Information Geometry (2019) 2:283–312
https://doi.org/10.1007/s41884-019-00024-z

RESEARCH PAPER

A class of non-parametric statistical manifolds modelled on Sobolev space

Nigel J. Newton[1]

Received: 13 December 2018 / Revised: 31 October 2019 / Published online: 25 November 2019

Abstract
We construct a family of non-parametric (infinite-dimensional) manifolds of finite measures on $\mathbb{R}^d$, each containing a smoothly embedded submanifold of *probability* measures. The manifolds are modelled on a variety of weighted Sobolev spaces, including Hilbert–Sobolev spaces and mixed-norm spaces, and support the Fisher–Rao metric as a weak Riemannian metric. Densities are expressed in terms of a *deformed exponential* function having linear growth. Unusually for the Sobolev context, and as a consequence of its linear growth, this "lifts" to a nonlinear superposition (Nemytskii) operator that acts continuously on a particular class of mixed-norm model spaces, and on the fixed norm space $W^{2,1}$; i.e. it maps each of these spaces continuously into itself. In contrast with non-parametric exponential manifolds, the density itself belongs to the model space, and the range of the chart is the whole of this space. Some of the results make essential use of a log-Sobolev embedding theorem, which also sharpens existing results concerning the regularity of statistical divergences on the manifolds. Applications to the stochastic partial differential equations of nonlinear filtering (and hence to the Fokker–Planck equation) are outlined.

Keywords Banach manifold · Fisher–Rao metric · Fokker–Planck equation · Log-Sobolev Embedding · Non-parametric statistics · Sobolev space

Mathematics Subject Classification 46N30 · 60D05 · 60H15 · 62B10 · 93E11

1 Introduction

This paper constructs variants of the statistical manifolds of [25,27], for which the sample space $\mathbb{X}$ is $\mathbb{R}^d$. The model spaces used incorporate *spatial derivatives* of density functions, and thereby subsume both sample space and information topologies. Apart

✉ Nigel J. Newton
njn@essex.ac.uk

[1] University of Essex, Colchester CO4 3SQ, UK

from applications, such as the Fokker–Planck equation, this construction sharpens some of the results in [25,27] concerning the regularity of statistical divergences; this is consequence of log-Sobolev embedding.

Beginning with Rao's observation that the Fisher information can be interpreted as a Riemannian metric [32], *information geometry* has exploited the formalism of smooth manifold theory in problems of statistical estimation. The finite-dimensional (parametric) theory is now mature, and is treated pedagogically in [1,2,6,12,22]. The archetypal example is the finite-dimensional *exponential model*, which is based on a finite set of real-valued random variables defined on an underlying probability space $(\mathbb{X}, \mathcal{X}, \mu)$. Affine combinations of these are exponentiated to yield probability density functions with respect to the reference measure μ. This construction induces an "information" topology on the resulting set of probability measures, that is compatible with the statistical divergences of estimation theory, derivatives of which can be used to define the Fisher–Rao metric and other geometric objects.

The first fully successful extension of these ideas to the non-parametric setting appeared in [31], and was further developed in [10,16,30]. These papers follow the formalism of the exponential model by using the log of the density as a chart. This approach requires a model space with a strong topology: the exponential Orlicz space. Amari's α-divergences (for $\alpha \in [-1, 1]$) are of class C^∞ on the exponential Orlicz manifold. In [20], the exponential function is replaced by the so-called *q-deformed exponential*, which has an important interpretation in statistical mechanics. (See chapter 7 in [23].) The model space used is $L^\infty(\mu)$. A more general class of deformed exponentials is used in [35] to construct families of probability measures dubbed *φ-families*, in which the model spaces are Musielak–Orlicz spaces. A deformed exponential function having linear growth is used in [25,27] to construct statistical manifolds modelled on the Lebesgue $L^\lambda(\mu)$ spaces, including the Hilbert space $L^2(\mu)$.

Many of these references take the classical differential geometric approach of constructing the tangent space at each point, P, in a set of measures, and then building towards a global geometry. With this approach it is natural to use local charts with model spaces defined with respect to P [eg. the exponential Orlicz space $L^\Phi(P)$]. However, the global geometry contains no surprises—each connected component of these manifolds is covered by a single chart.

A notion of central importance is that of a *sufficient statistic*, and so divergences that are invariant in this sense are of particular interest. These include Csiszàr's f-divergences, of which Amari's α-divergences are special cases. Any geometric objects (such as a Riemannian metric), defined through derivatives of invariant divergences, clearly retain this property. The question naturally arises whether or not these objects are *uniquely* determined by this property. The uniqueness of the Fisher–Rao metric under sufficient statistics was established for parametric manifolds by Chentsov [11], and more recently for a large class of non-parametric manifolds in [4,5]. Rather than constructing specific statistical manifolds, the latter references consider a parametrised statistical model to be a generic Banach manifold on which is defined a continuous map into a set of finite measures, having suitable properties. This approach admits manifolds in which the measures do not necessarily have the same null sets [5].

One of the most important statistical divergences is the Kullback–Leibler (KL) divergence (which corresponds to $\alpha = -1$). For probability measures P and Q having densities p and q with respect to μ, this is defined as follows:

$$\mathcal{D}(P|Q) = \int p \log(p/q) d\mu. \tag{1}$$

The KL divergence can be given the bilinear representation $\langle p, \log p - \log q \rangle$, in which probability densities and their logs take values in dual function spaces [for example, the Lebesgue spaces $L^\lambda(\mu)$ and $L^{\lambda/(\lambda-1)}(\mu)$]. Loosely speaking, in order for $\mathcal{D}$ to be smooth, the charts of any non-parametric manifold must "control" both the density p and its log, and this provides one explanation of the need for strong topologies in the model spaces of non-parametric exponential models. (They have to control the density through the exponential function.) This observation led in [25,27] to the introduction of a "balanced chart" (the sum of the density and its log), that directly controls both, thereby enabling the use of model spaces with weaker topologies—the Lebesgue $L^\lambda(\mu)$ spaces, including the Hilbert case $\lambda = 2$. The Amari α-divergences then exhibit increasing degrees of smoothness with increasing λ.

None of the non-parametric manifolds discussed above makes reference to any topology that the underlying sample space $\mathbb{X}$ may possess. Statistical divergences measure dependency between abstract random variables (those taking values in measurable spaces). Nevertheless, topologies, metrics and linear structures on $\mathbb{X}$ play important roles in many applications. For example, the Fokker–Planck and Boltzmann equations both quantify the evolution of probability density functions on $\mathbb{R}^d$, making direct reference to the latter's topology through differential operators. A natural direction for research in non-parametric information geometry is to adapt the manifolds outlined above to such problems by incorporating the topology of the sample space in the model space, and one way of achieving this is to use model spaces of Sobolev type. This is carried out in the context of the exponential Orlicz manifold in [19], where it is applied to the spatially homogeneous Boltzmann equation. Manifolds modelled on the Banach spaces $C_b^k(B; \mathbb{R})$, where B is an open subset of an underlying (Banach) sample space, are developed in [29], and manifolds modelled on Fréchet spaces of smooth densities are developed in [7,9,29]. In this context, the Fisher–Rao metric is shown in [7] to be the unique metric that is invariant under the action of the diffeomorphism group.

The aim of this paper is to develop Sobolev variants of the Lebesgue $L^\lambda(\mu)$ manifolds of [25,27] when the sample space $\mathbb{X}$ is $\mathbb{R}^d$. We construct, as a special case, a class of Hilbert–Sobolev manifolds. In developing these, the author was motivated by applications in *nonlinear filtering*. The equations of nonlinear filtering for diffusion processes generalise the Fokker–Planck equation by adding a term that accounts for partial observations of the diffusion. Let $(X_t, Y_t, t \geq 0)$ be a $d + 1$-vector Markov diffusion process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and satisfying the Itô stochastic differential equation

$$d\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} f(X_t) \\ h(X_t) \end{bmatrix} dt + \begin{bmatrix} g(X_t) & 0 \\ 0 & 1 \end{bmatrix} dV_t, \tag{2}$$

where $Y_0 = 0$, $(V_t, t \geq 0)$ is a $d+1$-vector standard Brownian motion, independent of X_0, and $f : \mathbb{R}^d \to \mathbb{R}^d$, $g : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $h : \mathbb{R}^d \to \mathbb{R}$ are suitably regular functions. The nonlinear filter for X computes, at each time t, the conditional probability distribution of X_t given the history of the *observations process* $(Y_s, 0 \leq s \leq t)$. Since X and Y are jointly Markov the nonlinear filter can be expressed in a time-recursive manner. Under suitable technical conditions, the observation-conditional distribution of X_t admits a density, p_t, (with respect to Lebesgue measure) satisfying the *Kushner Stratonovich* stochastic partial differential equation [14]

$$dp_t = \mathcal{A}p_t\,dt + p_t(h - \bar{h}_t)d\big(Y_t - \bar{h}_t dt\big), \tag{3}$$

where $\mathcal{A}$ is the Kolmogorov forward (Fokker–Planck) operator for X, and $\bar{h}_t$ is the $(Y_s, 0 \leq s \leq t)$-conditional mean of $h(X_t)$.

The exponential Orlicz manifold was proposed as an ambient manifold for partial differential equations of this type in [8] (and the earlier references therein), and methods of projection onto submanifolds were developed. Applications of the Hilbert manifold of [25] to nonlinear filtering were developed in [26,28], and information-theoretic properties were investigated.

It was argued in [27,28] that statistical divergences such as the KL divergence are natural measures of error for approximations to Bayesian conditional distributions such as those of nonlinear filtering. This is particularly so when the approximation constructed is used to estimate a number statistics of the process X, or when the dynamics of (X, Y) are significantly nonlinear. We summarise these ideas here since they motivate the developments that follow; details can be found in [28]. If our purpose is to estimate a single real-valued variate $v(X_t) \in L^2(\mu)$, then the estimate with the minimum mean-square error is the conditional mean $\bar{v}_t := \mathbf{E}_{\Pi_t} v = \mathbb{E}(v(X_t)|(Y_s, 0 \leq s \leq t))$, where $\mathbb{E}$ is expectation with respect to $\mathbb{P}$, and Π_t is the conditional distribution of X_t. If the estimate is based on a $(Y_s, 0 \leq s \leq t)$-measurable approximation to Π_t, $\hat{\Pi}_t$, then the mean-square error admits the orthogonal decomposition

$$\mathbb{E}\left(v(X_t) - \mathbf{E}_{\hat{\Pi}_t} v\right)^2 = \mathbb{E}\mathbf{E}_{\Pi_t}(v - \bar{v}_t)^2 + \mathbb{E}\Big(\bar{v}_t - \mathbf{E}_{\hat{\Pi}_t} v\Big)^2. \tag{4}$$

The first term on the right-hand side here is the *statistical error*, and is associated with the limitations of the observation Y; the second term is the *approximation error* resulting from the use of $\hat{\Pi}_t$ instead of Π_t. When comparing different approximations, it is appropriate to measure the second term *relative to the first*; if $\bar{v}_t$ is a poor estimate of $v(X_t)$ then there is no point in approximating it with great accuracy. Maximising these *relative* errors over all square-integrable variates leads to the (extreme) multi-objective measure of mean-square approximation errors $\mathcal{D}_{MO}(\hat{\Pi}_t|\Pi_t)$, where

$$\mathcal{D}_{MO}(Q|P) := \frac{1}{2}\sup_{v\in L^2(P)} \frac{\big(\mathbf{E}_Q v - \mathbf{E}_P v\big)^2}{\mathbf{E}_P(v - \mathbf{E}_P v)^2} = \frac{1}{2}\|dQ/dP - 1\|^2_{L^2(P)}. \tag{5}$$

$\mathcal{D}_{MO}$ is Pearson's χ^2-divergence. Although extreme, it illustrates an important feature of multi-objective measures of error—they require probabilities of events that are small to be approximated with greater absolute accuracy than those that are large. A less extreme multi-objective measure of mean-square errors is developed in [28]. This constrains the functions v of (5) to have exponential moments. The resulting measure of errors is shown to be of class C^1 on the Hilbert manifold of [25], and so has this same property on the manifolds developed here. See [28] for further discussion of these ideas.

The paper is structured as follows: Sect. 2 provides the technical background in mixed-norm weighted Sobolev spaces, where the L^λ spaces are based on a probability measure. Section 3 constructs (M, G, ϕ), a manifold of finite measures modelled on the general Sobolev space of Sect. 2. It outlines the properties of mixture and exponential representations of measures on the manifold, as well as those of the KL divergence. In doing so, it defines the Fisher–Rao metric and Amari–Chentsov tensor. Section 4.1 shows that a particular choice of mixed-norm Sobolev space is especially suited to the manifold in the sense that the density (and log-density) of any $P \in M$ also belong to the model space, and the associated nonlinear superposition operator is continuous—a rare property in the Sobolev context [34]. Section 4.2 shows that this property does not hold for fixed norm spaces, except in the special case $G = W^{2,1}$. It also develops a general class of fixed norm spaces, for which the continuity property can be retained if the Lebesgue exponent in the range space is suitably reduced. Section 5 develops an embedded submanifold of *probability* measures (M_0, G_0, ϕ_0), in which the charts are *centred* versions of ϕ. Section 6 outlines applications to the problem of nonlinear filtering for a diffusion process. Section 7 makes some concluding remarks, discussing, in particular, a variant of the results that uses the Kaniadakis deformed logarithm as a chart.

