp(z + 1), \\ \psi^{(1)}(z) &= \frac{1}{\theta^{(1)} \circ \psi(z)} = \frac{\psi(z)}{1 + \psi(z)} \in (0, 1), \end{aligned} \tag{12}$$

where $W : (0, \infty) \to (0, \infty)$ is the Lambert W function. In particular, ψ is strictly increasing, convex, and satisfies a linear growth condition. So, for any $\tilde{a} \in \tilde{G}$,

$$\mathbf{E}_\mu \psi(\tilde{a})^\lambda < K\left(1 + \mathbf{E}_\mu |\tilde{a}|^\lambda\right) < \infty \quad \text{and} \quad \mathbf{E}_\mu \left|\log \psi(\tilde{a})\right|^\lambda = \mathbf{E}_\mu \left|\tilde{a} - \psi(\tilde{a})\right|^\lambda < \infty.$$

Let P be the measure on $\mathcal{X}$ with density $p = \psi(\tilde{a})$; then P satisfies (M1)–(M3), and $\tilde{\phi}(P) = \tilde{a}$, and this completes the proof. □

This construction defines an infinite-dimensional manifold of measures, $(\tilde{M}, \tilde{G}, \tilde{\phi})$, with an atlas comprising the single chart $\tilde{\phi}$. The inverse map $\tilde{\phi}^{-1} : \tilde{G} \to \tilde{M}$ takes the form

$$\frac{\mathrm{d}\tilde{\phi}^{-1}(\tilde{a})}{\mathrm{d}\mu}(x) = \psi\big(\tilde{a}(x)\big), \tag{13}$$

where ψ is as defined in (12). (The definition of ψ used here is slightly different from that in [17]; in fact $\psi_{\text{here}}(z) = \psi_{\text{there}}(z + 1)$. The definition used here has the advantage that $\tilde{\phi}(\mu) = 0$.) As in Section 2, we consider a tangent vector U at $P \in \tilde{M}$ to be an equivalence class of differentiable curves at P: two curves, $(\tilde{\mathbf{a}}(t) \in \tilde{G}, t \in (-\varepsilon, \varepsilon))$ and $(\tilde{\mathbf{b}}(t) \in \tilde{G}, t \in (-\varepsilon, \varepsilon))$, being equivalent at P if $\tilde{\mathbf{a}}(0) = \tilde{\mathbf{b}}(0) = \tilde{\phi}(P)$ and $\dot{\tilde{\mathbf{a}}}(0) = \dot{\tilde{\mathbf{b}}}(0)$. We denote the tangent space at P by $T_P \tilde{M}$, and the

tangent bundle by $T\tilde{M} := \bigcup_{P\in\tilde{M}}(P, T_P\tilde{M})$. The latter admits the global chart $\tilde{\Phi} : T\tilde{M} \to \tilde{G} \times \tilde{G}$ where

$$\tilde{\Phi}(P, U) = \big(\tilde{\mathbf{a}}(0), \dot{\tilde{\mathbf{a}}}(0)\big), \tag{14}$$

and $\tilde{\mathbf{a}}$ is any differentiable curve in the equivalence class U. If $f : \tilde{M} \to Y$ is a map with range Y (a Banach space) and the map $f \circ \tilde{\phi}^{-1} : \tilde{G} \to Y$ is (Fréchet) differentiable, then we write

$$Uf := \frac{\mathrm{d}}{\mathrm{d}t}\big(f \circ \tilde{\phi}^{-1}\big)\big(\tilde{\mathbf{a}}(t)\big)\bigg|_{t=0} = D\big(f \circ \tilde{\phi}^{-1}\big)_{\tilde{a}}\tilde{u},$$

where $(\tilde{a}, \tilde{u}) = \tilde{\Phi}(P, U) = (\tilde{\phi}(P), U\tilde{\phi})$.

Remark 3.1. The weaker notion of *d-differentiability* is defined in [17]. In the present context, the map $f : \tilde{M} \to Y$ is d-differentiable if, for any $P \in \tilde{M}$, there exists a continuous linear map $d(f \circ \tilde{\phi}^{-1})_{\tilde{a}} : \tilde{G} \to Y$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(f \circ \tilde{\phi}^{-1}\big)\big(\tilde{\mathbf{a}}(t)\big)\bigg|_{t=0} = d\big(f \circ \tilde{\phi}^{-1}\big)_{\tilde{a}}\tilde{u},$$

for all differentiable curves $\tilde{\mathbf{a}}$ in the equivalence class U. (See Definition 3.1 in [17].) We then write $Uf = d(f \circ \tilde{\phi}^{-1})_{\tilde{a}}\tilde{u}$. Clearly, if f is Fréchet differentiable then it is also d-differentiable, and the derivatives are identical. However, the converse is not always true, as demonstrated by Example 3.1 in [17].

For any $\alpha \in [-1, 1]$, let $\xi_\alpha : \mathbb{R} \to \mathbb{R}$ be defined by

$$\xi_\alpha(z) = \begin{cases} \dfrac{2}{1-\alpha}\psi(z)^{(1-\alpha)/2}, & \text{if } \alpha \in [-1, 1), \\ \log\psi(z) = z + 1 - \psi(z), & \text{if } \alpha = 1, \end{cases} \tag{15}$$

where ψ is as defined in (12). Let $\xi_\alpha^{(i)}$ be the ith derivative of ξ_α. An induction argument shows that all such derivatives are bounded, the first being

$$\xi_\alpha^{(1)}(z) = \frac{\psi(z)^{(1-\alpha)/2}}{1 + \psi(z)} \in (0, 1). \tag{16}$$

For any $\alpha \in [-1, 1]$ and any $r \in [1, \lambda]$, let $\Xi_\alpha^r : \tilde{G} \to L^r(\mu)$ be defined by

$$\Xi_\alpha^r(\tilde{a})(x) = \xi_\alpha\big(\tilde{a}(x)\big). \tag{17}$$

Ξ_α^r is the *superposition* (*Nemytskij*) *operator* associated with the nonlinear function ξ_α, the domain $\tilde{G}$ and the range $L^r(\mu)$. The differentiability properties of such operators are developed in an abstract setting in Chapter 3 of [3]. In the present context, we are able to exploit the explicit nature of ξ_α to give a direct, self-contained proof of the following.

Lemma 3.1. (i) Ξ_α^r *is of class* $C^{\lceil\lambda/r\rceil-1}$, *where* $\lceil y\rceil := \min\{i \in \mathbb{Z}\colon\ y \le i\}$. *For any* $1 \le i \le \lceil\lambda/r\rceil - 1$, $D^i\,\Xi_\alpha^r : \tilde{G} \to L(\tilde{G}^i; L^r(\mu))$ *is given by*

$$D^i\,\Xi^r_{\alpha,\tilde{a}}(\tilde{u}_1,\dots,\tilde{u}_i)(x) = \xi_\alpha^{(i)}\big(\tilde{a}(x)\big)\tilde{u}_1(x)\cdots\tilde{u}_i(x). \tag{18}$$

(ii) Ξ_α^r *satisfies global Lipschitz continuity and linear growth conditions*, *and*, *for any* $1 \le i \le \lceil\lambda/r\rceil - 1$,

$$\sup_{\tilde{a}\in\tilde{G}} \big\| D^i\,\Xi^r_{\alpha,\tilde{a}} \big\| < \infty. \tag{19}$$

Proof. Let $l := \lceil\lambda/r\rceil - 1$, let $\tilde{a}, \tilde{u}_1, \dots, \tilde{u}_l \in \tilde{G}$, and let $(\tilde{a}_n \ne \tilde{a}, n \in \mathbb{N})$ be a sequence converging to $\tilde{a}$ in $\tilde{G}$. For convenience of notation, let $\xi_\alpha^{(0)} := \xi_\alpha$. If $l \ge 1$ then the mean value theorem applied on an x-by-x basis shows that, for any $0 \le i \le l-1$,

$$\big(\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a})\big)\tilde{u}_1\cdots\tilde{u}_i = \xi_\alpha^{(i+1)}(\tilde{a})\tilde{u}_1\cdots\tilde{u}_i(\tilde{a}_n - \tilde{a}) + R_n,$$

where $R_n := S_n\tilde{u}_1\cdots\tilde{u}_i(\tilde{a}_n - \tilde{a})$, and, for some $\beta = \beta(i, \tilde{a}_n(x), \tilde{a}(x)) \in [0,1]$,

$$S_n := \xi_\alpha^{(i+1)}\big((1-\beta)\tilde{a} + \beta\tilde{a}_n\big) - \xi_\alpha^{(i+1)}(\tilde{a}).$$

