Infinite-Dimensional Statistical Manifolds based on a Balanced Chart

NIGEL J. NEWTON

1 School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. 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 \(\alpha\)-divergences are of class \(C^{\lceil \lambda \rceil - 1}\), enabling the definition of the Fisher metric and \(\alpha\)-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 \(\alpha\)-covariant derivatives. \(\tilde{M}_\lambda\) and \(M_\lambda\) 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, Nonparametric 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 \(X : \Omega \to X\) and \(Y : \Omega \to Y\) are random variables defined on a common probability space \((\Omega, \mathcal{F}, \mathbb{P})\), taking values in metric spaces \(X\) and \(Y\), 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 \(X\). Under mild conditions (see, for example, [15]) an abstract Bayes formula defines a regular conditional distribution for \(X\) given \(Y\), \(\Pi : Y \to \mathcal{P}\). (For any Borel set \(B \subseteq X\), \(\Pi(\cdot)(B) : Y \to [0, 1]\) is measurable, and \(\mathbb{P}(X \in B \mid Y) = \Pi(Y)(B)\). In the applications we have in mind, \(\Pi(Y)\) is typically of infinite dimension and so we need to construct approximations of the form \(\hat{\Pi} : Y \to Q \subset \mathcal{P}\), where \(Q\) 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 \(P\), 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 \(\mu \in P\) 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 conditional mean of \(f\) under \(\hat{\Pi}(y)\) can be bounded by means of the Cauchy-Schwartz inequality:

\[
\left( \mathbb{E}_{\Pi(y)} f - \mathbb{E}_{\hat{\Pi}(y)} f \right)^2 \leq \mathbb{E}_\mu f^2 \mathbb{E}_\mu (\pi(y) - \hat{\pi}(y))^2.
\]

Although, in this context, the \(L^2\) metric on densities induces an appropriate topology on \(P\), 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 \(\mathbb{E}_{\Pi(Y)} f\) admits the orthogonal decomposition:

\[
\mathbb{E} \left( f(X) - \mathbb{E}_{\Pi(Y)} f \right)^2 = \mathbb{E} \mathbb{E}_{\Pi(Y)} (f - \mathbb{E}_{\Pi(Y)} f)^2 + \mathbb{E} \left( \mathbb{E}_{\Pi(Y)} f - \mathbb{E}_{\hat{\Pi}(Y)} f \right)^2.
\]

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 \(\Pi\). 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, \(\mathbb{E}_{\Pi(y)} f\), 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 \(P\):

\[
\mathcal{D}(Q \mid P) := \sup_{f \in \mathcal{L}^2(P)} \frac{(\mathbb{E}_P f - \mathbb{E}_Q f)^2}{\mathbb{E}_P (f - \mathbb{E}_P f)^2} = \sup_{f \in F} (\mathbb{E}_P f (1 - dQ/dP))^2 = \mathbb{E}_P (1 - dQ/dP)^2,
\]

where \(\mathcal{L}^2(P) = \{ f : X \to \mathbb{R} : \mathbb{E}_P f^2 < \infty \}\) and \(F\) is the subset of such functions having zero mean and unit variance. This is the \(\chi^2\)-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, for example, [12, 24].)
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 $\alpha$-embedding maps. Section 4 develops the family of manifolds of probability measures $((M_\lambda, \phi_\lambda), \lambda \in [2, \infty))$ in which the chart $\phi_\lambda$ is a “centred” version of $\tilde{\phi}_\lambda$. Section 5 studies the properties of the $\alpha$-divergences on $\tilde{M}_\lambda$ and $M_\lambda$, defining the Fisher metric, and a limited notion of $\alpha$-parallel transport on the tangent bundle. Some examples of finite-dimensional embedded submanifolds of $\tilde{M}_\lambda$ and $M_\lambda$ are outlined in section 6. A sketch of some of the results of sections 4 and 5.1 was given, without proofs, in [19].

### 2. Information Geometry

We begin by reviewing a classical finite-dimensional example: the exponential statistical manifold. (See, for example, [2].) Let $(\mathcal{X}, \mathcal{X}, \mu)$ 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 $L^0(\mu)$, i.e. $\mu(\alpha + \sum y^i \eta_i = 0) = 1$ if and only if $\alpha = 0$ and $\mathbb{R}^d \ni y = 0$; (ii) $\mathbb{E}_\mu \exp(\sum y^i \eta_i) < \infty$ 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
\[ \frac{dP_y}{d\mu} = \exp \left( \sum_i y^i \eta_i - c(y) \right), \tag{3} \]
where \( c(y) = \log E \exp(\sum_i y^i \eta_i) \), and let \( N := \{ P_y : y \in B \} \). 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 \((N,B,\theta)\) is an exponential statistical manifold, with an atlas comprising the single chart \( \theta \). 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), \((y(t) \in B, t \in (-\varepsilon,\varepsilon))\) and \((z(t) \in B, t \in (-\varepsilon,\varepsilon))\), being equivalent at \( P \) if \( y(0) = z(0) = \theta(P) \) and \( y(0) = z(0) \).

