A general spline representation for nonparametric and semiparametric density estimates using diffeomorphisms

Ethan Anderes[1] **and Marc Coram**[2]

[1]*Department of Statistics, University of California, Davis CA 95616, USA. e-mail:* anderes@stat.ucdavis.edu

[2]*Department of Health Research and Policy (Biostatistics), Stanford University, Palo Alto, CA 94305, USA. e-mail:* mcoram@stanford.edu

Abstract: A theorem of McCann [15] shows that for any two absolutely continuous probability measures on $\mathbb{R}^d$ there exists a monotone transformation sending one probability measure to the other. A consequence of this theorem, relevant to statistics, is that density estimation can be recast in terms of transformations. In particular, one can fix any absolutely continuous probability measure, call it $\mathbb{P}$, and then reparameterize the whole class of absolutely continuous probability measures as monotone transformations from $\mathbb{P}$. In this paper we utilize this reparameterization of densities, as monotone transformations from some $\mathbb{P}$, to construct semiparametric and nonparametric density estimates. We focus our attention on classes of transformations, developed in the image processing and computational anatomy literature, which are smooth, invertible and which have attractive computational properties. The techniques developed for this class of transformations allow us to show that a penalized maximum likelihood estimate (PMLE) of a smooth transformation from $\mathbb{P}$ exists and has a finite dimensional characterization, similar to those results found in the spline literature. These results are derived utilizing an Euler-Lagrange characterization of the PMLE which also establishes a surprising connection to a generalization of Stein's lemma for characterizing the normal distribution.

Keywords and phrases: Euler-Lagrange, density estimation, penalized maximum likelihood, diffeomorphism.

1. Introduction

Smooth invertible transformations, or deformations, are fast becoming important tools in modern data analysis. They have been used with spectacular success in the field of computational anatomy where time varying vector field flows, which generate deformations, are used to statistically analyze medical fMRI images and quantify abnormal morphological structure (see [1, 5, 6, 7, 8, 11, 13, 16, 17, 18, 23, 24, 25, 27, 28], and references therein). In cosmology, deformations are used to model gravitational distortions of the cosmic microwave background from dark matter density fluctuations and have resulted in a deeper understanding of cosmic structure [12]. Transformations or deformations also have the power to recast the generic problem of density estimation to that of deformation estimation. In particular, one can fix any absolutely continuous probability measure, call it $\mathbb{P}$, and then reparameterize the whole class of absolutely continuous probability measures as monotone transformations from $\mathbb{P}$ (this follows by results in [15]). The advantage of this new viewpoint is the flexibility in choosing the target measure $\mathbb{P}$ which can encode prior information on the shape of true sampling distribution. For example, if it is known that the data is nearly Gaussian then choosing a Gaussian $\mathbb{P}$ along with a strong penalty on transformations that are far from the identity allows one to construct penalized maximum likelihood estimates which effectively shrink the resulting nonparametric estimate in the direction of the Gaussian target $\mathbb{P}$. Moreover, when there is no knowledge about the true sampling measure, one can simply choose any absolutely continuous $\mathbb{P}$ and still construct a completely nonparametric density estimate.

Recent work in [2], [3] and [19] utilize this idea of representing classes of probability measures as deformations $\phi : \mathbb{R}^d \to \mathbb{R}^d$ of a known target probability measure $\mathbb{P}$ on $\mathbb{R}^d$. The principle difficulty when working with such models is the construction of a rich class of deformations which are non-

parametric, smooth, invertible and which are amenable to optimization. In [2] and [3] the authors the utilize the class of quasi-conformal maps to generate penalized maximum likelihood estimates of ϕ. However, these tools were only developed for $\mathbb{R}^2$ with no clear generalization for higher dimension. In [19] the authors use polynomial approximations to ϕ which could potentially violate the invertabliltiy requirements on ϕ. In this paper, we circumvent these challenges by adapting the powerful tools developed by Grenander, Miller, Younes, Trouvé and co-authors in the image processing and computational anatomy literature (see [27] and the references therein) to generate estimates of ϕ with all the required properties: nonparametric flexibility, smoothness, invertability and computational tractability. We establish the existence of a penalized maximum likelihood estimate of ϕ which has a finite dimensional characterization similar to those results found in the spline literature (see [26]). This finite dimensional characterization is a key component of the numerical computation of these estimates which are nominally defined as a infinite dimensional minimizer of a penalized likelihood. Moreover, our results are derived utilizing an Euler-Lagrange characterization of the PMLE which also establishes a surprising connection to a generalization of Stein's lemma for characterizing the normal distribution.