Now $r(i+1) < \lambda$ and so, setting $s = \lambda/(\lambda - r(i+1))$, we have $1/s + (i+1)r/\lambda = 1$ and Hölder's inequality shows that

$$\mathbf{E}_\mu|R_n|^r \le \big(\mathbf{E}_\mu|S_n|^{rs}\big)^{1/s}\big(\mathbf{E}_\mu|\tilde{u}_1|^\lambda\big)^{r/\lambda}\cdots\big(\mathbf{E}_\mu|\tilde{u}_i|^\lambda\big)^{r/\lambda}\big(\mathbf{E}_\mu|\tilde{a}_n - \tilde{a}|^\lambda\big)^{r/\lambda}.$$

So

$$\|\tilde{a}_n - \tilde{a}\|^{-1} \sup_{\|\tilde{u}_k\|=1} \|R_n\|_{L^r(\mu)} \le \|S_n\|_{L^{rs}(\mu)}.$$

Now $S_n \to 0$ in probability, and is bounded by $2\sup_z|\xi_\alpha^{(i+1)}(z)|$, and so it follows from the bounded convergence theorem that $\|S_n\|_{L^{rs}(\mu)} \to 0$. An induction argument on i thus establishes that Ξ_α^r admits Fréchet derivatives up to order l, and that these derivatives take the form (18).

For any $0 \le i \le l$, let $T_n := (\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a}))\tilde{u}_1\cdots\tilde{u}_i$. A similar argument to that used above shows that

$$\sup_{\|\tilde{u}_k\|=1} \|T_n\|_{L^r(\mu)} \le \big\|\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a})\big\|_{L^{rt}(\mu)}, \tag{20}$$

where $t = \lambda/(\lambda - ri)$. If $i = 0$ then the mean value theorem and Jensen's inequality show that

$$\big\|\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a})\big\|_{L^{rt}(\mu)} \le 2\sup_z\big|\xi_\alpha^{(i+1)}(z)\big|\,\|\tilde{a}_n - \tilde{a}\|_{\tilde{G}},$$

which shows that Ξ_α^r satisfies global Lipschitz continuity and linear growth conditions. If $i > 0$ then $|\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a})| \to 0$ in probability, and is bounded by $2\sup_z|\xi_\alpha^{(i)}(z)|$. In either case, the right-hand side of (20) converges to zero, and this shows that Ξ_α^r, and any derivatives it has, are

continuous. This completes the proof of part (i). The global boundedness of the derivatives in (18) follows from the boundedness of those of ξ_α, and this completes the proof of part (ii). □

Lemma 3.1 will be used in the following sections. It also determines the differentiability properties of Amari's α-embedding maps, F_α, [2]. In the present context $F_\alpha : \tilde{M} \to L^2(\mu)$, and is defined by

$$F_\alpha(P) = \Xi^2_\alpha\big(\tilde{\phi}(P)\big), \tag{21}$$

where Ξ^r_α is as defined in (17). The choice of $L^2(\mu)$ for the range of F_α is consistent with the latter's role in the definition of the α-divergences. (See (35) and the expressions for the derivatives of $\mathcal{D}_\alpha$ in Section 5.)

Corollary 3.1. *For any $\alpha \in [-1, 1]$, the map F_α is of class $C^{\lceil \lambda/2 \rceil - 1}$.*

Proof. This is an immediate consequence of Lemma 3.1 with $r = 2$. □

In the case $\lambda = 2$, where $\tilde{G}$ is a Hilbert space, F_α is continuous but not differentiable. It is, however, d-differentiable in the sense described in Remark 3.1. (See Proposition 3.1 in [17].) More generally, if $\lambda = 2n$ for some $n \in \mathbb{N}$, then F_α is of class C^{n-1}, and its highest Fréchet derivative is d-differentiable. However, the d-derivative may not be continuous.

4. The manifolds of probability measures

Let M be the subset of $\tilde{M}$ whose members are *probability* measures. These satisfy (M1)–(M3) and the additional hypothesis:

(M4) $\mathbf{E}_\mu p = 1$.

Let G $(= G_\lambda := L^\lambda_0(\mu))$ be the Lebesgue space of (equivalence classes of) random variables $a : \mathbb{X} \to \mathbb{R}$ for which $\mathbf{E}_\mu |a|^\lambda < \infty$ and $\mathbf{E}_\mu a = 0$, and let $\phi : M \to G$ be defined as follows:

$$\phi(P) = \tilde{\phi}(P) - \mathbf{E}_\mu \tilde{\phi}(P) = p - 1 + \log p - \mathbf{E}_\mu \log p. \tag{22}$$

Proposition 4.1.

(i) *ϕ is a bijection onto G.*
(ii) *(M, G, ϕ) is a $C^{\lceil \lambda \rceil - 1}$-embedded submanifold of $(\tilde{M}, \tilde{G}, \tilde{\phi})$.*
(iii) *Let $\rho := \tilde{\phi} \circ \phi^{-1}$ be the inclusion map $\iota : M \to \tilde{M}$ expressed in terms of the charts $\tilde{\phi}$ and ϕ. For any bounded set $B \subset G$,*

$$\sup_{a \in B} \big(\|\rho(a)\| + \|D\rho_a\| + \cdots + \|D^{\lceil \lambda \rceil - 1} \rho_a\| \big) < \infty. \tag{23}$$

Proof. Let $l := \lceil \lambda \rceil - 1$. Let $\Psi : G \times \mathbb{R} \to (0, \infty)$ be defined by

$$\Psi(a, z) = \mathbf{E}_\mu \psi(a + z) = \mathbf{E}_\mu \Xi^1_{-1}(a + z),$$

where ψ is as in (12) and Ξ_α^r is as in (17). It follows from Lemma 3.1, with $r = 1$, that Ψ is of class C^l and that, for any $u \in G$ and any $y \in \mathbb{R}$,

$$D\Psi_{a,z}(u, y) = \mathbf{E}_\mu \psi^{(1)}(a + z)u + \mathbf{E}_\mu \psi^{(1)}(a + z)y. \tag{24}$$

For any $a \in G$, let $\Psi_a : \mathbb{R} \to (0, \infty)$ be defined by $\Psi_a(z) = \Psi(a, z)$; then

$$\Psi_a^{(1)}(z) = \mathbf{E}_\mu \psi^{(1)}(a + z) > 0.$$

Since ψ is convex,

$$\sup_z \Psi_a \geq \sup_z \psi\big(\mathbf{E}_\mu(a + z)\big) = \sup_z \psi(z) = +\infty;$$

furthermore, the monotone convergence theorem shows that

$$\lim_{z \downarrow -\infty} \Psi_a = \mathbf{E}_\mu \lim_{z \downarrow -\infty} \psi(a + z) = 0.$$

Thus $\Psi_a : \mathbb{R} \to (0, \infty)$ is a bijection with strictly positive derivative, and the inverse function theorem shows that it is a C^l-isomorphism. The implicit function theorem shows that $Z : G \to \mathbb{R}$, defined by $Z(a) = \Psi_a^{-1}(1)$, is of class C^l. According to (24), its first derivative takes the form:

$$DZ_a u = -\frac{\mathbf{E}_\mu \psi^{(1)}(a + Z(a))u}{\mathbf{E}_\mu \psi^{(1)}(a + Z(a))}. \tag{25}$$

Let P be the probability measure on $\mathcal{X}$ with density $p = \psi(a + Z(a))$; then it follows from (12) and the mean value theorem that, for any $x \in \mathbb{X}$,

$$\big|p(x) - \psi\big(Z(a)\big)\big| \leq \big|a(x)\big| \quad \text{and} \quad \big|\log p(x) - \log \psi\big(Z(a)\big)\big| \leq \big|a(x)\big|.$$

So $P \in M$, and

$$\phi(P) = \theta \circ \psi\big(a + Z(a)\big) - \mathbf{E}_\mu \log \psi\big(a + Z(a)\big) = a + Z(a) - \mathbf{E}_\mu \log \psi\big(a + Z(a)\big).$$

Now $\phi(P) - a \in G$, and so

$$Z(a) = \mathbf{E}_\mu \log \psi\big(a + Z(a)\big) = -\mathcal{D}_{+1}(P \mid \mu), \tag{26}$$

and $\phi(P) = a$, which completes the proof of part (i).