2 The model spaces

For some $t \in (0, 2]$, let $\theta_t : [0, \infty) \to [0, \infty)$ be a strictly increasing function that is twice continuously differentiable on $(0, \infty)$, such that $\lim_{z \downarrow 0} \theta_t'(z) < \infty$, and

$$\theta_t(z) = \left\{ \begin{matrix} 0 & \text{if } z = 0 \\ c_t + z^t & \text{if } z \geq z_t \end{matrix} \right\}, \quad \text{where } z_t \geq 0, \text{ and } c_t \in \mathbb{R}. \tag{6}$$

If $t \in (1, 2]$ then we also require θ_t and $-\sqrt{\theta_t}$ to be convex.

Example 1 (i) Simple: $t \in [1, 2]$ and $z_t = c_t = 0$.
(ii) Smooth: $t \in (0, 2]$, $z_t = 2 - t$, $c_t = \alpha_t(1 - \cos(\beta_t z_t)) - z_t^t$, and

$$\theta_t(z) = \alpha_t(1 - \cos(\beta_t z)), \quad \text{for } z \in [0, z_t]. \tag{7}$$

Here, $\beta_t z_t$ is the unique solution in the interval $(0, \pi)$ of the equation

$$(t - 1)\tan(\beta_t z_t) = \beta_t z_t, \tag{8}$$

and $\alpha_t \beta_t \sin(\beta_t z_t) = t z_t^{t-1}$. (If $t = 1$ then $\beta_t = \beta_t z_t = \pi/2$.) The compound function $\mathbb{R} \ni z \mapsto \theta_t(|z|) \in \mathbb{R}$ is then of class C^2.

For some $d \in \mathbb{N}$ $(= \{1, 2, \ldots\})$, let $\mathcal{X}$ be the σ-algebra of Lebesgue measurable subsets of $\mathbb{R}^d$; let μ_t be the following probability measure on $\mathcal{X}$:

$$\mu_t(dx) = r_t(x)\,dx := \exp(l_t(x))dx, \quad \text{where } l_t(x) := \textstyle\sum_i (C_t - \theta_t(|x_i|)), \tag{9}$$

and $C_t \in \mathbb{R}$ is such that $\int \exp(C_t - \theta_t(|z|))dz = 1$. In what follows, we shall suppress the parameter t, and so l_t, r_t and μ_t will become l, r and μ, etc. μ is *mutually* absolutely continuous with respect to Lebesgue measure.

For any $1 \leq \lambda < \infty$, let $L^\lambda(\mu)$ be the Banach space of (equivalence classes of) measurable functions $u : \mathbb{R}^d \to \mathbb{R}$ for which $\|u\|_{L^\lambda(\mu)} := (\int |u|^\lambda d\mu)^{1/\lambda} < \infty$. Let $C^\infty(\mathbb{R}^d; \mathbb{R})$ be the space of continuous functions with continuous partial derivatives of all orders, and let $C_0^\infty(\mathbb{R}^d; \mathbb{R})$ be the subspace of those functions having compact support.

For $k \in \mathbb{N}$, let $S := \{0, \ldots, k\}^d$ be the set of d-tuples of integers in the range $0 \leq s_i \leq k$. For $s \in S$, we define $|s| = \sum_i s_i$, and denote by 0 the d-tuple for which $|s| = 0$. For any $0 \leq j \leq k$, $S_j := \{s \in S : j \leq |s| \leq k\}$ is the set of d-tuples of weight at least j and at most k. Let $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_k)$, where $1 \leq \lambda_k \leq \lambda_{k-1} \leq \cdots \leq \lambda_0 < \infty$, and let $W^{k,\Lambda}(\mu)$ be the mixed-norm, weighted Sobolev space comprising functions $a \in L^{\lambda_0}(\mu)$ that have weak partial derivatives $D^s a \in L^{\lambda_{|s|}}(\mu)$, for all $s \in S_1$. For $a \in W^{k,\Lambda}(\mu)$ we define

$$\|a\|_{W^{k,\Lambda}(\mu)} := \left(\sum_{s \in S_0} \|D^s a\|^{\lambda_0}_{L^{\lambda_{|s|}}(\mu)} \right)^{1/\lambda_0} < \infty. \tag{10}$$

The following theorem is a variant of a standard result in the theory of fixed-norm, unweighted Sobolev spaces.

Theorem 1 *The space* $(W^{k,\Lambda}(\mu), \|\cdot\|_{W^{k,\Lambda}(\mu)})$ *is a Banach space.*

Proof That $\|\cdot\|_{W^{k,\Lambda}(\mu)}$ satisfies the axioms of a norm is easily verified. Suppose that $(a_n \in W^{k,\Lambda}(\mu))$ is a Cauchy sequence in this norm; then, since the spaces $L^{\lambda_j}(\mu)$, $0 \leq j \leq k$ are all complete, there exist functions $v_s \in L^{\lambda_{|s|}}(\mu)$, $s \in S_0$ such that $D^s a_n \to v_s$ in $L^{\lambda_{|s|}}(\mu)$. For any $s \in S_0$, and any $\varphi \in C_0^\infty(\mathbb{R}^d; \mathbb{R})$,

$$\begin{aligned} \left| \int (D^s a_n - v_s)\varphi\,dx \right| &\leq \int |D^s a_n - v_s||\varphi|\,dx \\ &= \int |D^s a_n - v_s||\varphi| r^{-1} \mu(dx) \\ &\leq \sup_{x \in \text{supp}(\varphi)} (|\varphi|/r)(x) \|D^s a_n - v_s\|_{L^1(\mu)} \to 0, \end{aligned} \tag{11}$$

and so

$$\int v_s \varphi\,dx = \lim_n \int D^s a_n \varphi\,dx = (-1)^{|s|} \lim_n \int a_n D^s \varphi\,dx = (-1)^{|s|} \int v_0 D^s \varphi\,dx,$$

v_0 admits weak derivatives up to order k, and $D^s v_0 = v_s$. So $W^{k,\Lambda}(\mu)$ is complete. □

The following developments show that functions in $W^{k,\Lambda}(\mu)$ can be approximated by particular functions in $C^\infty(\mathbb{R}^d;\mathbb{R})$ or $C_0^\infty(\mathbb{R}^d;\mathbb{R})$. For any $z \in (0,\infty)$, let $B_z := \{x \in \mathbb{R}^d : |x| \le z\}$. Let $J \in C_0^\infty(\mathbb{R}^d;[0,\infty))$ be a function having the following properties: (i) $\mathrm{supp}(J) = B_1$; (ii) $\int J\,dx = 1$. For any $0 < \epsilon < 1$, let $J_\epsilon(x) := \epsilon^{-d} J(x/\epsilon)$; then J_ϵ also has unit integral, but is supported on B_ϵ. Since l is bounded on bounded sets, any $u \in L^1(\mu)$ is also in $L^1_{\mathrm{loc}}(dx)$, and we can define the *mollified* version $J_\epsilon * u \in C^\infty(\mathbb{R}^d;\mathbb{R})$ as follows:

$$(J_\epsilon * u)(x) := \int J_\epsilon(x-y)u(y)\,dy. \tag{12}$$

For any $m \in \mathbb{N}$, let $\mathcal{U}_m \subset L^1(\mu)$ comprise those functions that take the value zero on the complement of B_m. If $u \in \mathcal{U}_m$ then $J_\epsilon * u \in C_0^\infty(B_{m+1};\mathbb{R})$.

Lemma 1 (i) *For any* $\lambda \in [1,\infty)$, $m \in \mathbb{N}$ *and any* $u \in \mathcal{U}_m \cap L^\lambda(\mu)$, *there exists an* $0 < \epsilon < 1$ *such that*

$$\|J_\epsilon * u - u\|_{L^\lambda(\mu)} < 1/m. \tag{13}$$

(ii) *For any* $a \in W^{k,\Lambda}(\mu)$, $0 < \epsilon < 1$ *and* $s \in S_1$, $D^s(J_\epsilon * a) = J_\epsilon * (D^s a)$.

Proof It follows from Jensen's inequality that, for any $\lambda \in [1,\infty)$,

$$|(J_\epsilon * u)(x)|^\lambda \le (J_\epsilon * |u|^\lambda)(x) = \int J_\epsilon(x-y)|u(y)|^\lambda r(y)^{-1}\mu(dy).$$

Since l is uniformly continuous on B_{2m+1}, there exists a $0 < \alpha_m < 1$ such that $|l(x) - l(y)| \le \lambda \log 2$ for all $y \in B_{2m}$, $|x-y| \le \alpha_m$. So, for any $0 < \epsilon < \alpha_m$,

$$\begin{aligned}\|J_\epsilon * u\|^\lambda_{L^\lambda(\mu)} &\le \int\int J_\epsilon(x-y)|u(y)|^\lambda \exp(l(x)-l(y))\mu(dy)dx\\ &\le 2^\lambda \int\int J_\epsilon(x-y)dx|u(y)|^\lambda\mu(dy) = 2^\lambda\|u\|^\lambda_{L^\lambda(\mu)}.\end{aligned} \tag{14}$$

It is a standard result that there exists a $\varphi \in C_0^\infty(B_{2m};\mathbb{R})$ such that $\|u - \varphi\|_{L^\lambda(\mu)} < 1/6m$, which together with (14) shows that, for any $0 < \epsilon < \alpha_m$, $\|J_\epsilon * u - J_\epsilon * \varphi\|_{L^\lambda(\mu)} < 1/3m$. Furthermore,

$$|(J_\epsilon * \varphi)(x) - \varphi(x)| \le \int J_\epsilon(x-y)|\varphi(y) - \varphi(x)|\,dy \le \sup_{|x-y|\le\epsilon}|\varphi(y)-\varphi(x)|.$$

Since φ is uniformly continuous, there exists a $0 < \beta_u < 1$ such that, for any $0 < \epsilon < \beta_u$ and all x, $|(J_\epsilon * \varphi)(x) - \varphi(x)| < 1/3m$. We can now choose $0 < \epsilon < \min\{\alpha_m, \beta_u\}$, which completes the proof of part (i).

For a and s as in part (ii), and any $\varphi \in C_0^\infty(\mathbb{R}^d;\mathbb{R})$,

$$\begin{aligned}\int (J_\epsilon * a)(x) D^s\varphi(x)\,dx &= \int\int J_\epsilon(y)a(x-y)\,dy D^s\varphi(x)\,dx \\ &= \int\int a(x-y)D^s\varphi(x)\,dx J_\epsilon(y)\,dy \\ &= (-1)^{|s|}\int\int D^s a(x-y)\varphi(x)\,dx J_\epsilon(y)\,dy \\ &= (-1)^{|s|}\int (J_\epsilon * D^s a)(x)\varphi(x)\,dx,\end{aligned}$$

where we have used integration by parts $|s|$ times in the third step. This completes the proof of part (ii). □

For ease of notation in what follows, we shall abbreviate $J_\epsilon * u$ to Ju, where it is understood that ϵ has been chosen as in part (i). This can clearly be achieved *uniformly* across any finite subset of $\mathcal{U}_m$, including sets of functions together with their weak derivatives up to order k.

For any $a \in W^{k,\Lambda}(\mu)$ and $m \in \mathbb{N}$, let $a_m(x) := a(x)\rho(x/m)$, where $\rho \in C_0^\infty(\mathbb{R}^d;\mathbb{R})$ is such that

$$\rho(x) = 1 \text{ if } |x| \le 1/2 \quad \text{and} \quad \rho(x) = 0 \text{ if } |x| \ge 1. \tag{15}$$

Lemma 2 *$Ja_m \to a$ in $W^{k,\Lambda}(\mu)$, and so $C_0^\infty(\mathbb{R}^d;\mathbb{R})$ is dense in $W^{k,\Lambda}(\mu)$.*

Proof Since S_0 is finite we may choose $\epsilon > 0$ such that (13) is satisfied for all $u = D^s a_m$ with $s \in S_0$ and $\lambda = \lambda_{|s|}$. According to the Leibniz rule,

$$D^s a_m = \sum_{\sigma \le s} m^{-|s-\sigma|} D^\sigma a D^{s-\sigma}\rho \prod_{1\le i\le d}\binom{s_i}{\sigma_i}, \tag{16}$$

and so $|D^s a_m| \le K\sum_\sigma |D^\sigma a| \in L^{\lambda_{|s|}}(\mu)$. Since $D^s a_m \to D^s a$ for all x, it follows from the dominated convergence theorem that it also converges in $L^{\lambda_{|s|}}(\mu)$. Lemma 1 completes the proof. □

Remark 1 If $\lambda_j = 2$ for $0 \le j \le k$ then $H^k(\mu) := W^{k,\Lambda}(\mu)$ is a Hilbert Sobolev space with inner product

$$\langle a, b\rangle_{H^k(\mu)} = \sum_{s\in S_0}\left\langle D^s a, D^s b\right\rangle_{L^2(\mu)}. \tag{17}$$

3 The manifolds of finite measures

In this section, we construct manifolds of finite measures on $(\mathbb{R}^d, \mathcal{X})$ modelled on the Sobolev spaces of Sect. 2. The charts of the manifolds are based on the "deformed logarithm" $\log_d : (0, \infty) \to \mathbb{R}$, defined by

$$\log_d y = y - 1 + \log y. \tag{18}$$

Now $\inf_y \log_d y = -\infty$, $\sup_y \log_d y = +\infty$, and $\log_d \in C^\infty((0,\infty);\mathbb{R})$ with strictly positive first derivative $1 + y^{-1}$, and so, according to the inverse function theorem, $\log_d$ is a diffeomorphism from $(0,\infty)$ onto $\mathbb{R}$. Let ψ be its inverse. This can be thought of as a "deformed exponential" function [23]. We use $\psi^{(n)}$ to denote its nth derivative and, for convenience, set $\psi^{(0)} := \psi$.