The tangent space at \( P, T_PN \), is the linear space of all such tangent vectors, and is spanned by the vectors \((\partial_i; i = 1,\ldots,d)\), where \( \partial_i \) is the equivalence class containing the curve \((y_i(t) := \theta(P) + t\varepsilon_i), t \in (-\varepsilon,\varepsilon))\), and \( \varepsilon_i \) is equal to the Kronecker delta. The tangent bundle is the disjoint union \( TN := \bigcup_{P \in N} (P, T_PN) \), 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 \( U \in T_PN \), then we write
\[ Uf = u^i \partial_i f := u^i \frac{d}{dt}(f \circ \theta^{-1})(y_i(t)) \bigg|_{t=0} = u^i \frac{\partial(f \circ \theta^{-1})}{\partial y^i}(y), \tag{4} \]
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 alpha-divergence
\[ \mathcal{D}_\alpha(P|Q) := \begin{cases} E_Q \frac{dP}{dQ} \log \frac{dP}{dQ} & \text{if } \alpha = -1 \\ \frac{4}{1-\alpha^2} \left( 1 - E_Q \left( \frac{dP}{dQ} \right)^{(1-\alpha)/2} \right) & \text{if } \alpha \in (-1,1) \\ E_Q \log \frac{dQ}{dP} & \text{if } \alpha = 1. \end{cases} \tag{5} \]
(The Kullback-Leibler divergence corresponds to the case \( \alpha = -1 \).) These are of class \( C^\infty \); their mixed second derivatives define the Fisher metric as a Riemannian metric on \( N \): for any \( P \in N \), any \( U, V \in T_PN \), 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
\[ g(P)_{i,j} := (\partial_i, \partial_j)_P = E_P(\eta_i - E_P \eta_i)(\eta_j - E_P \eta_j). \tag{7} \]
The mixed third derivatives of the alpha-divergences define a family of covariant derivatives on \( N \). If \( U, V : N \to TN \) are sufficiently smooth vector fields of \( N \) then the Chentsov-Amari alpha-covariant derivative is defined as follows:
\[ \nabla_U^\alpha V(P) = U v^k(P) \partial_k + \Gamma_\alpha(P)_{i,j}^k u(P)^i v(P)^j \partial_k. \tag{8} \]
Here $u(P) = U(P)\theta$, $v(P) = V(P)\theta$, and the Christoffel symbols are as follows

$$\Gamma^k_{ij}(P) = -g^{k,l}(P)\partial_i\partial_jD_\alpha,$$

where $g^{k,l}(P)$ is the $(k,l)$ element of the inverse of the matrix $g(P)$, and $\partial_i$ acts on the first argument of $D_\alpha$, and $\partial_l$ acts on the second argument.

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

$$U \langle V, W \rangle_p = \langle \nabla^\alpha_U V, W \rangle_p + \langle V, \nabla^{-\alpha}_U W \rangle_p.$$

Each $\alpha$-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 $\nabla^\alpha$ and the other according to $\nabla^{-\alpha}$, then their Fisher dot product remains constant. The $\alpha$-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 [23]. 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 [20].

The literature on information geometry is dominated by the study of finite-dimensional manifolds of probability measures (parametric models) such as $(N, B, \theta)$ above. See [1, 2, 4, 7, 14, 16] 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 $\alpha$-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, $dP/dQ$, and its log (considered as elements of dual function spaces). In the series of papers [6, 10, 21, 22], G. Pistone and his co-workers developed an infinite-dimensional exponential statistical manifold on an abstract probability space $(X, X, \mu)$. (See, also, [11, 26].) Probability measures in the manifold are mutually absolutely continuous with respect to the reference measure $\mu$, and the manifold is covered by the charts $s_Q(P) = \log dP/dQ - E_Q \log dP/dQ$ for different “patch-centric” probability measures $Q$. These readily give log $dP/dQ$ the desired regularity, but require ranges that are subsets of exponential Orlicz spaces in order to do the same for $dP/dQ$. The exponential Orlicz manifold is the natural infinite-dimensional extension of the exponential manifold $(N, B, \theta)$ described above; it has a strong topology, under which the $\alpha$-divergences are of class $C^\infty$. 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 \( \alpha \)-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 [18] 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 \( \alpha \)-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 [18]. However, it is also useful for the \( \alpha \)-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 \((X, \mathcal{X}, \mu)\) be a probability space. For some \( \lambda \in [2, \infty) \), we consider the set, \( \tilde{M}(= \tilde{M}_\lambda) \), of finite measures on \( X \) satisfying the following conditions:

\begin{enumerate}[label=(M\arabic*), ref=(M\arabic*)]
  \item \( P \) is mutually absolutely continuous with respect to \( \mu \);
  \item \( E_\mu p^\lambda < \infty \);
  \item \( E_\mu |\log p|^\lambda < \infty \).
\end{enumerate}

(We denote measures in \( \tilde{M} \) by the upper-case letters \( P, Q, \ldots \), and their densities with respect to \( \mu \) by the corresponding lower case letters, \( p, q, \ldots \) ) 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 \( \tilde{a}: X \to \mathbb{R} \) for which \( E_\mu |\tilde{a}|^\lambda < \infty \), 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 \( \theta \) is of class \( C^\infty \) 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

\[
\psi(z) = \theta^{-1}(z) = W \circ \exp(z + 1)
\]

\[
\psi^{(1)}(z) = \frac{1}{\theta^{(1)} \circ \psi(z)} = \frac{\psi(z)}{1 + \psi(z)} \in (0, 1), \tag{12}
\]

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

\[
E_\mu \psi(\tilde{a})^\lambda < K(1 + E_\mu |\tilde{a}|^\lambda) < \infty \quad \text{and} \quad E_\mu |\log \psi(\tilde{a})|^\lambda = E_\mu |\tilde{a} - \psi(\tilde{a})|^\lambda < \infty.
\]
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, $(\tilde{M}, G, \tilde{\phi})$, with an atlas comprising the single chart $\tilde{a}$. The inverse map $\tilde{\phi}^{-1} : G \to \tilde{M}$ takes the form

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

where $\psi$ is as defined in (12). (NB. the definition of $\psi$ used here is slightly different from that in [18]; 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 $u \in \tilde{M}$ to be an equivalence class of differentiable curves at $P$; two curves, $(\tilde{a}(t)) \in \tilde{G}$, $t \in (-\epsilon, \epsilon)$, and $(\tilde{b}(t)) \in \tilde{G}$, $t \in (-\epsilon, \epsilon)$, being equivalent at $P$ if $\tilde{a}(0) = \tilde{b}(0) = \tilde{\phi}(P)$ and $\dot{\tilde{a}}(0) = \dot{\tilde{b}}(0)$. We denote the tangent space at $P$ by $T_P \tilde{M}$, and the tangent bundle by $TM := \cup_{P \in \tilde{M}} (P, T_P \tilde{M})$. The latter admits the global chart $\tilde{\Phi} : T \tilde{M} \to \tilde{G} \times \tilde{G}$ where

$$\tilde{\Phi}(P, U) = (\tilde{a}(0), \dot{\tilde{a}}(0), U),$$

(14)

and $\tilde{a}$ 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

$$U f := \left. \frac{d}{dt} (f \circ \tilde{\phi}^{-1})(\tilde{a}(t)) \right|_{t=0} = D(f \circ \tilde{\phi}^{-1})\tilde{a} \tilde{u},$$