According to (22) and (26), for any $a \in G$,

$$\rho(a) = a + \mathbf{E}_\mu \log \psi\big(a + Z(a)\big) = a + Z(a), \tag{27}$$

and so ρ is also of class C^l. ρ is injective, as is its first derivative; in fact, for any $\tilde{b} \in \rho(G)$ and any $\tilde{v} \in D\rho_a G$,

$$\rho^{-1}(\tilde{b}) = \tilde{b} - \mathbf{E}_\mu \tilde{b} \quad \text{and} \quad D\rho_a^{-1}\tilde{v} = \tilde{v} - \mathbf{E}_\mu \tilde{v},$$

from which it also follows that ρ and $D\rho_a$ are topological embeddings. Since $D\rho_a$ is also a linear map, it is a *toplinear isomorphism*, and its image $D\rho_a G$ is a closed linear subspace of $\tilde{G}$. Suppose, in the special case that $\lambda = 2$, that $\tilde{v} \in \tilde{G}$ $(= \tilde{G}_2)$ is orthogonal to $D\rho_a G$ $(= D\rho_a G_2)$. It is a consequence of the representation (25) that

$$\mathbf{E}_\mu \tilde{v}\left(u - \frac{\mathbf{E}_\mu \psi^{(1)}(\tilde{a})u}{\mathbf{E}_\mu \psi^{(1)}(\tilde{a})}\right) = \mathbf{E}_\mu \tilde{v} D\rho_a u = 0 \qquad \text{for all } u \in G_2,$$

where $\tilde{a} = \rho(a)$. It then readily follows that

$$\left\langle \tilde{v}\mathbf{E}_\mu \psi^{(1)}(\tilde{a}) - \psi^{(1)}(\tilde{a})\mathbf{E}_\mu \tilde{v}, u\right\rangle_{G_2} = 0 \qquad \text{for all } u \in G_2.$$

So the orthogonal complement of $D\rho_a G_2$ in $\tilde{G}_2$ is the one-dimensional subspace,

$$E_a := \left\{ y\psi^{(1)}(\tilde{a}), y \in \mathbb{R}\right\}. \tag{28}$$

Since $\psi^{(1)}$ is bounded, E_a is also a one-dimensional subspace of $\tilde{G}_\lambda$ for any $\lambda \in [2, \infty)$. Now $D\rho_a G_\lambda = \tilde{G}_\lambda \cap D\rho_a G_2$, and so $D\rho_a G_\lambda \oplus E_a = \tilde{G}_\lambda$ and $D\rho_a G_\lambda \cap E_a = \{0\}$. We have thus shown that $D\rho_a$ *splits* $\tilde{G}$ into the complementary closed subspaces $D\rho_a G$ and E_a. It thus follows from Proposition 2.3 of Chapter II in [13] that ρ is a C^l*-immersion*. Since ρ is a topological embedding it is also a C^l-embedding, and this completes the proof of part (ii).

It follows from Jensen's inequality and (12) that, for any $a \in G$,

$$\begin{aligned} -\log \mathbf{E}_\mu \psi^{(1)}(\tilde{a}) &\le -\mathbf{E}_\mu \log \psi^{(1)}(\tilde{a}) \\ &= -\mathbf{E}_\mu \log p + \mathbf{E}_\mu \log(1+p) \\ &\le \mathcal{D}_{+1}(P \mid \mu) + \log 2, \end{aligned} \tag{29}$$

where $\tilde{a} = \rho(a)$ and $P = \phi^{-1}(a)$. Now

$$\begin{aligned} \mathcal{D}_{-1}(P \mid \mu) + \mathcal{D}_{+1}(P \mid \mu) &= \mathbf{E}_\mu (p-1)\big(\log p + \mathcal{D}_{+1}(P \mid \mu)\big) \\ &\le \mathbf{E}_\mu \big(p - 1 + \log p + \mathcal{D}_{+1}(P \mid \mu)\big)^2/2 \\ &\le \|a\|_G^2/2, \end{aligned} \tag{30}$$

and so, since they are both non-negative, $\mathcal{D}_{-1}(\phi^{-1} \mid \mu)$ and $\mathcal{D}_{+1}(\phi^{-1} \mid \mu)$ are bounded on bounded sets. Together with (29), this proves that

$$\inf_{a \in B} \mathbf{E}_\mu \psi^{(1)}\big(\rho(a)\big) > 0. \tag{31}$$

The boundedness of the derivatives of ρ on bounded sets follows from (31), the boundedness of the derivatives of ψ, and an induction argument. The boundedness of ρ on bounded sets follows from (26), (27) and (30), and this completes the proof of part (iii). □

The tangent space at base point $P \in M$, $T_P M$, can be defined in the same way as was $T_P \tilde{M}$. The tangent bundle, $TM := \bigcup_{P \in M} (P, T_P M)$, admits the global chart $\Phi : TM \to G \times G$, where

$$\Phi(P, U) = \big(\mathbf{a}(0), \dot{\mathbf{a}}(0)\big) = \big(\phi(P), U\phi\big), \tag{32}$$

and $\mathbf{a}$ is any differentiable curve in the equivalence class U. For any $P \in M$, the tangent space $T_P M$ is a subspace of $T_P \tilde{M}$ of co-dimension 1; in fact

$$T_P \tilde{M} = T_P M \oplus \{yU_0, y \in \mathbb{R}\}, \tag{33}$$

where U_0 is the equivalence class of differentiable curves on $\tilde{M}$ containing the curve $(\tilde{\mathbf{a}}(t) := \tilde{a} + t\psi^{(1)}(\tilde{a}), t \in (-\varepsilon, \varepsilon))$, and $\tilde{a} = \tilde{\phi}(P)$. (See (28).)

Corollary 4.1. *The map $F_\alpha \circ \imath : M \to L^2(\mu)$, where F_α is as defined in (21), is of class $C^{\lceil \lambda/2 \rceil - 1}$.*

Proof. This follows from Corollary 3.1 and Proposition 4.1(ii). □

5. The α-divergences

We begin by investigating the regularity of the α-divergences on $\tilde{M}$. The usual extension of the α-divergences of (5) to sets of finite measures such as $\tilde{M}$ is as follows [2]:

$$\mathcal{D}_\alpha(P \mid Q) := \begin{cases} Q(\mathbb{X}) - P(\mathbb{X}) + \mathbf{E}_\mu\, p \log(p/q), & \text{if } \alpha = -1, \\ \dfrac{2}{1+\alpha} P(\mathbb{X}) + \dfrac{2}{1-\alpha} Q(\mathbb{X}) - \dfrac{4}{1-\alpha^2} \mathbf{E}_\mu\, p^{(1-\alpha)/2} q^{(1+\alpha)/2}, & \\ & \text{if } \alpha \in (-1, 1), \\ P(\mathbb{X}) - Q(\mathbb{X}) + \mathbf{E}_\mu\, q \log(q/p), & \text{if } \alpha = 1. \end{cases} \tag{34}$$

These can be represented in terms of the maps F_α of (21); for example,

$$\frac{4}{1-\alpha^2} \mathbf{E}_\mu\, p^{(1-\alpha)/2} q^{(1+\alpha)/2} = \big\langle F_\alpha(P), F_{-\alpha}(Q) \big\rangle_{L^2(\mu)}. \tag{35}$$

So, for any $\alpha \in [-1, 1]$ and any $P, Q \in \tilde{M}$, $\mathcal{D}_\alpha(P \mid Q) < \infty$, and we refer to elements of $\tilde{M}$ as "finite-entropy" measures. We could investigate the smoothness properties of $\mathcal{D}_\alpha$ starting from those of F_α. However, this approach would show only that the divergences are of class $C^{\lceil \lambda/2 \rceil - 1}$; a stronger result can be obtained by a more direct approach. The following lemma, which is similar in nature to Lemma 3.1, prepares the ground. For any $\alpha \in [-1, 1]$, let $\Upsilon_\alpha : \tilde{G} \times \tilde{G} \to L^1(\mu)$ be the following superposition operator:

$$\Upsilon_\alpha(\tilde{a}, \tilde{b})(x) = \xi_\alpha\big(\tilde{a}(x)\big)\xi_{-\alpha}\big(\tilde{b}(x)\big), \tag{36}$$

where ξ_α is as in (15).