Lemma 3 (i) *For any $n \in \mathbb{N}$:*

$$(1+\psi)\psi^{(n)} = \psi^{(n-1)} - \frac{1}{2}\sum_{j=1}^{n-1}\binom{n}{j}\psi^{(j)}\psi^{(n-j)}; \tag{19}$$

in particular $\psi^{(1)} = \psi/(1+\psi) > 0$ and $\psi^{(2)} = \psi/(1+\psi)^3 > 0$, and so ψ is strictly increasing and convex.

(ii) *For any $n \geq 2$,*

$$\psi^{(n)} = \frac{Q_{n-2}(\psi)}{(1+\psi)^{2(n-1)}}\psi^{(1)}, \tag{20}$$

where Q_{n-2} is a polynomial of degree no more than $n-2$. In particular, $\psi^{(n)}$, $\psi^{(n)}/\psi$ and $\psi^{(n)}/\psi^{(1)}$ are all bounded.

Proof That $\psi^{(1)}$ and $\psi^{(2)}$ are as stated is verified by a straightforward computation. Both (19) and (20) then follow by induction arguments. □

Let $G := W^{k,\Lambda}(\mu)$ be the general mixed-norm space of Sect. 2, and let M be the set of finite measures on $(\mathbb{R}^d, \mathcal{X})$ satisfying the following:

(M1) P is mutually absolutely continuous with respect to Lebesgue measure;
(M2) $\log_d p = \log_d(dP/d\mu) \in G$;

(We denote measures in M by the upper-case letters $P, Q, \ldots$, and their densities with respect to μ by the corresponding lower case letters, $p, q, \ldots$). In order to control both the density p and its log, we employ the "balanced" chart of [25,27], $\phi : M \to G$. This is defined by:

$$\phi(P) = \log_d p = p - 1 + \log p. \tag{21}$$

Proposition 1 *ϕ is a bijection onto G.*

Proof It follows from (M2) that, for any $P \in M$, $\phi(P) \in G$. Suppose, conversely, that $a \in G$; then since $\psi^{(1)}$ is bounded, $\psi(a) \in L^1(\mu)$, and so defines a finite measure $P(dx) = \psi(a(x))\mu(dx)$. Since ψ is strictly positive, P satisfies (M1); that it also satisfies (M2) follows from the fact that $\log_d \psi(a) = a \in G$. We have thus shown that $P \in M$ and clearly $\phi(P) = a$. □

The inverse map $\phi^{-1} : G \to M$ takes the form

$$P(dx) = \phi^{-1}(a) = \psi(a(x))\mu(dx). \tag{22}$$

In [25,27], tangent vectors were defined as equivalence classes of differentiable curves passing through a given base point, and having the same first derivative at this point. Here, we use a different (but equivalent) definition, which is closer to that of membership of M. For any $P \in M$, let $\tilde{P}_a(dx) := \psi^{(1)}(a(x))\mu(dx)$, where $a = \phi(P)$. ($\tilde{P}_a$ is *mutually* absolutely continuous with respect to μ since $\psi^{(1)}$ is strictly positive.) We define a tangent vector U at P to be a *signed* measure on $(\mathbb{R}^d, \mathcal{X})$ that is absolutely continuous with respect to $\tilde{P}_a$, with Radon–Nikodym derivative $dU/d\tilde{P}_a \in G$. (This definition is sound since, for every $u \in G$, $U(dx) := u(x)\tilde{P}_a(dx)$ defines such a measure.) The tangent space, $T_P M$, is the linear space comprising all such measures, and the tangent bundle is the disjoint union $TM := \cup_{P \in M}(P, T_P M)$. This is globally trivialised by the chart $\Phi : TM \to G \times G$, where

$$\Phi(P, U) = \Big(\phi(P), dU/d\tilde{P}_a\Big). \tag{23}$$

The derivative of a (Fréchet) differentiable, Banach-space-valued map $f : M \to \mathbb{Y}$ (at P and in the "direction" U) is defined in the obvious way:

$$Uf = \Big(f \circ \phi^{-1}\Big)_a^{(1)} u, \quad \text{where } (a, u) = \Phi(P, U). \tag{24}$$

Clearly $u = U\phi$. We shall also need a weaker notion of differentiability due to Leslie [17,18]. Let $A : G \to Y$ be a continuous linear map and, for fixed $a = \phi(P) \in G$, let $R : \mathbb{R} \times G \to Y$ be defined by

$$R(y, u) = \begin{cases} y^{-1}\big(f \circ \phi^{-1}(a + yu) - f \circ \phi^{-1}(a)\big) - Au & \text{if } y \neq 0, \\ 0 & \text{if } y = 0. \end{cases}$$

If R is continuous at $(0, u)$ for all $u \in G$, then we say that f is *Leslie differentiable* at P, with derivative

$$Uf = d\Big(f \circ \phi^{-1}\Big)_a u = Au. \tag{25}$$

If f is Leslie differentiable at all $P \in M$ then we say that it is Leslie differentiable. This is a slightly stronger property than the "d-differentiability" used in [25,27], which essentially demands continuity of R in the first argument only. Leslie differentiability lies between Fréchet differentiability and Gateaux differentiability. (The latter does not require the existence of the continuous linear map A.)

The construction above defines an infinite-dimensional smooth manifold of finite measures, (M, G, ϕ), with atlas comprising the single chart, ϕ. M is a subset of an instance of the manifold constructed in [27] (that in which the measurable space $\mathbb{X}$ of [27] is $\mathbb{R}^d$), but has a stronger topology than the relative topology. Results in [27] concerning the smoothness of maps defined on the model space $L^{\lambda_0}(\mu)$ are true *a-forteriori* when $L^{\lambda_0}(\mu)$ is replaced by G; in fact, even stronger results can be obtained under the following hypothesis:

(E1) $t \in (1, 2]$ and $\lambda_1 = \lambda_0$.

For some $1 \le \beta \le \lambda_0$, let $\Psi_\beta : G \to L^\beta(\mu)$ be the nonlinear superposition (Nemytskii) operator defined by $\Psi_\beta(a)(x) = \psi(a(x))$ (see [3]).

Lemma 4 (i) $\Psi_\beta \in C^N(G; L^\beta(\mu))$, *where*

$$N = N(\lambda_0, \lambda_1, \beta, t) := \begin{cases} \lceil \lambda_0/\beta \rceil - 1 & \textit{if (E1) does not hold}, \\ \lfloor \lambda_0/\beta \rfloor & \textit{if (E1) holds}. \end{cases} \tag{26}$$

For any $1 \le j \le N$, $\Psi_\beta^{(j)} : G \to L(G^j;\ L^\beta(\mu))$ *is as follows*

$$\Psi_{\beta,a}^{(j)}(u_1, \ldots, u_j)(x) = \psi^{(j)}(a(x))u_1(x) \cdots u_j(x). \tag{27}$$

(ii) *If* $\lambda_0/\beta \in \mathbb{N}$ *and (E1) does not hold, then the highest Fréchet derivative,* $\Psi_\beta^{(N)}$*, is Leslie differentiable, with derivative*

$$\left(d\Psi_{\beta,a}^{(N)} u_{N+1}\right)(u_1, \ldots, u_N)(x) = \psi^{(N+1)}(a(x))u_1(x) \cdots u_{N+1}(x). \tag{28}$$

(iii) Ψ_β *satisfies global Lipschitz continuity and linear growth conditions, and all its derivatives are globally bounded.*

Proof According to the mean value theorem, for any $a, b \in G$,

$$\psi(b) - \psi(a) = \psi^{(1)}(\alpha b + (1-\alpha)a)(b - a) \quad \text{for some } 0 \le \alpha(x) \le 1, \tag{29}$$

and so the Lipschitz continuity and linear growth of Ψ_β follow from the boundedness of $\psi^{(1)}$. Let $(a_n \in G\backslash\{a\})$ be a sequence converging to a in G. For any $1 \le j \le N$ let

$$\begin{aligned} \Delta_n &:= \psi^{(j-1)}(a_n) - \psi^{(j-1)}(a) - \psi^{(j)}(a)(a_n - a) \\ \Gamma_n &:= \psi^{(j)}(a_n) - \psi^{(j)}(a). \end{aligned} \tag{30}$$

According to the mean-value theorem $\Delta_n = \delta_n(a_n - a)$, where

$$\delta_n = \psi^{(j)}(\alpha_n a_n + (1-\alpha_n)a) - \psi^{(j)}(a) \quad \text{for some } 0 \le \alpha_n(x) \le 1.$$

Hölder's inequality shows that, for all $u_1, \ldots, u_j$ in the unit ball of G,

$$\|\Delta_n u_1 \cdots u_{j-1}\|_{L^\beta(\mu)} \le \|\Delta_n\|_{L^\nu(\mu)} \quad \text{and} \quad \|\Gamma_n u_1 \cdots u_j\|_{L^\beta(\mu)} \le \|\Gamma_n u_j\|_{L^\nu(\mu)},$$

where $\nu := \lambda_0\beta/(\lambda_0 - (N-1)\beta)$. In order to prove part (i), it thus suffices to show that

$$\|a_n - a\|_G^{-1}\|\varDelta_n\|_{L^\nu(\mu)} \to 0 \quad \text{and} \quad \sup_{\|u\|_G=1}\|\varGamma_n u\|_{L^\nu(\mu)} \to 0. \tag{31}$$

If $\nu < \lambda_0$ [eg. if (E1) does not hold] then Hölder's inequality shows that

$$\|\varDelta_n\|_{L^\nu(\mu)} \le \|\delta_n\|_{L^\zeta(\mu)}\|a_n - a\|_{L^{\lambda_0}(\mu)} \quad \text{and} \quad \|\varGamma_n u\|_{L^\nu(\mu)} \le \|\varGamma_n\|_{L^\zeta(\mu)}\|u\|_{L^{\lambda_0}(\mu)},$$

where $\zeta := \lambda_0\nu/(\lambda_0 - \nu)$. Now δ_n and $\varGamma_n$ are bounded and converge to zero in probability, and so the bounded convergence theorem establishes (31).