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 [18]. 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

$$\left. \frac{d}{dt} (f \circ \tilde{\phi}^{-1})(\tilde{a}(t)) \right|_{t=0} = d(f \circ \tilde{\phi}^{-1})_{\tilde{a}} \tilde{u},$$

for all differentiable curves $\tilde{a}$ in the equivalence class $U$. (See Definition 3.1 in [18].) We then write $U f = d(f \circ \tilde{\phi}^{-1})_{\tilde{a}} \tilde{u}$. 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 [18].

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

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

(15)

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

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

(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

\[
\Xi_\alpha^r(\tilde{a})(x) = \xi_\alpha(\tilde{a}(x)).
\]  

(17)

\( \Xi_\alpha^r \) is the superposition (Nemytskij) operator associated with the nonlinear function \( \xi_\alpha \), 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 \( \xi_\alpha \) to give a direct, self-contained proof of the following.

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

\[
D^i\Xi_\alpha^r(\tilde{a}_1, \ldots, \tilde{a}_i)(x) = \xi_\alpha^{(i)}(\tilde{a}(x))\tilde{a}_1(x) \cdots \tilde{a}_i(x).
\]  

(18)

(ii) \( \Xi_\alpha^r \) satisfies global Lipschitz continuity and linear growth conditions, and, for any \( 1 \leq i \leq [\lambda/r] - 1 \),

\[
\sup_{\tilde{a} \in \tilde{G}} \|D^i\Xi_\alpha^r(\tilde{a})\| < \infty.
\]  

(19)

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

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

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

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

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

\[
E_\mu|R_n|^r \leq (E_\mu|S_n|^r)^{1/s}(E_\mu|\tilde{a}_1|^\lambda)^{r/\lambda} \cdots (E_\mu|\tilde{a}_i|^\lambda)^{r/\lambda}(E_\mu|\tilde{a}_n - \tilde{a})^s|\tilde{a}|^{r/\lambda}.
\]

So

\[
\|\tilde{a}_n - \tilde{a}\|^{-1} \sup_{\|\tilde{a}_n\| = 1} \|R_n\|_{L^r(\mu)} \leq \|S_n\|_{L^r(\mu)}.
\]

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

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

\[
\sup_{\|\tilde{a}_n\| = 1} \|T_n\|_{L^r(\mu)} \leq \|\xi_\alpha^{(i)}(\tilde{a}_n) - \xi_\alpha^{(i)}(\tilde{a})\|_{L^r(\mu)}.
\]  

(20)
where $t = \lambda/(\lambda - r_i)$. If $i = 0$ then the mean value theorem and Jensen’s inequality show that

$$\|\xi^{(i)}_\alpha(\tilde{a}_n) - \xi^{(i)}_\alpha(\tilde{a})\|_{L^r(\mu)} \leq 2\sup_z |\xi^{(i+1)}_\alpha(z)|\|\tilde{a}_n - \tilde{a}\|_{\tilde{G}},$$

which shows that $\Xi^r_\alpha$ satisfies global Lipschitz continuity and linear growth conditions. If $i > 0$ then $|\xi^{(i)}_\alpha(\tilde{a}_n) - \xi^{(i)}_\alpha(\tilde{a})| \to 0$ in probability, and is bounded by $2\sup_z |\xi^{(i)}_\alpha(z)|$. In either case, the right-hand side of (20) converges to zero, and this shows that $\Xi^r_\alpha$, 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 $\xi_\alpha$, 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 $\alpha$-embedding maps, $F_\alpha$, [2]. In the present context $F_\alpha : \tilde{M} \to L^2(\mu)$, and is defined by

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

where $\Xi^r_\alpha$ is as defined in (17). The choice of $L^2(\mu)$ for the range of $F_\alpha$ is consistent with the latter’s role in the definition of the $\alpha$-divergences. (See (35) and the expressions for the derivatives of $D_\alpha$ in section 5.)

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

**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, $F_\alpha$ is continuous but not differentiable. It is, however, $d$-differentiable in the sense described in Remark 3.1. (See Proposition 3.1 in [18].) More generally, if $\lambda = 2n$ for some $n \in \mathbb{N}$, then $F_\alpha$ 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 $\tilde{M}$ 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 $G (= G_\lambda := L^2(\mu))$ 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:

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

**Proposition 4.1.** (i) $\phi$ is a bijection onto $G$.

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

\[
\sup_{a \in B} \left( \|\rho(a)\| + \|D\rho_a\| + \cdots + \|D^{[\lambda]}\rho_a\| \right) < \infty.
\]

(23)

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

\[
\Psi(a, z) = E_{\mu} \psi(\alpha + z) = E_{\mu} \Xi_{\alpha + z},
\]

where \( \psi \) is as in (12) and \( \Xi_{\alpha} \) is as in (17). It follows from Lemma 3.1, with \( r = 1 \), that \( \Psi \) is of class \( C^1 \) and that, for any \( u \in G \) and any \( y \in \mathbb{R} \),

\[
D\Psi_{a, z}(u, y) = E_{\mu} \psi^{(1)}(\alpha + z)u + E_{\mu} \psi^{(1)}(a + z)y.
\]

(24)

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

\[
\Psi_a^{(1)}(z) = E_{\mu} \psi^{(1)}(a + z) > 0.
\]