Lemma 5.1. *For any $0 \le i, j \le \lfloor\lambda\rfloor - 1$ with $i + j \le \lceil\lambda\rceil - 1$, the map Υ_α is of class $C^{i,j}$. Its partial derivatives, $\Upsilon_\alpha^{(i,j)} := D_1^i D_2^j \Upsilon_\alpha : \tilde{G} \times \tilde{G} \to L(\tilde{G}^{i+j}; L^1(\mu))$, are given by*

$$\begin{aligned}&\Upsilon_\alpha^{(i,j)}(\tilde{a}, \tilde{b})(\tilde{u}_1, \dots, \tilde{u}_i; \tilde{v}_1 \dots, \tilde{v}_j)(x)\\ &\quad = \xi_\alpha^{(i)}\big(\tilde{a}(x)\big)\xi_{-\alpha}^{(j)}\big(\tilde{b}(x)\big)\tilde{u}_1(x)\cdots\tilde{u}_i(x)\tilde{v}_1(x)\cdots\tilde{v}_j(x).\end{aligned} \tag{37}$$

Proof. Let $0 \le i \le \lfloor\lambda\rfloor - 2$, $0 \le j \le \lfloor\lambda\rfloor - 1$ and $i + j \le \lceil\lambda\rceil - 2$. Let $\tilde{a}, \tilde{b}$, $\tilde{u}_1, \dots, \tilde{u}_i, \tilde{v}_1, \dots, \tilde{v}_j \in \tilde{G}$ and let $(\tilde{a}_n \ne \tilde{a}, n \in \mathbb{N})$ be a sequence converging to $\tilde{a}$ in $\tilde{G}$. Applying the mean value theorem on an x-by-x basis, we obtain

$$\big(\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a})\big)\xi_{-\alpha}^{(j)}(\tilde{b})\tilde{u}_1\cdots\tilde{v}_j = \xi_\alpha^{(i+1)}(\tilde{a})\xi_{-\alpha}^{(j)}(\tilde{b})\tilde{u}_1\cdots\tilde{v}_j(\tilde{a}_n - \tilde{a}) + R_n,$$

where $R_n := S_n\tilde{u}_1\cdots\tilde{v}_j(\tilde{a}_n - \tilde{a})$ and, for some $\beta = \beta(i, j, \tilde{a}_n(x), \tilde{a}(x), \tilde{b}(x)) \in [0, 1]$,

$$S_n := \big(\xi_\alpha^{(i+1)}\big((1-\beta)\tilde{a} + \beta\tilde{a}_n\big) - \xi_\alpha^{(i+1)}(\tilde{a})\big)\xi_{-\alpha}^{(j)}(\tilde{b}).$$

Now $i + j + 1 < \lambda$ and so, setting $s = \lambda/(\lambda - i - j - 1)$, we have $1/s + (i + j + 1)/\lambda = 1$, and Hölder's inequality shows that

$$\mathbf{E}_\mu|R_n| \le \big(\mathbf{E}_\mu|S_n|^s\big)^{1/s}\big(\mathbf{E}_\mu|\tilde{u}_1|^\lambda\big)^{1/\lambda}\cdots\big(\mathbf{E}_\mu|\tilde{v}_j|^\lambda\big)^{1/\lambda}\big(\mathbf{E}_\mu|\tilde{a}_n - \tilde{a}|^\lambda\big)^{1/\lambda},$$

so that

$$\|\tilde{a}_n - \tilde{a}\|^{-1} \sup_{\|\tilde{u}_k\| = \|\tilde{v}_k\| = 1} \|R_n\|_{L^1(\mu)} \le \|S_n\|_{L^s(\mu)}.$$

Now $S_n \to 0$ in probability, and is dominated by $f := 2\sup_z |\xi_\alpha^{(i+1)}(z)||\xi_{-\alpha}^{(j)}(\tilde{b})|$. If $j = 0$ then $f \in \tilde{G}$ and $s \le \lambda$, whereas if $j \ge 1$ then $f \in L^\infty(\mu)$ and $s < \infty$. In either case the dominated convergence theorem shows that $\|S_n\|_{L^s(\mu)} \to 0$, so that $\Upsilon_\alpha^{(i,j)}$ is differentiable in its first argument, with derivative $\Upsilon_\alpha^{(i+1,j)}$. Similarly, if $0 \le i \le \lfloor\lambda\rfloor - 1$, $0 \le j \le \lfloor\lambda\rfloor - 2$, and $i + j \le \lceil\lambda\rceil - 2$ then $\Upsilon_\alpha^{(i,j)}$ is differentiable in its second argument, with derivative $\Upsilon_\alpha^{(i,j+1)}$. An induction argument on i and j thus establishes (37).

It remains to show that, for any $0 \le i, j \le \lfloor\lambda\rfloor - 1$ with $i + j = \lceil\lambda\rceil - 1$, $\Upsilon_\alpha^{(i,j)}$ is continuous. Now

$$\big(\Upsilon_\alpha^{(i,j)}(\tilde{a}_n, \tilde{b}_n) - \Upsilon_\alpha^{(i,j)}(\tilde{a}, \tilde{b})\big)(\tilde{u}_1, \dots, \tilde{v}_j) = (T_{1,n} + T_{2,n} + T_{3,n})\tilde{u}_1\cdots\tilde{v}_j,$$

where

$$\begin{aligned}T_{1,n} &:= \big(\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a})\big)\xi_{-\alpha}^{(j)}(\tilde{b}), \qquad T_{2,n} := \xi_\alpha^{(i)}(\tilde{a})\big(\xi_{-\alpha}^{(j)}(\tilde{b}_n) - \xi_{-\alpha}^{(j)}(\tilde{b})\big),\\ T_{3,n} &:= \big(\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a})\big)\big(\xi_{-\alpha}^{(j)}(\tilde{b}_n) - \xi_{-\alpha}^{(j)}(\tilde{b})\big),\end{aligned}$$

and similar arguments to those used above show that

$$\big\|\Upsilon_\alpha^{(i,j)}(\tilde{a}_n, \tilde{b}_n) - \Upsilon_\alpha^{(i,j)}(\tilde{a}, \tilde{b})\big\| \le \|T_{1,n}\|_{L^t(\mu)} + \|T_{2,n}\|_{L^t(\mu)} + \|T_{3,n}\|_{L^t(\mu)},$$

where $t = \lambda/(\lambda - i - j)$. We will thus have established the continuity of $\Upsilon_\alpha^{(i,j)}$ if we can show that

$$\|T_{k,n}\|_{L^t(\mu)} \to 0 \qquad \text{for } k = 1, 2, 3. \tag{38}$$

If $i = j = 0$, then $t = 1$ and (38) follows from the Cauchy–Schwarz inequality and the mean value theorem; for example,

$$\begin{aligned}\|T_{1,n}\|_{L^1(\mu)}^2 &\le \big\|\xi_\alpha(\tilde a_n) - \xi_\alpha(\tilde a)\big\|_{L^2(\mu)} \big\|\xi_{-\alpha}(\tilde b)\big\|_{L^2(\mu)} \\ &\le \sup_z \big|\xi_\alpha^{(1)}(z)\big| \|\tilde a_n - \tilde a\|_{L^2(\mu)} \sup_z \big|\xi_{-\alpha}^{(1)}(z)\big| \|\tilde b\|_{L^2(\mu)} \to 0.\end{aligned}$$

If $i, j > 0$, then $t < \infty$, and $T_{k,n} \to 0$ in probability and is bounded for all k; so (38) follows from the bounded convergence theorem. If $i = 0$ and $j > 0$, then $t \le \lambda$ and

$$\|T_{k,n}\|_{L^t(\mu)} \le 2 \sup_z \big|\xi_{-\alpha}^{(j)}(z)\big| \big\|\xi_\alpha(\tilde a_n) - \xi_\alpha(\tilde a)\big\|_{L^t(\mu)} \to 0 \qquad \text{for } k = 1, 3;$$

furthermore $T_{2,n} \to 0$ in probability and is dominated by $2 \sup_z |\xi_{-\alpha}^{(j)}(z)||\xi_\alpha(\tilde a)| \in \tilde G$, and so the dominated convergence theorem establishes (38). The case $i > 0$ and $j = 0$ can be treated in the same way, and this completes the proof. □