If $\nu = \lambda_0$ then (E1) holds. Suppose first that $\nu > 1$, and let $f_m \in C^\infty(\mathbb{R}^d;\mathbb{R})$ be a sequence converging in G to some $b \in G$. For some $1 \le i \le d$ and a weakly differentiable $g : \mathbb{R}^d \to \mathbb{R}$, let $g' := \partial g/\partial x_i$; then $(|f_m|^\nu)' = h(f_m)f_m'$ where $h(y) = \nu|y|^{\nu-1}\operatorname{sgn}(y)$. For any $0 < C < \infty$,

$$\mathbf{1}_{\{|h(f_m)f_m'|>C\}} \le \mathbf{1}_{\{|h(f_m)|^{\nu^*}>C\}} + \mathbf{1}_{\{|f_m'|^\nu>C\}},$$

where $\nu^* := \nu/(\nu-1)$. Together with Hölder's inequality, the uniform integrability of the sequences $|h(f_m)|^{\nu^*}$ and $|f_m'|^\nu$, and the continuity of h, this shows that $h(f_m)f_m' \to h(b)b'$ in $L^1(\mu)$. As in the proof of Theorem 1, this shows that $|b|^\nu$ is weakly differentiable with respect to x_i, with derivative

$$(|b|^\nu)' = h(b)b' \in L^1(\mu). \tag{32}$$

This enables the use of a log-Sobolev inequality. Let $\alpha := (t-1)/t$, and let $F_\alpha, G_\alpha : [0,\infty) \to [0,\infty)$ be the complementary *Young functions* defined by

$$F_\alpha(z) = \int_0^z \log^\alpha(y+1)\,dy \quad \text{and} \quad G_\alpha(z) = \int_0^z \Big(\exp(y^{1/\alpha}) - 1\Big)\,dy. \tag{33}$$

(see, for example, [33]). F_α is equivalent to any Young function $\tilde{F}_\alpha$, for which $\tilde{F}_\alpha(z) = z\log^\alpha z$ for $z \ge 3$, in the sense that there exist constants $0 < c_1 < c_2 < \infty$ such that, for all sufficiently large z, $F_\alpha(c_1 z) \le \tilde{F}_\alpha(z) \le F_\alpha(c_2 z)$. Similarly, G_α is equivalent to any Young function $\tilde{G}_\alpha$, for which $\tilde{G}_\alpha(z) = \exp(z^{1/\alpha})$ for $z \ge 1$. We denote the associated Orlicz spaces $L^1\log^\alpha L(\mu)$ and $\exp L^{1/\alpha}(\mu)$, respectively. $L^1\log^\alpha L(\mu)$ is equal (modulo equivalent norms) to the Lorentz–Zygmund space $L^{1,1;\alpha}(\mu)$, which in the context of the product probability space $(\mathbb{R}^d, \mathcal{X}, \mu)$ is a *rearrangement-invariant space* (see section 3 in [13]). It follows from Theorem 7.12 in [13], together with (32), that

$$\big\||b|^\nu\big\|_{L^1\log^\alpha L(\mu)} \le K\|b\|_G^\nu, \quad \text{for some } K < \infty.$$

This is clearly also true if $\nu = 1$. In the light of the generalised Hölder inequality, in order to prove (31) it now suffices to show that the sequences $|\delta_n|^\nu$ and $|\varGamma_n|^\nu$ converge to zero in $\exp L^{1/\alpha}(\mu)$, but this follows from their boundedness and convergence to zero in probability. This completes the proof of part (i).

With the hypotheses of part (ii), let $(t_n \in \mathbb{R}\backslash\{0\})$ and $(v_n \in G)$ be sequences converging to 0 and u_{N+1}, respectively, and let $a_n := a + t_n v_n$. Substituting this sequence into (30), we obtain

$$t_n^{-1}\Delta_n = \delta_n v_n = \delta_n(v_n - u_{N+1}) + \delta_n u_{N+1}.$$

Both terms on the right-hand side here converge to zero in $L^{\lambda_0}(\mu)$ since δ_n is bounded and converges to zero in probability. This completes the proof of part (ii). Part (iii) follows from (29) and the boundedness of the $\psi^{(j)}$. □

For $1 \leq \beta \leq \lambda_0$, let $m_\beta, e_\beta : M \to L^\beta(\mu)$ be defined by

$$m_\beta(P) = \Psi_\beta(\phi(P)) - 1 \quad \text{and} \quad e_\beta(P) = \imath \circ \phi(P) - m_\beta(P), \tag{34}$$

where $\imath : G \to L^\beta(\mu)$ is the inclusion map. These are injective and share the smoothness properties of Ψ_β developed in Lemma 4. In particular,

$$Um_\beta = \psi^{(1)}(a)\frac{dU}{d\tilde{P}_a} = \frac{dU}{d\mu} \quad \text{and} \quad Ue_\beta = \frac{\psi^{(1)}}{\psi}(a)\frac{dU}{d\tilde{P}_a} = \frac{dU}{dP}, \tag{35}$$

where $a = \phi(P)$, and the derivatives are Leslie derivatives if $\beta = \lambda_0$, and (E1) does not hold. These equations establish the measure-theoretic meaning of the (hitherto abstract) tangent vector U.

The maps m_β and e_β can be used to investigate the regularity of statistical divergences on M. The usual extension of the KL divergence to sets of finite measures, such as M, is [2]:

$$\begin{aligned}\mathcal{D}(P \mid Q) &= Q(\mathbb{R}^d) - P(\mathbb{R}^d) + \mathbf{E}_\mu p \log(p/q) \\ &= \mathbf{E}_\mu m_1(Q) - \mathbf{E}_\mu m_1(P) + \mathbf{E}_\mu(m_\alpha(P) + 1)\big(e_\beta(P) - e_\beta(Q)\big),\end{aligned} \tag{36}$$

where $1 < \alpha, \beta \leq \lambda_0$, $\alpha^{-1} + \beta^{-1} = 1$, and $\mathbf{E}_\mu$ is expectation (integration) with respect to μ. This clearly requires $\lambda_0 \geq 2$. The following corollary of Lemma 4(i) can be proved by induction, and careful use of Hölder's inequality.

Corollary 1 *If (E1) holds and $\lambda_0 \geq 2$ then, for any $0 \leq i, j \leq \lfloor\lambda_0\rfloor - 1$ with $1 \leq i + j \leq \lfloor\lambda_0\rfloor$, $\mathcal{D} \in C^{i,j}(M \times M; \mathbb{R})$ with derivatives:*

$$\mathcal{D}(\phi^{-1}|\phi^{-1})^{(i,j)}_{(a,b)}\big(u_1, \ldots, u_i; v_1, \ldots, v_j\big) = \mathbf{E}_\mu F_{i,j}(a, b)u_1 \cdots u_i v_1 \cdots v_j, \tag{37}$$

where void products take the value 1,

$$F_{i,0}(a,b) = \sum_{l=0}^{i} \binom{i}{l} \psi^{(l)}(a)\theta^{(i-l)}(a) - \psi^{(i)}(a)(1+\theta(b)) \tag{38}$$

$$F_{0,j}(a,b) = \psi^{(j)}(b) - \psi(a)\theta^{(j)}(b) \tag{39}$$

$$F_{i,j}(a,b) = -\psi^{(i)}(a)\theta^{(j)}(b) \quad \text{if } i, j \geq 1 \tag{40}$$

and $\theta : \mathbb{R} \to \mathbb{R}$ *is defined by* $\theta(z) = z - \psi(z) + 1 = \log \psi(z)$.

If (E1) does not hold then the condition on the sum $i + j$ becomes $1 \leq i + j \leq \lceil \lambda_0 \rceil - 1$ (see Corollary 5.1 in [27]). This is equivalent if λ_0 takes a non-integer value. If, on the other hand, $\lambda_0 \in \mathbb{N}$ then $\mathcal{D}$ admits fewer mixed Fréchet derivatives in the absence of (E1); however, the highest mixed derivatives in (40) (those for which $i + j = \lambda_0$) then exist in the Leslie sense. Similar results hold for Amari's α-divergences, for $\alpha \in [-1, 1]$.

We can now use Eguchi's characterisation of the Fisher–Rao metric on $T_P M$ [15]: for any $U, V \in T_P M$,

$$\langle U, V \rangle_P := -UV\mathcal{D} = \langle Um_2, Ve_2 \rangle_{L^2(\mu)} = \mathbf{E}_\mu \frac{p}{(1+p)^2} U\phi V\phi, \tag{41}$$

where U acts on the first argument of $\mathcal{D}$, and V acts on the second. It follows that $\langle V, U \rangle_P = \langle U, V \rangle_P$ and, for any $y \in \mathbb{R}$, $\langle yU, V \rangle_P = \langle U, yV \rangle_P = y\langle U, V \rangle_P$; furthermore,

$$\langle U, U \rangle_P \leq \mathbf{E}_\mu (U\phi)^2 \leq \|U\phi\|_G^2, \tag{42}$$

and $\langle U, U \rangle_P = 0$ if and only if $U\phi = 0$. So the metric is positive definite and dominated by the chart-induced norm on $T_P M$. However the Fisher–Rao metric and chart-induced norm are not equivalent, even when the model space is $L^2(\mu)$ [25]. In the general, infinite-dimensional case $(T_P M, \langle \cdot, \cdot \rangle_P)$ is not a Hilbert space; the Fisher–Rao metric is a *weak* Riemannian metric.

If $\lambda_0 \geq 3$ then M also admits the Amari–Chentsov tensor. This is the symmetric covariant 3-tensor field defined by

$$\tau_P(U, V, W) = \mathbf{E}_\mu Um_3 Ve_3 We_3 = \mathbf{E}_\mu \frac{p}{(1+p)^3} U\phi V\phi W\phi. \tag{43}$$

The Fisher–Rao metric and higher-order covariant tensor fields are smoother with increasing values of λ_0. Log-Sobolev embedding enhances this gain for particular integer values of λ_0. Suppose, for example, that $\lambda_0 = 2$. If (E1) holds then the metric is a continuous covariant 2-tensor field on M; however if (E1) does not hold then, although the composite map $M \ni P \mapsto \langle \mathbf{U}(P), \mathbf{V}(P) \rangle_P \in \mathbb{R}$ is continuous for all continuous vector fields $\mathbf{U}$, $\mathbf{V}$, the metric is not continuous in the sense of the operator norm.

If $\lambda_0 \geq 2$ then the variables m_2 and e_2 are *bi-orthogonal* representations of measures in M. This can be seen in the following *generalised cosine rule*:

$$\begin{aligned}\mathcal{D}(P|R) = \mathcal{D}(P|Q) + \mathcal{D}(Q|R)&\\ -\langle m_2(P) - m_2(Q), e_2(R) - e_2(Q)\rangle_{L^2(\mu)}.&\end{aligned} \tag{44}$$

Setting $R = P$ and using the fact that $m_2 + e_2 = \imath \circ \phi$, where $\imath : G \to L^2(\mu)$ is the inclusion map, we obtain the global bound

$$\begin{aligned}\mathcal{D}(P|Q) + \mathcal{D}(Q|P) &= \langle m_2(P) - m_2(Q)\,,\ e_2(P) - e_2(Q)\rangle_{L^2(\mu)}\\ &\leq \frac{1}{2}\|\phi(P) - \phi(Q)\|^2_{L^2(\mu)} \leq \frac{1}{2}\|\phi(P) - \phi(Q)\|^2_G.\end{aligned} \tag{45}$$

4 Special model spaces

The construction of M and TM in the previous section is valid for any of the weighted mixed-norm spaces developed in Sect. 2, including the *fixed norm* space $G_f := W^{k,(\lambda,\ldots,\lambda)}(\mu)$. However, certain spaces are particularly suited to the deformed exponential function ψ; these are introduced next. A special class of mixed-norm spaces, on which the nonlinear superposition operators associated with ψ *act continuously*, is developed in Sect. 4.1. Section 4.2 investigates fixed-norm spaces and shows that, with the exception of the cases $k = 1$, $\lambda \in [1, \infty)$ and $k = 2$, $\lambda = 1$, they do not share this property.

4.1 A family of mixed norm spaces

This section develops the mixed-norm space $G_m := W^{k,\Lambda}(\mu)$ with $\lambda_0 \geq \lambda_1 \geq k$ and $\lambda_j = \lambda_1/j$ for $2 \leq j \leq k$. Lemma 4 can be augmented as follows.

Proposition 2 (i) *For any $a \in G_m$, $\psi(a) \in G_m$.*
(ii) *The nonlinear superposition (Nemytskii) operator $\Psi_m : G_m \to G_m$, defined by $\Psi_m(a)(x) = \psi(a(x))$, is continuous.*
(iii) *$\Psi_m(G_m)$ is convex.*

Proof A *partition* of $s \in S_1$ is a set $\pi = \{\sigma_1, \ldots, \sigma_n \in S_1\}$ such that $\sum_i \sigma_i = s$. Let $\Pi(s)$ denote the set of distinct partitions of s and, for any $\pi \in \Pi(s)$, let $|\pi|$ denote the number of d-tuples in π. According to the Faá di Bruno formula, for any $s \in S_1$ and any $f \in C^\infty(\mathbb{R}^d; \mathbb{R})$,

$$D^s\psi(f) = F_s(f) := \sum_{\pi \in \Pi(s)} K_\pi \psi^{(|\pi|)}(f) \prod_{\sigma \in \pi} D^\sigma f, \tag{46}$$

where the $K_\pi < \infty$ are combinatoric constants. $D^s\psi(f) \in C^\infty(\mathbb{R}^d; \mathbb{R})$ since the derivatives of ψ are bounded and $D^\sigma f \in C^\infty(\mathbb{R}^d; \mathbb{R})$ for all $\sigma \in \pi$. We set $F_0 := \psi$, and extend the domain of F_s to G_m in the obvious way.