Since \( \psi \) is convex,

\[
\sup_z \Psi_a \geq \sup_z \psi(\mu(a + z)) = \sup_z \psi(z) = +\infty;
\]

furthermore, the monotone convergence theorem shows that

\[
\lim_{z \to -\infty} \Psi_a = E_{\mu} \lim_{z \to -\infty} \psi(\alpha + z) = 0.
\]

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 \( C^1 \)-isomorphism. The implicit function theorem shows that \( Z : G \to \mathbb{R} \), defined by \( Z(a) = \Psi_a^{-1}(1) \), is of class \( C^1 \). According to (24), its first derivative takes the form:

\[
DZ_a u = -\frac{E_{\mu} \psi^{(1)}(\alpha + Z(a))u}{E_{\mu} \psi^{(1)}(\alpha + Z(a))}.
\]

(25)

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

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

So \( P \in M \), and

\[
\phi(P) = \theta \circ \psi(a + Z(a)) - E_{\mu} \log \psi(a + Z(a)) = a + Z(a) - E_{\mu} \log \psi(a + Z(a)).
\]

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

\[
Z(a) = E_{\mu} \log \psi(a + Z(a)) = -D_{+1}(P | \mu),
\]

(26)
and \( \phi(P) = a \), which completes the proof of part (i).

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

\[
\rho(a) = a + E_\mu \log \psi(a + Z(a)) = a + Z(a),
\]

and so \( \rho \) is also of class \( C^1 \). \( \rho \) is injective, as is its first derivative; in fact, for any \( \tilde{b} \in \rho(G) \) and any \( \tilde{v} \in D\rho_a G \),

\[
\rho^{-1}(\tilde{b}) = \tilde{b} - E_\mu \tilde{b} \quad \text{and} \quad D\rho_a^{-1}\tilde{v} = \tilde{v} - E_\mu \tilde{v},
\]

from which it also follows that \( \rho \) 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 \( D\rho_a G (= D\rho_a \tilde{G}_2) \). It is a consequence of the representation (25) that

\[
E_\mu(\tilde{v} - \psi^{(1)}(\tilde{a})u) = E_\mu \tilde{v} D\rho_a u = 0 \quad \text{for all } u \in G_2,
\]

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

\[
\left\langle \tilde{v} - \psi^{(1)}(\tilde{a}) E_\mu \tilde{v} , u \right\rangle_{G_2} = 0 \quad \text{for all } u \in G_2.
\]

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

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

Since \( \psi^{(1)} \) is bounded, \( E_a \) is also a one-dimensional subspace of \( \tilde{G}_\lambda \) for any \( \lambda \in [2, \infty) \). Now \( D\rho_a G_\lambda = \tilde{G}_\lambda \cap D\rho_a \tilde{G}_2 \), 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 \( \rho \) is a \( C^1 \)-immersion. Since \( \rho \) is a topological embedding it is also a \( C^1 \)-embedding, and this completes the proof of part (ii).

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

\[
- \log E_\mu (\psi^{(1)}(\tilde{a})) \leq - E_\mu \log \psi(\tilde{a}) = - E_\mu \log p + E_\mu \log(1 + p) \leq D_{+1}(P \mid \mu) + \log 2,
\]

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

\[
D_{-1}(P \mid \mu) + D_{+1}(P \mid \mu) = E_\mu (p - 1) (\log p + D_{+1}(P \mid \mu)) \leq E_\mu (p - 1 + \log p + D_{+1}(P \mid \mu))^2 / 2 \leq \|a\|_{G_2}^2 / 2,
\]
and so, since they are both non-negative, $D_{-1}(\phi^{-1}|\mu)$ and $D_{+1}(\phi^{-1}|\mu)$ are bounded on bounded sets. Together with (29) this proves that
\[
\inf_{a \in B} E_{\mu}\psi^{(1)}(\rho(a)) > 0. \tag{31}
\]

The boundedness of the derivatives of $\rho$ on bounded sets follows from (31), the boundedness of the derivatives of $\psi$, and an induction argument. The boundedness of $\rho$ on bounded sets follows from (26), (27) and (30), and this completes the proof of part (iii).

The tangent space at base point $P \in M$, $T_PM$, can be defined in the same way as was $T_{\tilde{P}}\tilde{M}$. The tangent bundle, $TM := \cup_{P \in M}(P, T_PM)$, admits the global chart $\Phi : TM \rightarrow G \times G$, where $\Phi(P, U) = (a(0), \dot{a}(0)) = (\phi(P), U\phi)$, $\phi$ is any differentiable curve in the equivalence class $U$. For any $P \in \tilde{M}$, the tangent space $T_PM$ is a subspace of $T_{\tilde{P}}\tilde{M}$ of co-dimension 1; in fact
\[
T_PM = T_{\tilde{P}}\tilde{M} \oplus \{yU_0, y \in \mathbb{R}\}, \tag{33}
\]
where $U_0$ is the equivalence class of differentiable curves on $\tilde{M}$ containing the curve $(\tilde{a}(t) := \tilde{a} + t\psi^{(1)}(\tilde{a}))$, $t \in (-\epsilon, \epsilon)$, and $\tilde{a} = \tilde{\phi}(P)$. (See (28).)