Corollary 5.1. *For any $\alpha \in [-1, 1]$, and any $0 \le i, j \le \lfloor \lambda \rfloor - 1$ with $i + j \le \lceil \lambda \rceil - 1$, the α-divergence $\mathcal{D}_\alpha : \tilde M \times \tilde M \to [0, \infty)$ is of class $C^{i,j}$.*

Proof. It follows from (5), (16), (17) and (36) that

$$\mathcal{D}_\alpha(P \mid Q) := \begin{cases} \mathbf{E}_\mu\big(\Xi_{-1}^1(\tilde b) - \Xi_{-1}^1(\tilde a) + \Upsilon_{-1}(\tilde a, \tilde a) - \Upsilon_{-1}(\tilde a, \tilde b)\big), & \text{if } \alpha = -1, \\ \mathbf{E}_\mu\Big(\dfrac{2}{1+\alpha}\Xi_{-1}^1(\tilde a) + \dfrac{2}{1-\alpha}\Xi_{-1}^1(\tilde b) - \Upsilon_\alpha(\tilde a, \tilde b)\Big), & \text{if } \alpha \in (-1, 1), \\ \mathbf{E}_\mu\big(\Xi_{-1}^1(\tilde a) - \Xi_{-1}^1(\tilde b) + \Upsilon_1(\tilde b, \tilde b) - \Upsilon_1(\tilde a, \tilde b)\big), & \text{if } \alpha = 1, \end{cases}$$

where $\tilde a = \tilde\phi(P)$ and $\tilde b = \tilde\phi(Q)$. The corollary thus follows from Lemma 3.1 (with $r = 1$) and Lemma 5.1. □

Straightforward calculations show that, for any $\tilde a, \tilde b, \tilde u, \tilde v \in \tilde G$,

$$\begin{aligned} D_1\mathcal{D}_\alpha\big(\tilde\phi^{-1} \mid \tilde\phi^{-1}\big)_{\tilde a,\tilde b}\tilde u &= \mathbf{E}_\mu\big(\Upsilon_\alpha^{(1,0)}(\tilde a, \tilde a) - \Upsilon_\alpha^{(1,0)}(\tilde a, \tilde b)\big)\tilde u, \\ D_2\mathcal{D}_\alpha\big(\tilde\phi^{-1} \mid \tilde\phi^{-1}\big)_{\tilde a,\tilde b}\tilde v &= \mathbf{E}_\mu\big(\Upsilon_\alpha^{(0,1)}(\tilde b, \tilde b) - \Upsilon_\alpha^{(0,1)}(\tilde a, \tilde b)\big)\tilde v. \end{aligned} \tag{39}$$

If $\lambda > 2$, these admit the following representations

$$\begin{aligned} U\mathcal{D}_\alpha(\cdot \mid Q) &= \big\langle F_{-\alpha}(P) - F_{-\alpha}(Q), UF_\alpha\big\rangle_{L^2(\mu)}, \\ V\mathcal{D}_\alpha(P \mid \cdot) &= \big\langle F_\alpha(Q) - F_\alpha(P), VF_{-\alpha}\big\rangle_{L^2(\mu)}, \end{aligned} \tag{40}$$

and $\mathcal{D}_\alpha$ admits the following mixed second derivative

$$D_1 D_2 \mathcal{D}_\alpha\big(\tilde{\phi}^{-1} \mid \tilde{\phi}^{-1}\big)_{\tilde{a},\tilde{b}}(\tilde{u},\tilde{v}) = -\mathbf{E}_\mu \Upsilon_\alpha^{(1,1)}(\tilde{a},\tilde{b})(\tilde{u},\tilde{v}) = -\langle UF_\alpha, VF_{-\alpha}\rangle_{L^2(\mu)}, \tag{41}$$

where $(P,U) = \tilde{\Phi}^{-1}(\tilde{a},\tilde{u})$ and $(Q,V) = \tilde{\Phi}^{-1}(\tilde{b},\tilde{v})$. Setting $\tilde{b} = \tilde{a}$, we obtain the following definition of the (extended) Fisher metric on $T_P\tilde{M}$: for any $U, V \in T_P\tilde{M}$,

$$\langle U, V\rangle_P := -UV\mathcal{D}_\alpha = \langle UF_\alpha, VF_{-\alpha}\rangle_{L^2(\mu)}. \tag{42}$$

Remark 5.1. The representations in (40) and (41), and the definition in (42), are also valid for the case $\lambda = 2$ if the weaker notion of d-differentiability is used in the definitions of UF_α, $VF_{-\alpha}$ and $UV\mathcal{D}_\alpha$. (See [17].)

It follows from (16) and (41) that $\langle V, U\rangle_P = \langle U, V\rangle_P$, and that, for any $s \in \mathbb{R}$, $\langle sU, V\rangle_P = \langle U, sV\rangle_P = s\langle U, V\rangle_P$. Furthermore,

$$\langle U, U\rangle_P = \mathbf{E}_\mu \frac{p}{(1+p)^2}\tilde{u}^2 \le \mathbf{E}_\mu \tilde{u}^2 \le \|\tilde{u}\|_{\tilde{G}}^2, \tag{43}$$

where $\tilde{u} = U\tilde{\phi}$; in particular $\langle U, U\rangle_P = 0$ if and only if $\tilde{u} = 0$. Thus, $(T_P\tilde{M}, \langle\cdot\rangle_P)$ is an inner product space. As shown in (43), the Fisher norm is dominated by the natural Banach norm on $T_P\tilde{M}$. However, it is not equivalent to that norm, even in the case $\lambda = 2$. (See [17].) In the general, infinite-dimensional case $(T_P\tilde{M}, \langle\cdot\rangle_P)$ is not a Hilbert space; the Fisher metric is a pseudo-Riemannian metric but not a Riemannian metric.

If $\lambda > 3$, $\mathcal{D}_\alpha$ admits the following mixed third derivative

$$D_1^2 D_2 \mathcal{D}_\alpha\big(\tilde{\phi}^{-1} \mid \tilde{\phi}^{-1}\big)_{\tilde{a},\tilde{b}}(\tilde{u},\tilde{v};\tilde{w}) = -\mathbf{E}_\mu \Upsilon_\alpha^{(2,1)}(\tilde{a},\tilde{b})(\tilde{u},\tilde{v};\tilde{w}). \tag{44}$$

Setting $\tilde{b} = \tilde{a}$ and carrying out some straightforward calculations, we obtain

$$D_1^2 D_2 \mathcal{D}_\alpha\big(\tilde{\phi}^{-1} \mid \tilde{\phi}^{-1}\big)_{\tilde{a},\tilde{a}}(\tilde{u},\tilde{v};\tilde{w}) = -\mathbf{E}_\mu \frac{p}{(1+p)^2}\tilde{\Gamma}_\alpha(\tilde{a},\tilde{u},\tilde{v})\tilde{w}, \tag{45}$$

where $\tilde{\Gamma}_\alpha : \tilde{G}\times\tilde{G}\times\tilde{G} \to L^{\lambda/2}(\mu)$ is defined by

$$\tilde{\Gamma}_\alpha(\tilde{a},\tilde{u},\tilde{v})(x) = \frac{1-\alpha}{2}\frac{\tilde{u}(x)\tilde{v}(x)}{(1+p(x))^2} - \frac{1+\alpha}{2}p(x)\frac{\tilde{u}(x)\tilde{v}(x)}{(1+p(x))^2}. \tag{46}$$

If $\tilde{a} = \tilde{\phi}(P)$, and $\tilde{u}$ and $\tilde{v}$ are such that $\tilde{\Gamma}_\alpha(\tilde{a},\tilde{u},\tilde{v}) \in \tilde{G}$, then there exist tangent vectors $Y, W \in T_P\tilde{M}$ such that $\tilde{\Gamma}_\alpha(\tilde{a},\tilde{u},\tilde{v}) = Y\tilde{\phi}$ and $\tilde{w} = W\tilde{\phi}$. In this case

$$D_1^2 D_2 \mathcal{D}_\alpha\big(\tilde{\phi}^{-1} \mid \tilde{\phi}^{-1}\big)_{\tilde{a},\tilde{a}}(\tilde{u},\tilde{v};\tilde{w}) = -\langle Y, W\rangle_P. \tag{47}$$