Let $(f_n \in C^\infty(\mathbb{R}^d;\mathbb{R}))$ be a sequence converging in the sense of G_m to a. Since the first derivative of ψ is bounded, the mean value theorem shows that $\psi(f_n) \to \psi(a) = F_0(a)$ in the sense of $L^{\lambda_0}(\mu)$. Furthermore, for any $s \in S_1$,

$$\begin{aligned}\left|D^s\psi(f_n) - F_s(a)\right| \le K \sum_{\pi\in\Pi(s)} \left|\psi^{(|\pi|)}(f_n)\right| \Gamma_{\pi,n} \\ +K \sum_{\pi\in\Pi(s)} \left|\psi^{(|\pi|)}(f_n) - \psi^{(|\pi|)}(a)\right| \prod_{\sigma\in\pi} |D^\sigma a|, \qquad (47)\end{aligned}$$

where

$$\Gamma_{\pi,n} := \left|\prod_{\sigma\in\pi} D^\sigma f_n - \prod_{\sigma\in\pi} D^\sigma a\right| \le \sum_{\sigma\in\pi} |D^\sigma(f_n - a)| \prod_{\tau\in\pi\setminus\{\sigma\}} \big(|D^\tau f_n| + |D^\tau a|\big).$$

Now $\sum_{\sigma\in\pi} |\sigma| = |s|$, and so it follows from Hölder's inequality that

$$\|\Gamma_{\pi,n}\|_{L^{\lambda/|s|}(\mu)} \le \sum_{\sigma\in\pi} \|D^\sigma(f_n - a)\|_{L^{\lambda/|\sigma|}(\mu)} \prod_{\tau\in\pi\setminus\{\sigma\}} \big\||D^\tau f_n| + |D^\tau a|\big\|_{L^{\lambda/|\tau|}(\mu)},$$

which, together with the boundedness of the derivatives of ψ, shows that the first term on the right-hand side of (47) converges to zero in the sense of $L^{\lambda/|s|}(\mu)$. The second term converges to zero in probability and is dominated by the function $C\prod_{\sigma\in\pi} |D^\sigma a| \in L^{\lambda/|s|}(\mu)$ for some $C < \infty$, and so it also converges to zero in the sense of $L^{\lambda/|s|}(\mu)$. We have thus shown that, for any $s \in S_0$, $D^s\psi(f_n)$ converges to $F_s(a)$ in the sense of $L^{\lambda/|s|}(\mu)$. In particular, $F_s(a) \in L^{\lambda/|s|}(\mu)$. That $\psi(a)$ is weakly differentiable with derivatives $D^s\psi(a) = F_s(a)$, for all $s \in S_1$, follows from arguments similar to those in (11) with f_n playing the role of a_n, and this completes the proof of part (i).

Let $(a_n \in G_m)$ be a sequence converging to a in the sense of G_m. The above arguments, with a_n replacing f_n, show that, for any $s \in S_0$, $F_s(a_n) \to F_s(a)$ in the sense of $L^{\lambda/|s|}(\mu)$, and this completes the proof of part (ii).

For any $P_0, P_1 \in M$ and any $y \in (0,1)$, let $P_y := (1-y)P_0 + yP_1$. Clearly $p_y \in G_m$; we must show that $\log p_y \in G_m$. Let $f : (0,\infty) \to \mathbb{R}$ be defined by

$$f(z) = \mathbf{1}_{(0,1)}(z)(-\log z)^\lambda + \mathbf{1}_{[1,\infty)}(z)(z-1)^\lambda;$$

then $|\log z|^\lambda \le f(z)$, and f is of class C^2 with non-negative second derivative, and so is convex. It follows from Jensen's inequality that

$$\mathbf{E}_\mu|\log p_y|^\lambda \le \mathbf{E}_\mu f(p_y) \le (1-y)\mathbf{E}_\mu f(p_0) + y\mathbf{E}_\mu f(p_1) < \infty.$$

A further application of the Faá di Bruno formula shows that, for any $s \in S_1$,

$$\begin{aligned}|D^s \log p_y| &\le K_1 \sum_{\pi \in \Pi(s)} |\log^{(|\pi|)}(p_y)| \prod_{\sigma \in \pi} |D^\sigma p_y| \\ &\le K_2 \sum_{\pi \in \Pi(s)} \prod_{\sigma \in \pi} \left| \frac{D^\sigma p_0}{p_0} + \frac{D^\sigma p_1}{p_1} \right|.\end{aligned}$$

Now $p_i = \psi(a_i)$ for some $a_0, a_1 \in G_m$, and so $D^\sigma p_i / p_i = F_\sigma(a_i)/\psi(a_i)$. Since $\psi^{(n)}/\psi$ is bounded, the arguments used above to show that $D^s \psi(a_i) \in L^{\lambda/|s|}(\mu)$ can be used to show that $D^\sigma \psi(a_i)/\psi(a_i) \in L^{\lambda/|\sigma|}(\mu)$. Hölder's inequality then shows that $D^s \log p_y \in L^{\lambda/|s|}(\mu)$. We have thus shown that $\log p_y \in G_m$. So $P_y \in M$, and this completes the proof of part (iii). □

4.2 Fixed norm spaces

Proposition 2 shows that the function ψ defines a superposition operator that "acts continuously" on the mixed norm Sobolev space G_m. The question naturally arises whether or not it has this property with respect to any fixed norm spaces (other than $W^{1,(1,1)}(\mu)$, which is a special instance of G_m). Since, for $k \ge 2$ and $\lambda \ge \lambda_0$, the space $G_f := W^{k,(\lambda,\ldots,\lambda)}(\mu)$ is a subset of G_m and has a topology stronger than the relative topology, it is clear that $\Psi_m(G_f) \subset \Psi_m(G_m) \subset G_m$, and that the restriction of Ψ_m to G_f, is continuous. However, except in one specific case, ψ does not define a superposition operator with domain and range G_f, as the following proposition shows.

Proposition 3 *If $\lambda > 1$ and $k \ge 2$ then there exists an $a \in G_f$ for which $\psi(a) \notin G_f$.*

Proof (Adapted from Dahlberg's counterexample) Let $t \in (0, 2]$, $z_t \ge 0$ and $l_t : \mathbb{R}^d \to \mathbb{R}$ be as in Sect. 2, and let $\{B_n \subset \mathbb{R}^d, n \in \mathbb{N}\}$ be the sequence of closed spheres with centres $\sigma_n = (n^{1/t}, 0, \ldots, 0)$ and radii $1/n$. If $x \in B_n$ then $|l_t(x) - l_t(\sigma_n)| \le K/\sqrt{n}$ for some $K < \infty$. Let $\varphi \in C_0^\infty(\mathbb{R}^d; \mathbb{R})$ be such that

$$\varphi(y) = y_1 \quad \text{if} \quad |y| \le 1/2 \quad \text{and} \quad \varphi(y) = 0 \quad \text{if} \quad |y| \ge 1.$$

Since ψ is not a polynomial, its kth derivative $\psi^{(k)}$ is not identically zero, and we can choose $-\infty < \zeta_1 < \zeta_2 < \zeta_1 + 1$ such that $|\psi^{(k)}(z)| \ge \epsilon$ for all $z \in [\zeta_1, \zeta_2]$ and some $\epsilon > 0$. Finally, let $a : \mathbb{R}^d \to \mathbb{R}$ be defined by the sum

$$a(x) = \zeta_1 + \sum_{n=m}^{\infty} \alpha^n \varphi(n(x - \sigma_n)), \tag{48}$$

where $\alpha = \exp(2/((k+1)\lambda - 1))$ and $m = \lfloor z_t \rfloor + 16$. (The support of the nth term in the sum here is a subset of B_n, and so a is well defined and of class C^∞.) We claim that $a \in G_f$; in fact, for any $s \in S_1$ with $|s| = j$,

$$\begin{aligned}\mathbf{E}_\mu |D^s a|^\lambda &\le K \sum_{n=m}^{\infty} \alpha^{\lambda n} n^{j\lambda} \exp(-n) \int |D^s \varphi(n(x-\sigma_n))|^\lambda dx \\ &= K \sum_{n=m}^{\infty} \alpha^{\lambda n} n^{j\lambda - d} \exp(-n) \int |D^s \varphi(y)|^\lambda dy < \infty, \qquad (49)\end{aligned}$$

and a similar bound can be found for $\mathbf{E}_\mu |a - \zeta_1|^\lambda$. It now suffices to show that $D^s \psi(a) \notin L^\lambda(\mu)$, where $s = (k, 0, \ldots, 0)$. Let

$$T_n := \left\{ x \in \mathbb{R}^d : |x - \sigma_n| \le 1/2n \text{ and } 0 \le (x-\sigma_n)_1 \le (n\alpha^n)^{-1}(\zeta_2 - \zeta_1) \right\};$$

then, for any $x \in T_n$, $a(x) = \zeta_1 + n\alpha^n (x - \sigma_n)_1 \in [\zeta_1, \zeta_2]$, and so

$$\begin{aligned}\mathbf{E}_\mu |D^s \psi(a)|^\lambda &\ge \sum_{n=m}^{\infty} \alpha^{k\lambda n} n^{k\lambda} \int_{T_n} |\psi^{(k)}(\zeta_1 + n\alpha^n (x-\sigma_n)_1)|^\lambda r(x)\, dx \\ &\ge K_1 \epsilon^\lambda \sum_{n=m}^{\infty} \alpha^{k\lambda n} n^{k\lambda} \exp(-n) |T_n| \\ &= K_2 \epsilon^\lambda \sum_{n=m}^{\infty} \alpha^{(k\lambda - 1)n} n^{k\lambda - d} \exp(-n) = +\infty, \qquad (50)\end{aligned}$$

where $|T_n|$ is the Lebesgue measure of T_n, and this completes the proof. □

As (49) shows, no amount of "derivative sacrifice" will overcome this property of G_f: there is no choice of $k \le m < \infty$ such that $\psi(a) \in G_f$ for all $a \in W^{m,(\lambda,\ldots,\lambda)}(\mu)$. However, we are able to prove the following, which includes the case $k = 2, \lambda = \nu = 1$.

Proposition 4 *Let $k \ge 2$, let $\lambda \ge k - 1$ and let $\nu := (\lambda + 1)/k$.*

(i) *For any $a \in G_f$, $\psi(a) \in W^{k,(\nu,\ldots,\nu)}(\mu)$.*
(ii) *The nonlinear superposition operator $\Psi_f : G_f \to W^{k,(\nu,\ldots,\nu)}(\mu)$, defined by $\Psi_f(a)(x) = \psi(a(x))$, is continuous.*

Proof As in the proof of Proposition 2, it suffices to show that, for any $a \in G_f$, any sequence $(a_n \in G_f)$ converging to a in G_f, and any $s \in S_0$, $F_s(a_n) \to F_s(a)$ in $L^\nu(\mu)$, where F_s is as defined in (46). For any s with $|s| < k$ this can be accomplished by means of Hölder's inequality, as in the proof of Proposition 2. Furthermore, even if $|s| = k$, all terms in the sum on the right-hand side of (46) for which $|\pi| < k$ can be treated in the same way. (There are no more than $k - 1$ factors in the product, each of which is in $L^\lambda(\mu)$, and $\lambda/(k-1) \ge \nu$.) This leaves the terms for which $|\pi| = |s| = k$; in order to show that these converge in $L^\nu(\mu)$ it suffices to show that, for any $1 \le i \le d$, the sequence $(|\psi^{(k)}(a_n)(a_n')^k|^\nu)$ is uniformly integrable, where, for any weakly differentiable $g : \mathbb{R}^d \to \mathbb{R}$, $g' := \partial g / \partial x_i$.

Let $\rho \in C_0^\infty(\mathbb{R}^d;\mathbb{R})$ be as defined in (15), and let

$$K_\rho := \sup_x \big(\rho(x) + 2|\rho'(x)| + |\rho''(x)|\big).$$

Let $h : G_f \to L^1(\mu)$ be defined by $h(a) = |a| + |a'| + (|a| + |a'| + |a''|)^\lambda$; then $h(K_\rho a_n) \to h(K_\rho a)$ in $L^1(\mu)$ and so, according to the Lebesgue–Vitaly theorem, $(h(K_\rho a_n))$ is a uniformly integrable sequence. So, according to the de la Vallée Poussin theorem, there exists a convex increasing function $\tilde{F} : [0,\infty) \to [0,\infty)$ such that $\tilde{H}(z) := \tilde{F}(z)/z$ is an unbounded, non-decreasing function and $\sup_n \mathbf{E}_\mu \tilde{F}(h(K_\rho a_n)) < \infty$. Let $H : [0,\infty) \to [0,\infty)$ be defined by

$$H(z) = \left\{\begin{matrix} 0 & \text{if } z = 0 \\ z^{-1}\int_0^z \tilde{H}(y)\,dy & \text{otherwise.} \end{matrix}\right\} \le \tilde{H}(z) \tag{51}$$

For any $y \in [0,\infty)$, let $z_y := \inf\{z \in [0,\infty) : \tilde{H}(z) \ge y\}$; for any $z > 2z_y$,

$$H(z) = z^{-1}\int_0^{z_y} \tilde{H}(t)\,dt + z^{-1}\int_{z_y}^{z} \tilde{H}(t)\,dt \ge (z - z_y)y/z \ge y/2,$$

and so H is also unbounded. Furthermore

$$zH^{(1)}(z) = \tilde{H}(z) - H(z) \in [0, \tilde{H}(z)]. \tag{52}$$

Summarising the above, H is unbounded, non-decreasing and differentiable, and so $F : [0,\infty) \to [0,\infty)$, defined by $F(z) = zH(z)$ is another de la Vallée Poussin function for which $\sup_n \mathbf{E}_\mu F(h(K_\rho a_n)) < \infty$.