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

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

## 5. The $\alpha$-Divergences

We begin by investigating the regularity of the $\alpha$-divergences on $\tilde{M}$. The usual extension of the $\alpha$-divergences of (5) to sets of finite measures such as $\tilde{M}$ is as follows [2]:
\[
D_\alpha(P|Q) := \begin{cases} 
\frac{Q(X) - P(X) + E_{\mu}p\log(p/q)}{1 - \alpha} & \text{if } \alpha = -1 \\
\frac{2}{1-\alpha} P(X) + \frac{2}{1-\alpha} Q(X) - \frac{4}{1-\alpha} E_{\mu}p^{(1-\alpha)/2}q^{(1+\alpha)/2} & \text{if } \alpha \in (-1, 1) \\
\frac{4}{1-\alpha} E_{\mu}p^{(1-\alpha)/2}q^{(1+\alpha)/2} & \text{if } \alpha = 1.
\end{cases}
\tag{34}
\]

These can be represented in terms of the maps $F_\alpha$ of (21); for example
\[
\frac{4}{1-\alpha} E_{\mu}p^{(1-\alpha)/2}q^{(1+\alpha)/2} = \langle F_\alpha(P), F_{-\alpha}(Q) \rangle_{L^2(\mu)}. \tag{35}
\]

So, for any $\alpha \in [-1, 1]$ and any $P, Q \in \tilde{M}$, $D_\alpha(P|Q) < \infty$, and we refer to elements of $\tilde{M}$ as “finite-entropy” measures. We could investigate the smoothness properties of $D_\alpha$.
Infinite-Dimensional Manifolds

starting from those of \( F_\alpha \). However, this approach would show only that the divergences
are of class \( C^{(\lambda/2)-1} \); 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:

\[
\Upsilon_\alpha(\tilde{a}, \tilde{b})(x) = \xi_\alpha(\tilde{a}(x))\xi_{-\alpha}(\tilde{b}(x)),
\]

(36)

where \( \xi_\alpha \) is as in (15).

**Lemma 5.1.** For any \( 0 \leq i, j \leq |\lambda| - 1 \) with \( i+j \leq |\lambda| - 1 \), the map \( \Upsilon_\alpha \) is of class
\( C^{i,j} \). Its partial derivatives, \( \Upsilon_\alpha^{(i,j)} := D_i D_j \Upsilon_\alpha : \tilde{G} \times \tilde{G} \to L(\tilde{G}^{i+j}; L^1(\mu)) \), are given by

\[
\Upsilon_\alpha^{(i,j)}(\tilde{a}, \tilde{b})(\tilde{u}_1, \ldots, \tilde{u}_i; \tilde{v}_1, \ldots, \tilde{v}_j)(x) = \xi^{(i)}_\alpha(\tilde{a}(x))\xi^{(j)}_{-\alpha}(\tilde{b}(x))\tilde{u}_1(x) \cdots \tilde{u}_i(x)\tilde{v}_1(x) \cdots \tilde{v}_j(x).
\]

(37)

**Proof.** Let \( 0 \leq i \leq |\lambda| - 2 \), \( 0 \leq j \leq |\lambda| - 1 \) and \( i+j \leq |\lambda| - 2 \). Let \( \tilde{a}, \tilde{b}, \tilde{u}_1, \ldots, \tilde{u}_i, \tilde{v}_1, \ldots, \tilde{v}_j \in \tilde{G} \) and let \( (\tilde{a}_n) \neq \tilde{a}, n \in \mathbb{N} \) be a sequence converging to \( \tilde{a} \) in \( \tilde{G} \). Applying the mean value theorem on an \( x \)-by-\( x \) basis, we obtain

\[
(\xi^{(i)}_\alpha(\tilde{a}_n) - \xi^{(i)}_\alpha(\tilde{a}))\xi^{(j)}_{-\alpha}(\tilde{b})(\tilde{u}_1 \cdots \tilde{u}_i \cdots \tilde{v}_j)(\tilde{u}_1(\tilde{a}_n) - \tilde{u}_1(\tilde{a})) + R_n,
\]

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

\[
S_n := (\xi^{(i+1)}_\alpha((1-\beta)\tilde{a} + \beta\tilde{a}_n) - \xi^{(i+1)}_\alpha(\tilde{a}))\xi^{(j)}_{-\alpha}(\tilde{b}).
\]

Now \( i+j+1 < |\lambda| \) and so, setting \( s = \lambda/(\lambda - i - j - 1) \), we have \( 1/s + (i+j+1)/\lambda = 1 \), and H"older’s inequality shows that

\[
E_{max}|R_n| \leq (E_{max}|S_n|^s)^{1/s}(E_{max}|\tilde{u}_1|^s)^{1/s} \cdots (E_{max}|\tilde{v}_j|^s)^{1/s}(E_{max}|\tilde{a}_n - \tilde{a}|^s)^{1/s}.
\]

(38)

so that

\[
\|\tilde{a}_n - \tilde{a}\|^{-1} \sup_{\|\tilde{u}_n\| = \|\tilde{v}_n\| = 1} \|R_n\|_{L^s(\mu)} \leq \|S_n\|_{L^s(\mu)}.
\]

Now \( S_n \to 0 \) in probability, and is dominated by \( f := 2\sup_{s} |\xi^{(i+1)}_\alpha(\tilde{a})||\xi^{(j)}_{-\alpha}(\tilde{b})| \). If \( j = 0 \) then \( f \in \tilde{G} \) and \( s \leq \lambda \), whereas if \( j \geq 1 \) 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 \( 0 \leq i \leq |\lambda| - 1, 0 \leq j \leq |\lambda| - 2 \), and \( i+j \leq |\lambda| - 2 \) 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 \( 0 \leq i, j \leq |\lambda| - 1 \) with \( i+j = |\lambda| - 1 \), \( \Upsilon_\alpha^{(i,j)} \) is
continuous. Now

\[
(\Upsilon_\alpha^{(i,j)}(\tilde{a}_n, \tilde{b}_n) - \Upsilon_\alpha^{(i,j)}(\tilde{a}, \tilde{b}))(\tilde{u}_1, \ldots, \tilde{v}_j) = (T_{i,n} + T_{2,n} + T_{3,n})\tilde{u}_1 \cdots \tilde{v}_j,
\]
For any $\alpha \in [-1, 1]$, and any $0 \leq i, j \leq |\lambda| - 1$ with $i + j \leq |\lambda| - 1$, the $\alpha$-divergence $D_\alpha : \tilde{M} \times \tilde{M} \to [0, \infty)$ is of class $C^{i,j}$.