For any $l \in \mathbb{N}_0$ and any $s \in [1,\infty]$, let $\tilde{\mathcal{V}}_s^l$ be the set of vector fields $\mathbf{V} : \tilde{M} \to T\tilde{M}$ for which $\tilde{\mathbf{v}}(P)(:= \mathbf{V}(P)\tilde{\phi}) \in L^{s\lambda}(\mu)$ for all $P \in \tilde{M}$, and $\tilde{\mathbf{v}} : \tilde{M} \to L^{s\lambda}(\mu)$ is of class C^l. For any $\mathbf{U} \in \tilde{\mathcal{V}}_s^0$, we can use (47) and the Eguchi relations [9] to define an "α-derivative" $\tilde{\nabla}_{\mathbf{U}}^\alpha : \tilde{\mathcal{V}}_{s/(s-1)}^1 \to \tilde{\mathcal{V}}_1^0$, as

follows

$$\tilde{\nabla}^{\alpha}_{\mathbf{U}}\mathbf{V} := \tilde{\Phi}^{-1}\big(\tilde{\phi}, \mathbf{U}\tilde{\mathbf{v}} + \tilde{\Gamma}_{\alpha}(\tilde{\phi}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}})\big). \tag{48}$$

However, this does not define an operator, $\tilde{\nabla}^{\alpha}$, with domain $\tilde{\mathcal{V}}^0_1 \times \tilde{\mathcal{V}}^1_1$, and so it does not define a full covariant derivative on $T\tilde{M}$. With the exception of the $+1$ connection on the exponential Orlicz manifold, this appears to be an insuperable problem in infinite dimensions. In order for the divergences to be sufficiently smooth, the tangent space must be given a stronger topology than that generated by the Fisher metric, and so it is incomplete with respect to the latter. This creates difficulties with the projection methods at the heart of the definition of α-covariant derivatives. In the special case that $s = \infty$, $\tilde{\nabla}^{\alpha}_{\mathbf{U}}\mathbf{V}$ is well defined for all C^1 vector fields $\mathbf{V}$, and thus provides a limited notion of α-parallel transport on the tangent bundle. (See [11] for a similar result on the exponential Orlicz manifold.)

A straightforward calculation shows that, for any $\alpha \in [-1, 1]$, $\mathbf{U} \in \tilde{\mathcal{V}}^0_s$ and $\mathbf{V}, \mathbf{W} \in \tilde{\mathcal{V}}^1_{s/(s-1)}$,

$$\mathbf{U}\langle \mathbf{V}, \mathbf{W}\rangle_P = \big\langle \tilde{\nabla}^{\alpha}_{\mathbf{U}}\mathbf{V}, \mathbf{W}\big\rangle_P + \big\langle \mathbf{V}, \tilde{\nabla}^{-\alpha}_{\mathbf{U}}\mathbf{W}\big\rangle_P, \tag{49}$$

reflecting the duality (10) of the finite-dimensional case.

5.1. The Fisher metric and α-derivatives on (M, G, ϕ)

In the above, we used the α-divergences and Eguchi relations to define the extended Fisher metric and α-derivatives on the manifold $\tilde{M}$. Clearly, we could follow the same approach with the submanifold M. (It follows from Proposition 4.1(ii) and Corollary 5.1 that the α-divergences have the same smoothness properties on M as they have on $\tilde{M}$.) For any $P \in M$, the Fisher metric on $T_P M$, thus obtained, is a restriction of the extended Fisher metric on $T_P\tilde{M}$, as defined above. (See (33).) On the other hand, the definition of the α-derivative involves second derivatives of $\mathcal{D}_{\alpha}$ in one variable, and so the correspondence between $\tilde{M}$ and M is not so transparent.

For some $s \in [1, \infty]$, let $\mathbf{U} \in \tilde{\mathcal{V}}^0_s$ and $\mathbf{V} \in \tilde{\mathcal{V}}^1_{s/(s-1)}$ be vector fields on $\tilde{M}$, whose restrictions to M are vector fields of M; then, for any $P \in M$, $\tilde{\Phi}(\mathbf{U}(P)) = (\rho(a), D\rho_a\mathbf{u}(P))$ and $\tilde{\Phi}(\mathbf{V}(P)) = (\rho(a), D\rho_a\mathbf{v}(P))$, where $(a, \mathbf{u}(P)) = \Phi(\mathbf{U}(P))$ and $(a, \mathbf{v}(P)) = \Phi(\mathbf{V}(P))$. So, according to (48),

$$\begin{aligned}
\tilde{\nabla}^{\alpha}_{\mathbf{U}}\mathbf{V} &= \tilde{\Phi}^{-1}\big(\tilde{\phi}, \mathbf{U}(D\rho\mathbf{v}) + \tilde{\Gamma}_{\alpha}(\tilde{\phi}, D\rho\mathbf{u}, D\rho\mathbf{v})\big) \\
&= \tilde{\Phi}^{-1}\big(\tilde{\phi}, D\rho\mathbf{U}\mathbf{v} + D^2\rho(\mathbf{u}, \mathbf{v}) + \tilde{\Gamma}_{\alpha}(\tilde{\phi}, D\rho\mathbf{u}, D\rho\mathbf{v})\big) \\
&= \tilde{\Phi}^{-1}\left(\tilde{\phi}, D\rho\mathbf{U}\mathbf{v} + \frac{1-\alpha}{2}\gamma - \frac{1+\alpha}{2}\eta\right),
\end{aligned} \tag{50}$$

where $\gamma, \eta : \tilde{M} \to \tilde{G}$ are defined by

$$\begin{aligned}
\gamma(P)(x) &= \frac{D\rho_a\mathbf{u}(P)(x)D\rho_a\mathbf{v}(P)(x)}{(1+p(x))^2} + D^2\rho_a\big(\mathbf{u}(P), \mathbf{v}(P)\big)(x), \\
\eta(P)(x) &= p(x)\frac{D\rho_a\mathbf{u}(P)(x)D\rho_a\mathbf{v}(P)(x)}{(1+p(x))^2} - D^2\rho_a\big(\mathbf{u}(P), \mathbf{v}(P)\big)(x),
\end{aligned} \tag{51}$$

and $a = \phi(P)$. It follows from (25) and (27) that, for any $u \in G$,

$$D\rho_a u = u - \frac{\mathbf{E}_\mu \psi^{(1)}(\rho(a))u}{\mathbf{E}_\mu \psi^{(1)}(\rho(a))} = u - \frac{\mathbf{E}_\mu D\Xi^1_{-1,\rho(a)}u}{\mathbf{E}_\mu D\Xi^1_{-1,\rho(a)}1},$$

and so, according to the quotient and chain rules of differentiation, and Lemma 3.1,

$$D^2\rho_a(u,v) = -\frac{\mathbf{E}_\mu \psi^{(2)}(\rho(a))D\rho_a u D\rho_a v}{\mathbf{E}_\mu \psi^{(1)}(\rho(a))} = -\frac{1}{\mathbf{E}_\mu \psi^{(1)}(\rho(a))}\mathbf{E}_\mu \psi^{(1)}\big(\rho(a)\big)\frac{D\rho_a u D\rho_a v}{(1+p)^2}.$$