Let $G : [0,\infty) \to [0,\infty)$ be defined by $G(z) = zH(|z/C|^{1/k\nu})$, where $C := \sup_z |\psi^{(k)}(z)|^\nu$; then, for any $f \in C_0^\infty(\mathbb{R}^d;\mathbb{R})$,

$$\begin{aligned}
\mathbf{E}_\mu G\Big(|\psi^{(k)}(f)|^\nu |f'|^{k\nu}\Big) &\le K_1 \mathbf{E}_\mu |\psi^{(k)}(f)||f'|^{k\nu} H(|f'|) \\
&\le K_2 \mathbf{E}_\mu \frac{\psi^{(1)}(f)}{(1+\psi(f))^k} |f'|^{k\nu} H(|f'|) \\
&= K_2 \int \frac{\psi(f)'}{(1+\psi(f))^k} f'|f'|^{k\nu-2} H(|f'|) r\,dx \\
&= K_3 \int (1+\psi(f))^{1-k} \Big[|f'|^{k\nu-2} f'' H(|f'|) \\
&\qquad + |f'|^{k\nu-1} H^{(1)}(|f'|) f'' + f'|f'|^{k\nu-2} H(|f'|) l' \Big] r\,dx \\
&\le K_3(R(f) + S(f) + T(f)),
\end{aligned} \tag{53}$$

where K_1 and K_2 depend only on the function ψ, $K_3/K_2 = (k\nu - 1)/(k-1)$ and $R(f)$, $S(f)$ and $T(f)$ are as follows:

$$\begin{aligned} R(f) &:= \mathbf{E}_\mu |f'|^{k\nu-2}|f''|H(|f'|), \\ S(f) &:= \mathbf{E}_\mu |f'|^{k\nu-1}H^{(1)}(|f'|)|f''| \le \mathbf{E}_\mu |f'|^{k\nu-2}|f''|\tilde{H}(|f'|), \\ T(f) &:= \mathbf{E}_\mu |f'|^{k\nu-1}H(|f'|)|l'|. \end{aligned} \tag{54}$$

In (53), we have used the boundedness of $\psi^{(k)}$ in the first step, Lemma 3(ii) in the second step and integration by parts with respect to x_i in the fourth step. (If $t = 1$ in Example 1(i), then $\theta_t(|\cdot|)$ is not differentiable at 0 and the integration by parts has to be accomplished separately on the two sub-intervals $(-\infty, 0)$ and $(0, \infty)$.) In (54), we have used (52). Let $a_{m,n} := a_n(x)\rho(x/m) \in \mathcal{U}_m$; then, with J as defined in section 2,

$$\begin{aligned} R(Ja_{m,n}) &\le \mathbf{E}_\mu F\big(|(Ja_{m,n})'| + (|(Ja_{m,n})'| + |(Ja_{m,n})''|)^\lambda\big) \\ &= \mathbf{E}_\mu F\big(|Ja'_{m,n}| + (|Ja'_{m,n}| + |Ja''_{m,n}|)^\lambda\big) \\ &\le \mathbf{E}_\mu JF\big(|a'_{m,n}| + (|a'_{m,n}| + |a''_{m,n}|)^\lambda\big) \\ &\le \mathbf{E}_\mu F\big(|a'_{m,n}| + (|a'_{m,n}| + |a''_{m,n}|)^\lambda\big) + 1/m \\ &\le \mathbf{E}_\mu F(h(K_\rho a_n)) + 1, \end{aligned}$$

where we have used the definition of F in the first step, Lemma 1(ii) in the second step, Jensen's inequality in the third step, Lemma 1(i) in the fourth step and (16) in the final step. Similar bounds can be found for $S(Ja_{m,n})$ and, if $t \in (0, 1]$ (so that l' is bounded), $T(Ja_{m,n})$.

If $t \in (1, 2]$ we note that

$$\big(|f'|^\lambda H(|f'|)\big)' = \lambda |f'|^{\lambda-1}\mathrm{sgn}(f')f''H(|f'|) + |f'|^\lambda H^{(1)}(|f'|)\mathrm{sgn}(f')f'',$$

so that $\mathbf{E}_\mu |(|f'|^\lambda H(|f'|))'| \le \lambda R(f) + S(f)$, and

$$\| |(Ja_{m,n})'|^\lambda H(|(Ja_{m,n})'|)\|_{W^{1,(1,1)}(\mu)} \le (\lambda + 2)(\mathbf{E}_\mu \tilde{F}(h(K_\rho a_n)) + 1).$$

Let $\alpha := (t-1)/t$, and let $L^1 \log^\alpha L(\mu)$ and $\exp L^{1/\alpha}(\mu)$ be the complementary Orlicz spaces defined in the proof of Lemma 4. It follows from Theorem 7.12 in [13] that, for some $K_4 < \infty$ not depending on m or n,

$$\| |(Ja_{m,n})'|^\lambda H(|(Ja_{m,n})'|)\|_{L^1 \log^\alpha L(\mu)} \le K_4(\lambda + 2)(\mathbf{E}_\mu \tilde{F}(h(K_\rho a_n)) + 1).$$

For any $|x_i| > z_t$, $l'(x) = -t|x_i|^{t-1}\mathrm{sgn}(x_i)$, and so $l' \in \exp L^{1/\alpha}(\mu)$, and the generalised Hölder inequality shows that, for some $K_5 < \infty$

$$T(Ja_{m,n}) \le K_5 \| |(Ja_{m,n})'|^\lambda H(|(Ja_{m,n})'|)\|_{L^1 \log^\alpha L(\mu)} \|l'\|_{\exp L^{1/\alpha}(\mu)}.$$

We have thus shown that, for any $t \in (0, 2]$,

$$\mathbf{E}_\mu G\Big(|\psi^{(k)}(Ja_{m,n})|^\nu |(Ja_{m,n})'|^{k\nu}\Big) \le K_6(\mathbf{E}_\mu \tilde{F}(h(K_\rho a_n)) + 1), \tag{55}$$

where $K_6 < \infty$ does not depend on m or n. Since G is a de la Vallée Poussin function, the sequence $(|\psi^{(k)}(Ja_{m,n})|^\nu |(Ja_{m,n})'|^{k\nu},\ m \in \mathbb{N})$, for any fixed n, is uniformly integrable and so, according to Lemma 2, converges in $L^1(\mu)$ to $|\psi^{(k)}(a_n)|^\nu |a_n'|^{k\nu}$. Fatou's theorem now shows that

$$\mathbf{E}_\mu G\Big(|\psi^{(k)}(a_n)|^\nu |a_n'|^{k\nu}\Big) \le K_6(\mathbf{E}_\mu \tilde{F}(h(K_\rho a_n)) + 1),$$

which in turn shows that the sequence $(|\psi^{(k)}(a_n)|^\nu |a_n'|^{k\nu},\ n \in \mathbb{N})$ is uniformly integrable. □

If we want all spatial derivatives of $\psi(a)$ to be continuous maps from G_f to $L^\nu(\mu)$ (for some $\nu \ge 1$) then the fixed norm space G_f should have Lebesgue exponent $\lambda = \max\{2, \nu k - 1\}$. (The resulting manifold will not have a strong enough topology for global information geometry unless $\lambda \ge 2$.) The mixed norm space G_m requires $\lambda_1 = \nu k, \lambda_2 = \nu k/2, \ldots, \lambda_k = \nu$. This places a slightly higher integrability constraint on the first derivative, but lower constraints on all other derivatives (significantly lower if $k \ge 3$). Furthermore, if G_f is used as a model space, then $\psi(a)$ and its first partial derivatives actually belong to $L^\lambda(\mu)$, and so the true range of the superposition operator in this context is a mixed norm space, whether or not we choose to think about it in this way.

The case in which $\lambda = 1$ is of particular interest. Proposition 4 then shows that ψ defines a nonlinear superposition operator that acts continuously on $G_s := W^{2,(1,1,1)}(\mu)$. The use of such a low Lebesgue exponent precludes the results in Sect. 3 concerning the smoothness of the KL-divergence. In particular, we cannot expect to retain global geometric constructs such as the Fisher–Rao metric. However, $\mathcal{D}(\mu\,|\,\cdot\,) : M_s \to [0, \infty)$ is still continuous for all $t \in (0, 2]$, and $\mathcal{D}(\,\cdot\,|\,\mu)$ is finite if $t = 2$. Since $\psi^{(1)}$ is bounded, there is no difficulty in extending these results as follows.

Corollary 2 *For any $\lambda_0 \in [1, \infty)$, ψ defines a nonlinear superposition operator that acts continuously on $G_{ms} := W^{2,(\lambda_0,1,1)}(\mu)$.*

Remark 2 When the model space, G, is G_m, G_s or G_{ms}, then condition (M2) can be replaced by: (M2') $p, \log p \in G$.

5 The manifolds of probability measures

In this section we shall assume that $\lambda_0 > 1$, or that $\lambda_0 = 1$ and the embedding hypothesis (E1) holds. Let $M_0 \subset M$ be the subset of the general manifold of Sect. 3 (that modelled on $G := W^{k,\Lambda}(\mu)$), whose members are *probability* measures. These satisfy the additional hypothesis:

(M3) $\mathbf{E}_\mu p = 1$.

The co-dimension 1 subspaces of $L^\lambda(\mu)$ and G, whose members, a, satisfy $\mathbf{E}_\mu a = 0$ will be denoted $L_0^\lambda(\mu)$, and G_0. Let $\phi_0 : M_0 \to G_0$ be defined by

$$\phi_0(P) = \phi(P) - \mathbf{E}_\mu \phi(P) = \log_d p - \mathbf{E}_\mu \log_d p. \tag{56}$$

Proposition 5 (i) *ϕ_0 is a bijection onto G_0. Its inverse takes the form*

$$P(dx) = \psi(a(x) + Z(a))\mu(dx), \tag{57}$$

where $Z \in C^N(G_0; \mathbb{R})$ is an (implicitly defined) normalisation function, and $N = N(\lambda_0, \lambda_1, 1, t)$ is as defined in (26).

(ii) *The first (and if $N \geq 2$ second) derivative of Z is as follows:*

$$\begin{aligned} Z_a^{(1)} u &= -\mathbf{E}_{P_a} u \\ Z_a^{(2)}(u, v) &= -\frac{\mathbf{E}_\mu \psi^{(2)}(a + Z(a))(u - \mathbf{E}_{P_a} u)(v - \mathbf{E}_{P_a} v)}{\mathbf{E}_\mu \psi^{(1)}(a + Z(a))}, \end{aligned} \tag{58}$$

where $P_a := \tilde{P}_{a+Z(a)} / \tilde{P}_{a+Z(a)}(\mathbb{R}^d)$ and $\tilde{P}_a(dx) = \psi^{(1)}(a(x))\mu(dx)$.

(iii) *If $\lambda_0 - 1 \in \mathbb{N}$ and (E1) does not hold then $Z^{(\lambda_0 - 1)}$ is Leslie differentiable (with derivative is as in (58) if $\lambda_0 = 2$).*

Proof Let $\Upsilon : G_0 \times \mathbb{R} \to (0, \infty)$ be defined by

$$\Upsilon(a, z) = \mathbf{E}_\mu \psi(a + z) = \mathbf{E}_\mu \Psi_1(a + z), \tag{59}$$

where Ψ_β is as in Lemma 4. It follows from Lemma 4, that Υ is of class C^N and that, for any $u \in G_0$,

$$\Upsilon_{a,z}^{(1,0)} u = \mathbf{E}_\mu \psi^{(1)}(a + z) u \quad \text{and} \quad \Upsilon_{a,z}^{(0,1)} = \mathbf{E}_\mu \psi^{(1)}(a + z) > 0. \tag{60}$$

Since ψ is convex,

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

furthermore, the monotone convergence theorem shows that

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

So $\Upsilon(a, \cdot)$ is a bijection with strictly positive derivative, and the inverse function theorem shows that it is a C^N-isomorphism. The implicit mapping theorem shows that $Z : G_0 \to \mathbb{R}$, defined by $Z(a) = \Upsilon(a, \cdot)^{-1}(1)$, is of class C^N. For some $a \in G_0$, let P be the probability measure on $\mathcal{X}$ with density $p = \psi(a + Z(a))$; then $\phi_0(P) = a$ and $P \in M_0$, which proves part (i).

That the first derivative of Z is as in (58) follows from (60). Since $\mathbf{E}_\mu \psi^{(1)}(a + Z(a)) > 0$, parts (ii) and (iii) follow from Lemma 4 and the chain and quotient rules of differentiation (which hold for Leslie derivatives). □

Expressed in charts, the inclusion map $\imath : M_0 \to M$ is as follows

$$\rho(a) := \phi \circ \phi_0^{-1}(a) = a + Z(a), \tag{61}$$

and has the same smoothness properties as Z. The following goes further.