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

$$D_\alpha(P | Q) := \begin{cases} E_{\mu} \left( \mathcal{L}_{-1}^{-1}(\tilde{b}) - \mathcal{L}_{-1}^{-1}(\tilde{a}) + \Upsilon_{-1}(\tilde{a}, \tilde{a}) - \Upsilon_{-1}(\tilde{b}, \tilde{b}) \right) & \text{if } \alpha = -1 \\ E_{\mu} \left( \frac{2}{1+\alpha} \mathcal{L}_{-1}^{-1}(\tilde{a}) + \frac{2}{1-\alpha} \mathcal{L}_{-1}^{-1}(\tilde{b}) - \Upsilon_{\alpha}(\tilde{a}, \tilde{b}) \right) & \text{if } \alpha \in (-1, 1) \\ E_{\mu} \left( \mathcal{L}_{-1}^{-1}(\tilde{a}) - \mathcal{L}_{-1}^{-1}(\tilde{b}) + T_1(\tilde{b}, \tilde{b}) - T_1(\tilde{a}, \tilde{b}) \right) & \text{if } \alpha = 1, \end{cases}$$

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.\qed

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

$$D_1 D_\alpha(\tilde{\phi}^{-1} | \tilde{\phi}^{-1})_{\tilde{a}, \tilde{b}} \tilde{u} = E_{\mu}(\Upsilon_\alpha^{(1,0)}(\tilde{a}, \tilde{u}) - \Upsilon_\alpha^{(1,0)}(\tilde{a}, \tilde{b})) \tilde{u}$$

$$D_2 D_\alpha(\tilde{\phi}^{-1} | \tilde{\phi}^{-1})_{\tilde{a}, \tilde{b}} \tilde{v} = E_{\mu}(\Upsilon_\alpha^{(0,1)}(\tilde{b}, \tilde{v}) - \Upsilon_\alpha^{(0,1)}(\tilde{a}, \tilde{b})) \tilde{v}.$$ (39)
Infinite-Dimensional Manifolds

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

$$UD_\alpha(\cdot|Q) = \langle F_{-\alpha}(P) - F_{-\alpha}(Q), UF_\alpha \rangle_{L^2(\mu)}$$
$$VD_\alpha(P|\cdot) = \langle F_\alpha(Q) - F_\alpha(P), VF_{-\alpha} \rangle_{L^2(\mu)},$$

(40)

and $D_\alpha$ admits the following mixed second derivative

$$D_1D_2D_\alpha(\tilde{\phi}^{-1}|\tilde{\phi}^{-1})_{\tilde{a},\tilde{b}}(\tilde{u}, \tilde{v}) = -E_\mu Y^{(1,1)}_\alpha(\tilde{a}, \tilde{b})(\tilde{u}, \tilde{v}) = -\langle UF_\alpha, VF_{-\alpha} \rangle_{L^2(\mu)},$$

(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 $T_p\hat{M}$: for any $U, V \in T_p\hat{M}$,

$$\langle U, V \rangle_p := -UVD_\alpha = \langle UF_\alpha, VF_{-\alpha} \rangle_{L^2(\mu)}.$$  

(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 $UF_\alpha$, $VF_{-\alpha}$ and $UVD_\alpha$. (See [18].)

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,

$$\langle U, U \rangle_p = E_\mu \frac{p}{(1 + p)^2} \tilde{u}^2 \leq E_\mu \tilde{u}^2 \leq \|\tilde{u}\|_{L_2}^2,$$

(43)

where $\tilde{u} = U\tilde{\phi}$; in particular $\langle U, U \rangle_p = 0$ if and only if $\tilde{u} = 0$. Thus $(T_p\hat{M}, \langle \cdot, \cdot \rangle_p)$ is an inner product space. As shown in (43), the Fisher norm is dominated by the natural Banach norm on $T_p\hat{M}$. However, it is not equivalent to that norm, even in the case $\lambda = 2$. (See [18].) In the general, infinite-dimensional case $(T_p\hat{M}, \langle \cdot, \cdot \rangle_p)$ is not a Hilbert space; the Fisher metric is a pseudo-Riemannian metric but not a Riemannian metric.

If $\lambda > 3$, $D_\alpha$ admits the following mixed third derivative

$$D_1^2D_2D_\alpha(\tilde{\phi}^{-1}|\tilde{\phi}^{-1})_{\tilde{a},\tilde{b}}(\tilde{u}, \tilde{v}; \tilde{w}) = -E_\mu Y^{(2,1)}_\alpha(\tilde{a}, \tilde{b})(\tilde{u}, \tilde{v}; \tilde{w}).$$

(44)

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

$$D_1^2D_2D_\alpha(\tilde{\phi}^{-1}|\tilde{\phi}^{-1})_{\tilde{a},\tilde{a}}(\tilde{u}, \tilde{v}; \tilde{w}) = -E_\mu \frac{p}{(1 + p)^2} \tilde{\Gamma}_\alpha(\tilde{u}, \tilde{v}; \tilde{w}),$$

(45)

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

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

(46)

If $\tilde{u} = \tilde{\phi}(P)$, and $\tilde{u}$ and $\tilde{v}$ are such that $\tilde{\Gamma}_\alpha(\tilde{u}, \tilde{v}, \tilde{w}) = 0$, then there exist tangent vectors $Y, W \in T_P\hat{M}$ such that $\Gamma_\alpha(\tilde{u}, \tilde{v}, \tilde{w}) = Y\tilde{\phi}$ and $\tilde{w} = W\tilde{\phi}$. In this case

$$D_1^2D_2D_\alpha(\tilde{\phi}^{-1}|\tilde{\phi}^{-1})_{\tilde{a},\tilde{a}}(\tilde{u}, \tilde{v}; \tilde{w}) = -\langle Y, W \rangle_p.$$  

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

\[
\mathring{\nabla}^\alpha_U V := \Phi^{-1} \left( \mathpmb{\phi}, U\mathpmb{v} + \mathpmb{r}_\alpha(\mathpmb{\phi}, \mathpmb{u}, \mathpmb{v}) \right).
\]  

However, this does not define an operator, \( \mathring{\nabla}^\alpha \), with domain \( \mathring{V}_1^0 \times \mathring{V}_1^1 \), and so it does not define a full covariant derivative on \( T\mathring{M} \). 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 \( \alpha \)-covariant derivatives. In the special case that \( s = \infty \), \( \mathring{\nabla}^\alpha_U V \) is well defined for all \( C^1 \) vector fields \( V \), and thus provides a limited notion of \( \alpha \)-parallel transport on the tangent bundle. (See [11] for a similar result on the exponential Orlicz manifold.)