From these derivatives and (51), we conclude that

$$\begin{aligned}
\gamma(P) &= D\rho_a\left(\frac{\tilde{\mathbf{u}}(P)\tilde{\mathbf{v}}(P)}{(1+p)^2} - \mathbf{E}_\mu \frac{\tilde{\mathbf{u}}(P)\tilde{\mathbf{v}}(P)}{(1+p)^2}\right),\\
\eta(P) &= D\rho_a\left(p\frac{\tilde{\mathbf{u}}(P)\tilde{\mathbf{v}}(P)}{(1+p)^2} - p\big\langle \mathbf{U}(P), \mathbf{V}(P)\big\rangle_P\right)\\
&\quad + \big\langle \mathbf{U}(P), \mathbf{V}(P)\big\rangle_P\left(p - \frac{\mathbf{E}_\mu \psi^{(1)}(\tilde{a})p}{\mathbf{E}_\mu \psi^{(1)}(\tilde{a})}\right)\\
&\quad + \frac{1}{\mathbf{E}_\mu \psi^{(1)}(\tilde{a})}\mathbf{E}_\mu \psi^{(1)}(\tilde{a})\left(p\frac{\tilde{\mathbf{u}}(P)\tilde{\mathbf{v}}(P)}{(1+p)^2} + \frac{\tilde{\mathbf{u}}(P)\tilde{\mathbf{v}}(P)}{(1+p)^2}\right)\\
&= D\rho_a\left(p\frac{\tilde{\mathbf{u}}(P)\tilde{\mathbf{v}}(P)}{(1+p)^2} - p\big\langle \mathbf{U}(P), \mathbf{V}(P)\big\rangle_P\right)\\
&\quad + \big\langle \mathbf{U}(P), \mathbf{V}(P)\big\rangle_P\left(p - \frac{1}{\mathbf{E}_\mu \psi^{(1)}(\tilde{a})} + 1\right) + \frac{1}{\mathbf{E}_\mu \psi^{(1)}(\tilde{a})}\mathbf{E}_\mu p\frac{\tilde{\mathbf{u}}(P)\tilde{\mathbf{v}}(P)}{(1+p)^2}\\
&= D\rho_a\left(p\frac{\tilde{\mathbf{u}}(P)\tilde{\mathbf{v}}(P)}{(1+p)^2} - p\big\langle \mathbf{U}(P), \mathbf{V}(P)\big\rangle_P\right) + (1+p)\big\langle \mathbf{U}(P), \mathbf{V}(P)\big\rangle_P,
\end{aligned}$$

where $\tilde{a} = \tilde{\phi}(P)$, $\tilde{\mathbf{u}}(P) = D\rho_a \mathbf{u}(P)$, $\tilde{\mathbf{v}}(P) = D\rho_a \mathbf{v}(P)$, and we have used the fact that $\psi^{(1)}\psi = \psi - \psi^{(1)}$ in the second step. We have thus shown that

$$\tilde{\nabla}^\alpha_{\mathbf{U}}\mathbf{V}(P) = \tilde{\Phi}^{-1}\left(\tilde{a}, D\rho_a\big(\mathbf{U}\mathbf{v}(P) + \Gamma_\alpha\big(a, \mathbf{u}(P), \mathbf{v}(P)\big)\big) - \frac{1+\alpha}{2}(1+p)\big\langle \mathbf{U}(P), \mathbf{V}(P)\big\rangle_P\right),$$

where $\Gamma_\alpha : G \times G \times G \to L_0^{\lambda/2}(\mu)$ is defined by

$$\begin{aligned}
\Gamma_\alpha(a,u,v)(x) &= \frac{1-\alpha}{2}\left(\frac{D\rho_a u(x) D\rho_a v(x)}{(1+p(x))^2} - \mathbf{E}_\mu \frac{D\rho_a u(x) D\rho_a v(x)}{(1+p(x))^2}\right)\\
&\quad - \frac{1+\alpha}{2}p(x)\left(\frac{D\rho_a u(x) D\rho_a v(x)}{(1+p(x))^2} - \langle U, V\rangle_P\right).
\end{aligned} \tag{52}$$

For any $W \in T_P M$

$$\begin{aligned}\big\langle \tilde{\nabla}^{\alpha}_{\mathbf{U}}\mathbf{V}(P), W\big\rangle_P &= \mathbf{E}_\mu \frac{p}{(1+p)^2} D\rho_a\big(\mathbf{U}\mathbf{v}(P) + \Gamma_\alpha\big(a, \mathbf{u}(P), \mathbf{v}(P)\big)\big) D\rho_a w \\ &\quad - \frac{1+\alpha}{2}\big\langle \mathbf{U}(P), \mathbf{V}(P)\big\rangle_P \mathbf{E}_\mu \frac{p}{(1+p)^2}(1+p) D\rho_a w \\ &= \big\langle \nabla^{\alpha}_{\mathbf{U}}\mathbf{V}(P), W\big\rangle_P,\end{aligned} \tag{53}$$

where $\nabla^{\alpha}_{\mathbf{U}}\mathbf{V} : M \to TM$ is the vector field on M defined by

$$\nabla^{\alpha}_{\mathbf{U}}\mathbf{V} = \Phi^{-1}\big(\phi, \mathbf{U}\mathbf{v} + \Gamma_\alpha(\phi, \mathbf{u}, \mathbf{v})\big). \tag{54}$$

As (53) shows, $\nabla^{\alpha}_{\mathbf{U}}\mathbf{V}(P)$ is the projection of $\tilde{\nabla}^{\alpha}_{\mathbf{U}}\mathbf{V}(P)$ onto $T_P M$, in the sense of the Fisher metric. The map $\nabla^{\alpha}_{\mathbf{U}} : \mathcal{V}^1_{s/(s-1)} \to \mathcal{V}^0_1$, thus defined, is the α-derivative on M, which could also found by direct calculation in the same way as was $\tilde{\nabla}^{\alpha}_{\mathbf{U}}$.

6. Finite-dimensional submanifolds

For some $d \in \mathbb{N}$ and $n \in \mathbb{N} \cup \{\infty\}$, let (N, B, θ) be a d-dimensional C^n-embedded submanifold of $\tilde{M}$. By this, we mean that $N \subset \tilde{M}$, B is a non-empty open subset of $\mathbb{R}^d$, $\theta : N \to B$ is a bijection, and the inclusion map $\tilde{\iota} : N \to \tilde{M}$ is both a topological embedding and a C^n-immersion. (See, e.g., [13].) As in Section 2, the tangent space at base point $P \in N$, $T_P N$, is spanned by the vectors $(\partial_i, i = 1, \ldots, d)$, where ∂_i is the equivalence class of differentiable curves on N containing the curve $(\mathbf{y}_i(t) := \theta(P) + t\mathbf{e}_i, t \in (-\varepsilon, \varepsilon))$. The matrix form of the (extended) Fisher metric is

$$g(P)_{i,j} := \langle \partial_i, \partial_j \rangle_P = \mathbf{E}_\mu \frac{p}{(1+p)^2} \tilde{w}_i \tilde{w}_j, \tag{55}$$

where $\tilde{w}_i = \partial_i \tilde{\phi}$.

Since $(T_P N, \langle\cdot\rangle_P)$ is a *finite-dimensional* inner-product space it is also a Euclidean space, and the Fisher metric is a Reimannian metric on N. If $\lambda > 3$ and $n \geq 2$, the full theory of α-covariant derivatives and their associated geometries can thus be developed on N. According to the Eguchi relations, the Christoffel symbols for the α-covariant derivative on (N, θ) are

$$\Gamma^N_\alpha(P)^k_{i,j} := -g(P)^{k,l}\partial_i \partial_j \partial_l \mathcal{D}_\alpha = g(P)^{k,l}\mathbf{E}_\mu \frac{p}{(1+p)^2} \tilde{\Gamma}_\alpha\big(\tilde{\phi}(P), \tilde{w}_i, \tilde{w}_j\big)\tilde{w}_l, \tag{56}$$

where $g(P)^{k,l}$ is the (k,l) element of the inverse of the matrix $g(P)$, ∂_i and ∂_j act on the first argument of $\mathcal{D}_\alpha$, ∂_l acts on the second argument of $\mathcal{D}_\alpha$, and $\tilde{\Gamma}_\alpha$ is as defined in (46).

If N is a *statistical* manifold (it is also a subset of M) then the inclusion map, $\iota : N \to M$, takes the form $\iota = \pi \circ \tilde{\iota}$, where $\pi = \phi^{-1} \circ \bar{\rho} \circ \tilde{\phi}$ and $\bar{\rho} : \tilde{G} \to G$ is defined by $\bar{\rho}(\tilde{a}) = \tilde{a} - \mathbf{E}_\mu \tilde{a}$. Clearly π is of class C^∞, and so ι is of class C^n. Furthermore, $\partial_i \tilde{\iota} \in T_P M$ for all i, and the restriction of the pushforward π_* to $T_P M$ is the identity map of $T_P M$, and so the derivative of ι is injective.

Since $\tilde{\imath}$ is a topological embedding and the map ρ of Proposition 4.1 is continuous, $\imath$ is also a topological embedding. It thus follows that N is also a C^n-embedded submanifold of M.