Proposition 6 *(M_0, G_0, ϕ_0) is a C^N-embedded submanifold of (M, G, ϕ), where $N = N(\lambda_0, \lambda_1, 1, t)$ is as defined in (26).*

Proof Let $\eta : G \to G_0$ be the superposition operator defined by $\eta(a)(x) = a(x) - \mathbf{E}_\mu a$; then η is of class C^∞, has first derivative $\eta_a^{(1)} u = u - \mathbf{E}_\mu u$, and zero higher derivatives. Now $\eta \circ \rho$ is the identity map of G_0, which shows that ρ is homeomorphic onto its image, $\rho(G_0)$, endowed with the relative topology. Furthermore, for any $u \in G_0$,

$$u = (\eta \circ \rho)_a^{(1)} u = \eta_{\rho(a)}^{(1)} \rho_a^{(1)} u,$$

and so $\rho_a^{(1)}$ is a toplinear isomorphism, and its image, $\rho_a^{(1)} G_0$, is a closed linear subspace of G. Let E_a be the one dimensional subspace of G defined by $E_a = \{y\psi^{(1)}(\rho(a)) : y \in \mathbb{R}\}$. If $u \in E_a$ and $v \in \rho_a^{(1)} G_0$ then there exist $y \in \mathbb{R}$ and $w \in G_0$ such that

$$\mathbf{E}_\mu uv = y\mathbf{E}_\mu \psi^{(1)}(\rho(a))(w - \mathbf{E}_{P_a} w) = 0.$$

So $E_a \cap \rho_a^{(1)} G_0 = \{0\}$, and $\rho_a^{(1)}$ *splits* G into the direct sum $E_a \oplus \rho_a^{(1)} G_0$. We have thus shown that ρ is a C^N-immersion, and this completes the proof. □

For any $P \in M_0$, the tangent space $T_P M_0$ is a subspace of $T_P M$ of co-dimension 1; in fact, as shown in the proof of Proposition 6,

$$T_P M = T_P M_0 \oplus \{y\hat{U},\ y \in \mathbb{R}\}, \quad \text{where } \hat{U}\phi = \psi^{(1)}(\phi(P)). \tag{62}$$

Let $\Phi_0 : TM_0 \to G_0 \times G_0$ be defined as follows:

$$\Phi_0(P, U) = \Phi(P, U) - \mathbf{E}_\mu \Phi(P, U). \tag{63}$$

Then $\Phi \circ \Phi_0^{-1}(a, u) = (\rho(a), \rho_a^{(1)} u)$. For any $(P, U) \in TM_0$, $U\phi = \rho_a^{(1)} u = u - \mathbf{E}_{P_a} u$, and so tangent vectors in $T_P M_0$ are distinguished from those merely in $T_P M$ by the fact that their total mass is zero:

$$U(\mathbb{R}^d) = \int (u - \mathbf{E}_{P_a} u) d\tilde{P}_a = 0. \tag{64}$$

The map Z of (57) is (the negative of) the additive normalisation function, α, associated with the interpretation of M_0 as a *generalised exponential model* with deformed exponential function ψ. (See Chapter 10 in [23]. We use the symbol Z rather than $-\alpha$ for reasons of consistency with [25,27].) In this context, the probability measure P_a of (58) is called the *escort* measure. In [21], the authors considered *local* charts on the Hilbert manifold of [25]. In the present context, these take the form $\phi_P : M_0 \to G_P$, where G_P is the subspace of G whose members, b, satisfy $\mathbf{E}_{P_a} b = 0$. This amounts to re-defining the origin of G as $\phi(P)$, and using the co-dimension 1 subspace that is tangential to the image $\phi(M_0)$ at this new origin as the model space. This local chart is *normal* at P for the Riemannian metric and Levi–Civita parallel transport induced by the global chart ϕ on M. However, the metric differs from the Fisher–Rao metric on all fibres of the tangent bundle except that at μ. The local model space, G_P, is based on the reference measure μ, not the local measure P, as is the case with the exponential Orlicz manifold.

The equivalent on M_0 of the maps m_β and e_β of Sect. 3 are the maps $m_{\beta,0}, e_{\beta,0} : M_0 \to L_0^\beta(\mu)$, defined by

$$m_{\beta,0}(P) = m_\beta(P) \quad \text{and} \quad e_{\beta,0}(P) = e_\beta(P) - \mathbf{E}_\mu e_\beta(P). \tag{65}$$

Their properties follow from those of m_β and e_β.

6 Application to nonlinear filtering

We sketch here an application of the manifolds of Sects. 3 and 5 to the nonlinear filtering problem discussed in Sect. 1. An abstract filtering problem (in which X is a Markov process evolving on a measurable space) was investigated in [26]. Under suitable technical conditions, it was shown that the $(Y_s,\ 0 \le s \le t)$-conditional distribution of X_t, Π_t, satisfies an infinite-dimensional stochastic differential equation on the Hilbert manifold of [25], and this representation was used to study the filter's information-theoretic properties. This equation involves the normalisation constant Z, which is difficult to use since it is implicitly defined, and so it is of interest to use a manifold of finite measures not involving Z, such as M of Sect. 3.

If the conditional distribution Π_t has a density with respect to Lebesgue measure, p_t, satisfying the Kushner–Stratonovich Eq. (3), then its density with respect to μ, $\pi_t = p_t/r$, also satisfies (3), but with the transformed forward operator:

$$\mathcal{A}\pi = \frac{1}{2r}\frac{\partial^2 \Gamma^{ij} r\pi}{\partial x^i \partial x^j} - \frac{1}{r}\frac{\partial f^i r\pi}{\partial x^i}, \tag{66}$$

where $\Gamma = gg^*$ and we have used the Einstein summation convention. The density π_t also satisfies

$$d\pi_t = \mathcal{A}\pi_t\, dt + \pi_t\big(h - \bar{h}(\pi_t)\big)\big(dY_t - \bar{h}(\pi_t)dt\big), \tag{67}$$

where, for appropriate densities p, $\bar{h}(p) := (\mathbf{E}_\mu p)^{-1}\mathbf{E}_\mu ph$. This equation is *homogeneous*, in the sense that if π_t is a solution then so is $\alpha\pi_t$, for any $\alpha > 0$. A straightforward formal calculation shows that $\log_d \pi_t$ satisfies the following stochastic partial differential equation

$$da_t = \mathbf{u}(\,\cdot\,, a_t)dt + \mathbf{v}(\,\cdot\,, a_t)\big(dY_t - \bar{h}(\psi(a_t))dt\big), \tag{68}$$

where

$$\begin{aligned}
\mathbf{v}(x,a) &= (1+\psi(a(x)))(h(x)-\bar{h}(\psi(a))),\\
\mathbf{u}(x,a) &= \frac{1}{2}\Gamma^{ij}(x)\left[\frac{\partial^2 a}{\partial x^i \partial x^j}(x) + (1+\psi(a(x)))^{-2}\frac{\partial a}{\partial x^i}(x)\frac{\partial a}{\partial x^j}(x)\right]\\
&\quad + F^i(x)\frac{\partial a}{\partial x^i}(x) + (1+\psi(a(x)))F^0(x) - \frac{1}{2}\big[h(x)-\bar{h}(\psi(a))\big]^2,
\end{aligned} \tag{69}$$

and

$$\begin{aligned}
F^i &= \Gamma^{ij}\frac{\partial l}{\partial x^j} + \frac{\partial \Gamma^{ij}}{\partial x^j} - f^i,\\
F^0 &= \frac{1}{2}\frac{\partial^2 \Gamma^{ij}}{\partial x^i \partial x^j} + \frac{\partial \Gamma^{ij}}{\partial x^i}\frac{\partial l}{\partial x^j} + \frac{1}{2}\Gamma^{ij}\left[\frac{\partial^2 l}{\partial x^i \partial x^j} + \frac{\partial l}{\partial x^i}\frac{\partial l}{\partial x^j}\right] - f^i\frac{\partial l}{\partial x^i} - \frac{\partial f^i}{\partial x^i}.
\end{aligned}$$

In order to make sense of (68) and (69), we need further hypotheses. The following are used for illustration purposes, and are not intended to be ripe.

(F1) $G = G_m$ (the mixed norm space of Sect. 4.1) with $\lambda_0 = \lambda_1 \geq 2k \geq 4$; $t = 1$ and $\theta_1(|\cdot|) \in C^\infty(\mathbb{R}; [0,\infty))$.
(F2) The functions f, g and h are of class $C^\infty(\mathbb{R}^d)$.
(F3) The functions f and h, and all their derivatives, satisfy polynomial growth conditions in $|x|$.
(F4) The function g and all its derivatives are bounded.

In particular, these allow $\bar{h}$, $\mathbf{u}$ and $\mathbf{v}$ to be defined on M in a precise way.

Proposition 7 (i) *The functional $\bar{H} : G_m \to \mathbb{R}$, defined by $\bar{H}(a) = \bar{h}(\psi(a))$, is of class $C^{\lceil\lambda_0\rceil-1}$.*
(ii) *Let $k \geq 2$ and $\lambda_1 \geq 2k$. If $a \in G_m$ then $\mathbf{u}(\,\cdot\,,a), \mathbf{v}(\,\cdot\,,a) \in H^{k-2}(\mu)$, where $H^{k-2}(\mu)$ is the Hilbert Sobolev space of Remark 1.*
(iii) *The superposition operators $\mathbf{U}, \mathbf{V} : G_m \to H^{k-2}(\mu)$, defined by $\mathbf{U}(a)(x) = \mathbf{u}(x,a)$ and $\mathbf{V}(a)(x) = \mathbf{v}(x,a)$, are continuous.*

Proof It follows from (F1–F4) that

$$F^i, F^0, h, \in W^{k,(\lambda,\lambda,\ldots,\lambda)}(\mu) \quad \text{for every } k \in \mathbb{N}, \text{ and every } \lambda \in [1,\infty). \tag{70}$$

Lemma 4 shows that, for any $\epsilon > 0$, $\Psi_{1+\epsilon}$ is of class $C^{\lceil\lambda_0/(1+\epsilon)\rceil-1}$. For any $\lambda_0 \in [2,\infty)$ there exists an $\epsilon > 0$ such that $\lceil\lambda_0/(1+\epsilon)\rceil = \lceil\lambda_0\rceil$ and so with this choice,

$\Psi_{1+\epsilon}$ is of class $C^{\lceil\lambda_0\rceil-1}$. Hölder's inequality shows that, for any $0 \le i \le \lceil\lambda_0\rceil - 2$, any $a, b \in G_m$ and any $u_1, \ldots, u_i$ in the unit ball of G_m,

$$\begin{aligned}
&\mathbf{E}_\mu\big|(\Psi^{(i)}_{1+\epsilon,b} - \Psi^{(i)}_{1+\epsilon,a} - \Psi^{(i+1)}_{1+\epsilon,a}(b-a))(u_1, \ldots, u_i)h\big| \\
&\quad \le \big\|\Psi^{(i)}_{1+\epsilon,b} - \Psi^{(i)}_{1+\epsilon,a} - \Psi^{(i+1)}_{1+\epsilon,a}(b-a)\big\|_{L(G^i_m; L^{1+\epsilon}(\mu))}\|h\|_{L^{(1+\epsilon)/\epsilon}(\mu)}, \\
&\mathbf{E}_\mu\big|(\Psi^{(i+1)}_{1+\epsilon,b} - \Psi^{(i+1)}_{1+\epsilon,a})h\big| \le \big\|\Psi^{(i+1)}_{1+\epsilon,b} - \Psi^{(i+1)}_{1+\epsilon,a}\big\|_{L(G^{i+1}_m; L^{1+\epsilon}(\mu))}\|h\|_{L^{(1+\epsilon)/\epsilon}(\mu)},
\end{aligned}$$

which shows that the map $G_m \ni a \mapsto \mathbf{E}_\mu\psi(a)h \in \mathbb{R}$ is of class $C^{\lceil\lambda_0\rceil-1}$. The quotient rule of differentiation and the fact that $\mathbf{E}_\mu\psi(a) > 0$ complete the proof of part (i).

Parts (ii) and (iii) can be proved by applying Hölder's inequality to the weak derivatives of the various components of $\mathbf{u}(\,\cdot\,, a)$ and $\mathbf{v}(\,\cdot\,, a)$. The quadratic term in $\mathbf{u}$ is the most difficult to treat, and so we give a detailed proof for this. We begin by noting that $(1 + \psi(a))^{-1}\partial a/\partial x^i = \partial(a - \psi(a))/\partial x^i$. For any $|s| \le k - 2$

$$D^s \frac{\partial\psi(a)}{\partial x^i}\frac{\partial\psi(a)}{\partial x^j} = \sum_{\sigma\le s} D^\sigma \frac{\partial\psi(a)}{\partial x^i} D^{s-\sigma}\frac{\partial\psi(a)}{\partial x^j} \prod_{1\le l\le d}\binom{s_l}{\sigma_l}. \tag{71}$$

According to Proposition 2, the nonlinear superposition operator $\Psi_{\sigma,i} : G_m \to L^{\lambda_1/(|\sigma|+1)}(\mu)$ defined by $\Psi_{\sigma,i}(a) = D^\sigma(\partial\psi(a)/\partial x^i)$ is continuous, and so it follows from Hölder's inequality that the same is true of $\Upsilon_{s,i,j} : G_m \to L^{\lambda_1/(|s|+2)}(\mu)$ defined by the right-hand side of (71). Together with (F4), this shows that $\Upsilon : G_m \to H^{k-2}(\mu)$ defined by $\Upsilon(a) = \Gamma^{ij}(\partial\psi(a)/\partial x^i)(\partial\psi(a)/\partial x^j)$ is continuous.