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

\[
U(V, W)_P = \langle \mathring{\nabla}^\alpha_U V, W \rangle_P + \langle V, \mathring{\nabla}^\alpha_U W \rangle_P,
\]

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

### 5.1. The Fisher Metric and \( \alpha \)-Derivatives on \((M,G,\phi)\)

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

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

\[
\mathring{\nabla}^\alpha_U V = \Phi^{-1} \left( \mathpmb{\phi}, U(D\rho \mathpmb{v}) + \mathpmb{r}_\alpha(\mathpmb{\phi}, D\rho \mathpmb{u}, D\rho \mathpmb{v}) \right) = \Phi^{-1} \left( \mathpmb{\phi}, D\rho U \mathpmb{v} + D^2 \rho(\mathpmb{u}, \mathpmb{v}) + \mathpmb{r}_\alpha(\mathpmb{\phi}, D\rho \mathpmb{u}, D\rho \mathpmb{v}) \right) = \Phi^{-1} \left( \mathpmb{\phi}, D\rho U \mathpmb{v} + \frac{1 - \alpha}{2} \gamma - \frac{1 + \alpha}{2} \eta \right).
\]
where $\gamma, \eta : \tilde{M} \to \tilde{G}$ are defined by

$$
\gamma(P)(x) = \frac{D\rho_a u(P)(x)D\rho_a v(P)(x)}{(1 + p(x))^2} + D^2\rho_a(u(P), v(P))(x),
$$
(51)

$$
\eta(P)(x) = \frac{p(x)D\rho_a u(P)(x)D\rho_a v(P)(x)}{(1 + p(x))^2} - D^2\rho_a(u(P), v(P))(x),
$$
and $a = \phi(P)$. It follows from (25) and (27) that, for any $u \in G$,

$$
D\rho_a u = u - \frac{E_\mu \psi^{(1)}(\rho(a))u}{E_\mu \psi^{(1)}(\rho(a))} = u - \frac{E_\mu D\Xi_1 - 1, \rho(a) u}{E_\mu D\Xi_1 - 1, \rho(a) 1},
$$
and so, according to the quotient and chain rules of differentiation, and Lemma 3.1,

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

From these derivatives and (51) we conclude that

$$
\gamma(P) = D\rho_a \frac{\tilde{u}(P)\bar{v}(P)}{(1 + p)^2} - E_\mu \frac{\tilde{u}(P)\bar{v}(P)}{(1 + p)^2},
$$
$$
\eta(P) = D\rho_a \frac{\tilde{u}(P)\bar{v}(P)}{(1 + p)^2} - p(U(P), V(P)) P
$$
$$
+ \langle U(P), V(P) \rangle P \left( p - \frac{E_\mu \psi^{(1)}(\tilde{a}) p}{E_\mu \psi^{(1)}(\tilde{a})} \right)
$$
$$
+ \frac{1}{E_\mu \psi^{(1)}(\tilde{a})} E_\mu \psi^{(1)}(\tilde{a}) \left( p \tilde{u}(P)\bar{v}(P) (1 + p)^2 + \tilde{u}(P)\bar{v}(P) (1 + p)^2 \right)
$$
$$
= D\rho_a \frac{\tilde{u}(P)\bar{v}(P)}{(1 + p)^2} - p(U(P), V(P)) P
$$
$$
+ \langle U(P), V(P) \rangle P \left( p - \frac{1}{E_\mu \psi^{(1)}(\tilde{a})} + 1 \right) + \frac{1}{E_\mu \psi^{(1)}(\tilde{a})} E_\mu \psi^{(1)}(\tilde{a}) (1 + p)^2
$$
$$
= D\rho_a \frac{\tilde{u}(P)\bar{v}(P)}{(1 + p)^2} - p(U(P), V(P)) P + (1 + p)\langle U(P), V(P) \rangle P,
$$
where $\tilde{a} = \tilde{\phi}(P), \tilde{u}(P) = D\rho_a u(P), \tilde{v}(P) = D\rho_a v(P)$, and we have used the fact that $\psi^{(1)}(\psi) = \psi - \psi^{(1)}$ in the second step. We have thus shown that

$$
\tilde{\nabla}_U^n V(P) = \tilde{\phi}^{-1} \left( \tilde{a}, D\rho_a \left( U v(P) + \Gamma_\alpha(a, u(P), v(P)) \right) - \frac{1 + \alpha}{2} (1 + p)\langle U(P), V(P) \rangle P \right)
$$

where $\Gamma_\alpha : G \times G \times G \to E_0^{1/2}(\mu)$ is defined by

$$
\Gamma_\alpha(a, u, v)(x) = \frac{1 - \alpha}{2} \left( D\rho_a u(x)D\rho_a v(x) \right) - \frac{1 + \alpha}{2} p(x) \left( D\rho_a u(x)D\rho_a v(x) \right).
$$
(52)
For any \( W \in T_P M \)

\[
\langle \nabla^a_U V(P), W \rangle_P = E_\mu \frac{p}{(1 + p)^2} D_{\rho_\alpha} (Uv(P) + \Gamma_\alpha(a, u(P), v(P))) D_{\rho_\alpha} w
\]

\[
= \frac{1 + \alpha}{2} \langle U(P), V(P) \rangle_P E_\mu \frac{p}{(1 + p)^2} (1 + p) D_{\rho_\alpha} w
\]

(53)

where \( \nabla^a_U V : M \rightarrow TM \) is the vector field on \( M \) defined by

\[
\nabla^a_U V = \Phi^{-1} (\phi, Uv + \Gamma_\alpha(\phi, u, v)).
\]

(54)

As (53) shows, \( \nabla^a_U V(P) \) is the projection of \( \tilde{\nabla}^a_U V(P) \) onto \( TP M \), in the sense of the Fisher metric. The map \( \nabla^a_U : Y^1 / (s-1) \rightarrow Y^0 \), thus defined, is the \( \alpha \)-derivative on \( M \), which could also found by direct calculation in the same way as was \( \tilde{\nabla}^a_U \).