We finish with two examples of finite-dimensional submanifolds that illustrate the foregoing developments.

Example 6.1. Let $\eta_1, \ldots, \eta_d$ be linearly independent elements of $\tilde{G}$, let $\gamma : \mathbb{R}^d \to \tilde{G}$ be defined by $\gamma(y) = y^i \eta_i$, and let $N := \tilde{\phi}^{-1} \circ \gamma(\mathbb{R}^d)$. Since the η_i are linearly independent γ is an injection, and $(N, \mathbb{R}^d, \theta)$, with $\theta := \gamma^{-1} \circ \phi$, is a d-dimensional manifold. It is trivially a C^∞-embedded submanifold of $\tilde{M}$.

Example 6.2. Let (N, B, θ) be the d-dimensional exponential statistical manifold defined in Section 2, where the underlying space $(\mathbb{X}, \mathcal{X}, \mu)$ coincides with that of Sections 3–5, and suppose that the η_i and B are such that $\theta^{-1}(B) \subseteq M$. It is shown in Theorem 5.1 of [17] that N, thus defined, is a C^∞-embedded submanifold of M. (Strictly speaking, Theorem 5.1 in [17] addresses only the case $\lambda = 2$; however, the same proof carries over to the more general setting where $\lambda \in [2, \infty)$.)

7. Concluding remarks

Because of their role in the definition of the Kullback–Leibler divergence, it is natural to regard the density, p, and its log as belonging to dual function spaces. The choice of the exponential Orlicz space for $\log p$ (and, implicitly, its dual for p) yields the manifold of [21], comprising all probability measures in an absolute continuity equivalence class. The choice in [17] of the Hilbert space $L_0^2(\mu)$ for both p and $\log p$ leads to a significantly simpler construction, but at a cost to inclusiveness. This is also true of the Banach space approach taken here. However, this is unimportant in many applications (and may even be beneficial). In problems of Bayesian estimation, for example, we do not need manifolds to contain more than the posterior distributions associated with the various observations, and some finite-dimensional structures on which approximations can be based.

The choice of reference measure μ is important. The use of a *finite* measure is natural in the context of (M2) and (M3), and since the elements and topologies of $\tilde{M}$ and M are not affected by its total mass, it is also natural to assume that μ is a *probability* measure. If $\mathbb{X} = \mathbb{R}^n$ then (M1) is satisfied by all measures that are mutually absolutely continuous with respect to Lebesgue measure if, for example, μ is a non-singular multi-variate Gaussian measure. It may seem that one could construct larger manifolds by piecing together coordinate patches $(\tilde{M}_i, \tilde{G}_i, \tilde{\phi}_i, \mu_i)$, defined as in Section 3 but with differing patch-centric measures, μ_i. However, this is not possible since the requirement that $\mathrm{d}P/\mathrm{d}\mu_i \in L^\lambda(\mu_i)$ for each i is incompatible with the regularity of the transition maps $\tilde{\phi}_i \circ \tilde{\phi}_j^{-1}$ in all but trivial cases (such as that in which $\mathrm{d}\mu_i/\mathrm{d}\mu_j \in L^\infty(\mu_j)$ for all i, j). The requirement that $\mathrm{d}P/\mathrm{d}\mu \in L^\lambda(\mu)$ is stronger than needed for pure information geometry (even in the Hilbert case, $\lambda = 2$). However, it is useful in its own right. For example, in the context of Bayesian estimation, it yields bounds such as (1).

The role played by the exponential function in an exponential family, such as that of [21], is played here by the function ψ. In this sense, $\tilde{M}$ and M are extreme examples of *general deformed*

families, as defined in Chapter 10 of [16]. (They are extreme in the sense that ψ satisfies a linear growth condition.) General deformed families of probability measures are also developed and generalised in [24], where they are referred to as *φ-families*. The function φ is used there in the definition of the (Musielak–Orlicz) model spaces, and gives rise to dual divergence functions distinct from $\mathcal{D}_\alpha$. Here, our aim is somewhat different from those of [16] and [24]. We provide a simple framework for the classical information geometry in infinite dimensions; this requires a stronger topology on the model space than that associated with the φ-function ψ. (The Musielak–Orlicz spaces associated with the φ-function ψ have topologies that are too weak, even for the definition of the Fisher metric.)

Acknowledgement

The author would like to thank an anonymous referee for suggesting the offset -1 in the definition of $\tilde{\phi}$ in (11), which introduces a number of advantages.

References

[1] Amari, S.-I., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L. and Rao, C.R. (1978). *Differential Geometry and Statistical Inference*. *Lecture Note Monograph Series* **10**. Hayward, CA: IMS.
[2] Amari, S.-I. and Nagaoka, H. (2000). *Methods of Information Geometry*. *Translations of Mathematical Monographs* **191**. Providence, RI: Amer. Math. Soc. MR1800071
[3] Appell, J. and Zabrejko, P.P. (1990). *Nonlinear Superposition Operators*. Cambridge: Cambridge Univ. Press. MR1066204
[4] Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Chichester: Wiley. MR0489333
[5] Brigo, D., Hanzon, B. and Le Gland, F. (1999). Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli* **5** 495–534. MR1693600
[6] Cena, A. and Pistone, G. (2007). Exponential statistical manifold. *Ann. Inst. Statist. Math.* **59** 27–56. MR2396032
[7] Čencov, N.N. (1982). *Statistical Decision Rules and Optimal Inference*. *Translations of Mathematical Monographs* **53**. Providence, RI: Amer. Math. Soc. MR0645898
[8] Crisan, D. and Rozovskiĭ, B. (2011). *The Oxford Handbook of Nonlinear Filtering*. Oxford: Oxford Univ. Press. MR2882749
[9] Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. *Ann. Statist.* **11** 793–803. MR0707930
[10] Gibilisco, P. and Pistone, G. (1998). Connections on non-parametric statistical manifolds by Orlicz space geometry. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* **1** 325–347. MR1628177
[11] Grasselli, M.R. (2010). Dual connections in nonparametric classical information geometry. *Ann. Inst. Statist. Math.* **62** 873–896. MR2669742
[12] Grimmer, J. (2011). An introduction to Bayesian inference via variational approximations. *Polit. Anal.* **19** 32–47.
[13] Lang, S. (1999). *Fundamentals of Differential Geometry*. *Graduate Texts in Mathematics* **191**. New York: Springer. MR1666820
[14] Liptser, R.S. and Shiryayev, A.N. (1977). *Statistics of Random Processes. I*. New York: Springer. MR0474486

[15] Murray, M.K. and Rice, J.W. (1993). *Differential Geometry and Statistics*. *Monographs on Statistics and Applied Probability* **48**. London: Chapman & Hall. MR1293124
[16] Naudts, J. (2011). *Generalised Thermostatistics*. London: Springer. MR2777415
[17] Newton, N.J. (2012). An infinite-dimensional statistical manifold modelled on Hilbert space. *J. Funct. Anal.* **263** 1661–1681. MR2948226
[18] Newton, N.J. (2013). Infinite-dimensional manifolds of finite-entropy probability measures. In *Geometric Science of Information*. *Lecture Notes in Computer Science* **8085** 713–720. Heidelberg: Springer. MR3126105
[19] Nielsen, F. and Barbaresco, F., eds. (2013). Geometric science of information. In *Proceedings of the First International Conference*, *GSI* 2013, *Paris*, *France*, *August* 2013. *Lecture Notes in Computer Science* **8085**. Heidelberg: Springer.
[20] Pistone, G. and Rogantin, M.P. (1999). The exponential statistical manifold: Mean parameters, orthogonality and space transformations. *Bernoulli* **5** 721–760. MR1704564
[21] Pistone, G. and Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.* **23** 1543–1561. MR1370295
[22] Radhakrishna Rao, C. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** 81–91. MR0015748
[23] Šmidl, V. and Quinn, A. (2006). *The Variational Bayes Method in Signal Processing*. Berlin: Springer.
[24] Vigelis, R.F. and Cavalcante, C.C. (2013). On ϕ-families of probability distributions. *J. Theoret. Probab.* **26** 870–884. MR3090555
[25] Zhang, J. and Hästö, P. (2006). Statistical manifold as an affine space: A functional equation approach. *J. Math. Psych.* **50** 60–65. MR2208065

Received September 2013 and revised August 2014