The other components of $\mathbf{u}(\,\cdot\,, a)$ and the only component of $\mathbf{v}(\,\cdot\,, a)$ can be shown to have the stated continuity by similar arguments. These make use of (70), Proposition 2 and part (i) here. □

Remark 3 There are many variants of this proposition, corresponding to different choices of the domain and range of $\mathbf{U}$ and $\mathbf{V}$. If λ_0 and λ_1 are suitably large, then $\mathbf{U}$ and $\mathbf{V}$ admit various derivatives on M.

One application of Proposition 7 is in the development of *projective approximations*, as proposed in the context of the exponential Orlicz manifold in [8] and the earlier references therein. As a particular instance, suppose that $k \ge 2$ and $\lambda_1 \ge 2k$; let $(\eta_i \in C^k(\mathbb{R}^d) \cap G_m,\ 1 \le i \le m)$ be linearly independent, and define

$$G_{m,\eta} = \Big\{a \in G_m : a = \alpha^i\eta_i \text{ for some } \alpha \in \mathbb{R}^m\Big\}. \tag{72}$$

This is an m-dimensional linear subspace of both G_m and $H^{k-2}(\mu)$. We can use the inner product of $H^{k-2}(\mu)$ to project members of $H^{k-2}(\mu)$ onto $G_{m,\eta}$. In particular, we can project $\mathbf{U}(a)$ and $\mathbf{V}(a)$ onto $G_{M,\eta}$ for any $a \in G_{m,\eta}$ to obtain continuous vector fields of the finite-dimensional submanifold of M defined by $M_\eta = \phi^{-1}(G_{m,\eta})$. Since the model space norms of $H^{k-2}(\mu)$ dominate the Fisher–Rao metric on every fibre of the tangent bundle (42), the projection takes account of the information theoretic cost of approximation, as well as controlling the derivatives of the conditional density π_t.

M_η is a finite-dimensional *deformed exponential model*, and is trivially a C^∞-embedded submanifold of M. Many other classes of finite-dimensional manifold also have this property. For example, since $\Psi(G_m)$ is convex, certain finite-dimensional mixture manifolds modelled on the space $G_{m,\eta}$, where $\eta_i \in \Psi(G_m)$, are also C^∞-embedded submanifolds of M. This is also true of particular finite-dimensional exponential models.

7 Concluding remarks

This paper has developed a class of infinite-dimensional statistical manifolds that use the balanced chart of [25,27] in conjunction with a variety of probability spaces of Sobolev type. It has shown that the mixed-norm space of Sect. 4.1 is especially suited to the balanced chart (and any other chart with similar properties), in the sense that densities (and log-densities) then also belong to this space and vary continuously on the manifolds. It has shown that this property is also true of a particular fixed norm space involving two derivatives, but can be retained for fixed norm spaces with more than two derivatives only with the loss of Lebesgue exponent. The paper has outlined an application of the manifolds to nonlinear filtering (and hence to the Fokker–Planck equation). Although motivated by problems of this type, the manifolds are clearly applicable in other domains, the Boltzmann equation of statistical mechanics being an obvious candidate.

The deformed exponential function used in the construction of M has *linear growth*, a feature that has recently been shown to be advantageous in *quantum information geometry* [24]. The linear growth arises from the deformed logarithm of (21), which is dominated by the density, p, when the latter is large. As recently pointed out in [21], this property is shared by other deformed exponentials, notably the Kaniadakis 1-exponential $\psi_K(z) = z + \sqrt{1 + z^2}$. The corresponding deformed logarithm is $\log_K(y) = (y^2 - 1)/2y$, and so the density is controlled (when close to zero) by the term $-1/p$ rather than $\log p$, as used here. In the non-parametric setting, the need for both p and $1/p$ to be in $L^{\lambda_0}(\mu)$ places significant restrictions on membership of the manifold. If, for example, the reference measure of Example 1(i) is used, and $t = 1$, then the measure having density $C \exp(-\alpha|x|)$ (with respect to Lebesgue measure) belongs to the manifold only if $|\alpha - 1| < 1/\lambda_0$.

The Kaniadakis 1-exponential shares the properties of ψ used in this paper; these are summarised in Lemma 5, which is easily proved by induction.

Lemma 5 (i) *The Kaniadakis 1-exponential* $\psi_K : \mathbb{R} \to (0, \infty)$ *is diffeomorphic; in particular*

$$\psi_K^{(1)} = \frac{2\psi_K^2}{1 + \psi_K^2} > 0 \quad \text{and} \quad \psi_K^{(2)} = \frac{8\psi_K^3}{\left(1 + \psi_K^2\right)^3} > 0, \tag{73}$$

and so ψ_K *is strictly increasing and convex.*

(ii) *For any $n \geq 2$,*

$$\psi_K^{(n)} = \frac{Q_{3(n-2)}(\psi_K)}{\left(1+\psi_K^2\right)^{2(n-1)}}\psi_K\psi_K^{(1)}, \tag{74}$$

where $Q_{3(n-2)}$ is a polynomial of degree no more than $3(n-2)$. In particular, $\psi_K^{(n)}$, $\psi_K^{(n)}/\psi_K$ and $\psi_K^{(n)}/(\psi_K\psi_K^{(1)})$ are all bounded.

We can therefore construct a manifold of finite measures M_K, as in Sect. 3, substituting the chart of (21) by $\phi_K : M_K \to G$, defined by $\phi_K(P) = \log_K p$. The only properties of ψ used in Sect. 3 are its strict positivity, and the boundedness of its derivatives, properties shared by ψ_K. The results in Sect. 4 carry over to M_K with the exception of Proposition 2(iii). Most of these depend only on the boundedness of the derivatives of ψ; however, the integration by parts in (53) uses (20), which can be substituted by (74) in the case of M_K. The results of Sect. 5 all carry over to M_K.

M_K is a subset of M. Let $\tau : \mathbb{R} \to \mathbb{R}$ be the "transition function" $\tau(z) = \log_d \psi_K(z)$. All derivatives of τ are bounded, which explains why the regularity of the KL-divergence on M carries over to M_K. Furthermore, it follows from arguments similar to those used in the proof of Proposition 2 that the superposition operator $T_m : G_m \to G_m$ defined by $T_m(a)(x) = \tau(a(x))$ is continuous for any of the mixed norm model spaces of Sect. 4.1.

The deformed logarithm of (21) was chosen in [25] because the resulting manifold is highly inclusive, and suited to the Shannon–Fisher–Rao information geometry. In this context, it yields the global bound (45). Condition (6) (on the reference measure μ) has to be considered in the context of (M2), which places upper and lower bounds on the rate at which the densities of measures in M can decrease as $|x|$ becomes large. For example, if all nonsingular Gaussian measures are to belong to M, then (M2) requires r to decay more slowly than a Gaussian density, but more rapidly than a Cauchy density. Variants of the reference measure μ with $t \in [1, 2)$ may be good choices for such applications.

Compliance with ethical standards

Conflict of interest The author states that there is no conflict of interest.

References

1. Amari, S.-I., Barndorff-Nielsen, O.E., Kass, R.E., Lauritzen, S.L., Rao, C.R.: Differential geometry and statistical inference, Lecture Note Monograph Series, vol. 10. Institute of Mathematical Statistics, Hayward CA (1978)

2. Amari, S.-I., Nagaoka, H.: Methods of Information Geometry, Translations of Mathematical Monographs, vol. 191. American Mathematical Society, Providence (2000)
3. Appell, J., Zabrejko, P.P.: Nonlinear Superposition Operators. Cambridge University Press, Cambridge (1990)
4. Ay, N., Jost, J., Van Lê, H., Schwachhöfer, L.: Information geometry and sufficient statistics. Probab. Theory Relat. Fields **162**, 327–364 (2015)
5. Ay, N., Jost, J., Van Lê, H., Schwachhöfer, L.: Parametrized measure models. In: Information Geometry. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge/A Series of Modern Surveys in Mathematics, vol. 64, pp. 121–184 . Springer (2017)
6. Barndorff-Nielsen, O.E.: Information and Exponential Families in Statistical Theory. Chichester (1978)
7. Bauer, M., Bruveris, M., Michor, P.W.: Uniqueness of the Fisher–Rao metric on the space of smooth densities. Bull. Lond. Math. Soc. **48**, 499–506 (2016)
8. Brigo, D., Pistone, G.: Dimensionality reduction for measure valued evolution equations in statistical manifolds. In: Nielsen, F., Critchley, F., Dodson, C.T.J. (eds.) Computational Information Geometry for Image and Signal Processing. Springer, Cham (2017)
9. Bruveris, M., Michor, P.W.: Geometry of the Fisher–Rao metric on the space of smooth densities on a compact manifold (2016). arXiv:1607.04550
10. Cena, A., Pistone, G.: Exponential statistical manifold. Ann. Inst. Stat. Math. **59**, 27–56 (2007)
11. Chentsov, N.N.: Algebraic foundation of mathematical statistics. Math. Operationsforsch. Stat. Ser. Stat. **9**, 267–276 (1978)
12. Chentsov, N.N.: Statistical Decision Rules and Optimal Inference, Translations of Mathematical Monographs, vol. 53. American Mathematical Society, Providence (1982)
13. Cianchi, A., Pick, L., Slavíková, L.: Higher-order Sobolev embeddings and isoperimetric inequalities. Adv. Math. **273**, 568–650 (2015)
14. Crisan, D., Rozovskiĭ, B.: The Oxford Handbook of Nonlinear Filtering. Oxford University Press, Oxford (2011)
15. Eguchi, S.: Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Stat. **11**, 793–803 (1983)
16. Gibilisco, P., Pistone, G.: Connections on non-parametric statistical manifolds by Orlicz space geometry. Infin. Dimens. Anal. Quantum. Prob Relat. Top. **1**, 325–347 (1998)
17. Leslie, J.A.: On a differential structure for the group of diffeomorphisms. Topology **46**, 263–271 (1967)
18. Leslie, J.A.: Some Frobenius theorems in global analysis. J. Differ. Geom. **42**, 279–297 (1968)
19. Lods, B., Pistone, G.: Information geometry formalism for the spatially homogeneous Boltzmann equation. Entropy **17**, 4323–4363 (2015)
20. Loaiza, G., Quiceno, H.R.: A q-exponential statistical Banach manifold. J. Math. Anal. Appl. **398**, 466–476 (2013)
21. Montrucchio, L., Pistone, G.: A class of non-parametric deformed exponential statistical models. In: Nielsen, F. (ed.) Geometric Structures of Information, Signals and Communication Technology, pp. 15–35. Springer, Cham (2019)
22. Murray, M.K., Rice, J.W.: Differential Geometry and Statistics, Monographs in Statistics and Applied Probability, vol. 48. Chapman Hall, London (1993)
23. Naudts, J.: Generalised Thermostatistics. Springer, London (2011)
24. Naudts, J.: Quantum statistical manifolds: the linear growth case (2018). arXiv:1801.07642
25. Newton, N.J.: An infinite-dimensional statistical manifold modelled on Hilbert space. J. Funct. Anal. **263**, 1661–1681 (2012)
26. Newton, N.J.: Information geometric nonlinear filtering. Infin. Dimens. Anal. Quantum Prob. Relat. Top. **18**, 1550014 (2015)
27. Newton, N.J.: Infinite-dimensional statistical manifolds based on a balanced chart. Bernoulli **22**, 711–731 (2016)
28. Newton, N.J.: Nonlinear filtering and information geometry: a Hilbert manifold approach. In: Ay, N., Gibilisco, P., Matúš, F. (eds.) Information Geometry and its Applications, Proceedings in Mathematics and Statistics, vol. 252, pp. 189–208. Springer, Cham (2018)
29. Newton, N.J.: Manifolds of differentiable densities. ESAIM Probab. Stat. **22**, 19–34 (2018). https://doi.org/10.1051/ps/2018003
30. Pistone, G., Rogantin, M.P.: The exponential statistical manifold: mean parameters, orthogonality and space transformations. Bernoulli **5**, 721–760 (1999)

31. Pistone, G., Sempi, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. Ann. Stat. **23**, 1543–1561 (1995)
32. Rao, C.R.: Information and accuracy obtainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–91 (1945)
33. Rao, M.M., Ren, Z.D.: Theory of Orlicz Spaces. Marcel Dekker, New York (1991)
34. Runst, T., Sickel, W.: Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations. de Gruyter, Berlin (2011)
35. Vigelis, R.F., Cavalcante, C.C.: On φ-families of probability distributions. J. Theor. Probab. **26**, 870–884 (2013)