6. Finite-Dimensional Submanifolds

For some \( d \in \mathbb{N} \) and \( n \in \mathbb{N} \cup \{ \infty \} \), let \( (N, B, \theta) \) be a \( d \)-dimensional \( C^n \)-embedded submanifold of \( \tilde{M} \). By this, we mean that \( N \subset M \), \( B \) is a non-empty open subset of \( \mathbb{R}^d \), \( \theta : N \rightarrow B \) is a bijection, and the inclusion map \( i : N \rightarrow \tilde{M} \) is both a topological embedding and a \( C^n \)-immersion. (See, for example, [13].) As in section 2, the tangent space at base point \( P \in N \), \( TP N \), is spanned by the vectors \( (\partial_i, i = 1, \ldots, d) \), where \( \partial_i \) is the equivalence class of differentiable curves on \( N \) containing the curve \( (y_i(t) := \theta(P) + te_i, t \in (-\epsilon, \epsilon)) \). The matrix form of the \( (\text{extended}) \) Fisher metric is

\[
(g(P)_{i,j}) := (\partial_i, \partial_j)_P = E_\mu \frac{p}{(1 + p)^2} \tilde{w}_i \tilde{w}_j,
\]

(55)

where \( \tilde{w}_i = \partial_i \tilde{\phi} \).

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

\[
\Gamma^N_{\alpha}(P)_{i,j} := -g^{-1}(P)^{k,l} \partial_i \partial_j \partial_l \theta(g(P)^{k,l} \partial_\mu \frac{p}{(1 + p)^2} \tilde{\Gamma}_\alpha(\tilde{\phi}(P), \tilde{w}_i, \tilde{w}_j) \tilde{w}_l,
\]

(56)

where \( g^{-1} \) is the \( (k, l) \) element of the inverse of the matrix \( g(P) \), \( \partial_i \) and \( \partial_j \) act on the first argument of \( \partial_{\alpha} \), \( \partial_\mu \) acts on the second argument of \( \partial_{\alpha} \), 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, \( i : N \rightarrow M \), takes the form \( i = \pi \circ \tilde{i} \), where \( \pi = \phi^{-1} \circ \tilde{\rho} \circ \tilde{\phi} \) and \( \tilde{\rho} : \tilde{G} \rightarrow G \) is defined by \( \tilde{\rho}(\tilde{a}) = \tilde{a} - E_\mu \tilde{\phi} \). Clearly \( \pi \) is of class \( C^\infty \), and so \( i \) is of class \( C^n \). Furthermore, \( \partial_\mu \tilde{i} \in TP M \)
for all $i$, and the restriction of the pushforward $\pi_*$ to $T_P M$ is the identity map of $T_P M$, and so the derivative of $\iota$ is injective. Since $\iota$ is a topological embedding and the map $\rho$ of Proposition 4.1 is continuous, $\iota$ 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 $\gamma(y) = y^* \eta_i$, and let $N := \tilde{\phi}^{-1} \circ \gamma(\mathbb{R}^d)$. Since the $\eta_i$ are linearly independent $\gamma$ is an injection, and $(N, \mathbb{R}^d, \theta)$, with $\theta := \gamma^{-1} \circ \phi$, is a $d$-dimensional manifold. It is trivially a $C^\infty$-embedded submanifold of $\tilde{M}$.

**Example 6.2.** Let $(N, B, \theta)$ be the $d$-dimensional exponential statistical manifold defined in section 2, where the underlying space $(\mathcal{X}, \mathcal{X}, \mu)$ coincides with that of sections 3–5, and suppose that the $\eta_i$ and $B$ are such that $\theta^{-1}(B) \subseteq M$. It is shown in Theorem 5.1 of [18] that $N$, thus defined, is a $C^\infty$-embedded submanifold of $M$. (Strictly speaking, Theorem 5.1 in [18] 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 [22], comprising all probability measures in an absolute continuity equivalence class. The choice in [18] of the Hilbert space $L^2_0(\mu)$ 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 $\mu$ 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 $\mu$ is a probability measure. If $\mathcal{X} = \mathbb{R}^n$ then (M1) is satisfied by all measures that are mutually absolutely continuous with respect to Lebesgue measure if, for example, $\mu$ is a non-singular multivariate Gaussian measure. It may seem that one could construct larger manifolds by piecing together coordinate patches $(M_i, \tilde{G}_i, \tilde{\phi}_i, \mu_i)$, defined as in section 3 but with differing patch-centric measures, $\mu_i$. However, this is not possible since the requirement that $dP/d\mu_i \in L^\lambda(\mu_i)$ 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 $d\mu_i/d\mu_j \in L^\infty(\mu_j)$ for all $i, j$). The requirement that $dP/d\mu \in L^\lambda(\mu)$ 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 [22], is played here by the function \( \psi \). In this sense, \( M \) and \( M' \) are extreme examples of general deformed families, as defined in Chapter 10 of [17]. (They are extreme in the sense that \( \psi \) satisfies a linear growth condition.) General deformed families of probability measures are also developed and generalised in [25], where they are referred to as \( \varphi \)-families. The function \( \varphi \) is used there in the definition of the (Musielak-Orlicz) model spaces, and gives rise to dual divergence functions distinct from \( D_\alpha \). Here, our aim is somewhat different from those of [17] and [25]. 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 \( \varphi \)-function \( \psi \). (The Musielak-Orlicz spaces associated with the \( \varphi \)-function \( \psi \) 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**