We start the paper in Section 2 with an overview of using the dynamics of time varying vector field flows to generate a rich class of diffeomorphisms. Then in sections 3 and 4 we define our penalized maximum likelihood estimate (PMLE) of ϕ and prove not only existence, but also establish a finite dimensional characterization which is key for numerically computing the resulting density estimate. In Section 5 we notice a surprising connection with the Euler-Lagrange equation for the PMLE of ϕ and a generalization of Stein's lemma for characterizing the normal distribution (see [22]). In sections 6 and 7 we give examples of our new density estimate, first as a nonparametric density estimate and second as a semiparametric density estimate where a finite dimensional model is used for the target probability measure $\mathbb{P}$. We finish the paper with an appendix which contains some technical details used for the proofs of the existence of the PMLE and for the finite dimensional characterization.

2. A rich class of diffeomorphisms

In this section we give a brief overview of using the dynamics of time varying vector field flows to generate rich classes of diffeomorphisms. These time varying flows have been utilized in the field of computational anatomy and image processing (see [28], and references therein) and have been shown to be very powerful for developing algorithms which optimize a diverse range of objective functions defined over classes of diffeomorphism. It is these tools that we adapt for density estimation and statistics.

A map $\phi\colon \Omega \to \mathbb{R}^d$ is said to be a $C^k(\Omega, \mathbb{R}^d)$ diffeomorphism of the open set $\Omega \subset \mathbb{R}^d$ if ϕ is one-to-one, maps onto Ω, and $\phi, \phi^{-1} \in C^k(\Omega, \mathbb{R}^d)$. In what follows we generate classes of diffeomorphisms by time varying vector field flows. In particular, let $\{v_t\}_{t\in[0,1]}$ be a time varying vector field in $\mathbb{R}^d$, where t denotes 'time' so that for each t, v_t is a function mapping Ω into $\mathbb{R}^d$. Under mild smoothness conditions there exists a unique class of diffeomorphisms of Ω, denoted $\{\phi^v_t\}_{t\in[0,1]}$, which satisfy the following ordinary differential equation

$$\partial_t \phi^v_t(x) = v_t(\phi^v_t(x)) \tag{1}$$

with boundary condition $\phi^v_0(x) = x$, for all $x \in \Omega$ (see Theorem 1 below). The interpretation of these flows is that $\phi_t(x)$ represents the position of a particle at time t, which originated from location x at time $t = 0$, and flowed according the the instantaneous velocity given by v_t. It will be convenient to consider the diffeomorphism that maps time t to some other time s, this will be denoted $\phi^v_{ts}(x) \equiv \phi^v_s(\phi^{v\,-1}_t(x))$.

For the remainder of the paper we will assume that at each time t, v_t will be a member of a Hilbert space of vector fields mapping Ω into $\mathbb{R}^d$ with inner product denoted by $\langle \cdot, \cdot \rangle_V$ and norm by $\| \cdot \|_V$. Indeed, how one chooses the Hilbert space V will determine the smoothness properties of the resulting class of deformations $\{\phi_t\}_{t\in[0,1]}$. Once the Hilbert space V is fixed we can define the following set of time varying vector fields.

Definition 1. *Let $V^{[0,1]}$ denote the space of measurable functions $v_t(x)\colon [0,1]\times\Omega\to\mathbb{R}^d$ such that $v_t\in V$ for all $t\in[0,1]$ and $\int_0^1\|v_t\|_V^2dt<\infty$.*

One clear advantage of this class is that it can be endowed with a Hilbert space inner product if V is a Hilbert space. Indeed, $V^{[0,1]}$ is a Hilbert space with inner product defined by $\langle v,h\rangle_{V^{[0,1]}}\equiv\int_0^1\langle v_t,h_t\rangle_V dt$ (see Proposition 1 in the Appendix or Proposition 8.17 in [27]). For the remainder of the paper we typically use v or w to denote elements of $V^{[0,1]}$ and v_t or w_t to denote the corresponding elements of V at any fixed time t. An important theorem found in [27] relates the smoothness of V to the smoothness of the resulting diffeomorphism. Before we state the theorem, some definitions will be prudent. The Hilbert space V is said to be continuously embedded in another normed space H (denoted $V\hookrightarrow H$) if $V\subset H$ and there exists a constant c such that

$$\|v\|_H\le c\|v\|_V$$

for all $v\in V$ where $\|\cdot\|_H$ denotes the norm in H. Also we let $C_0^k(\Omega,\mathbb{R}^d)$ denote the subset of $C^k(\Omega,\mathbb{R}^d)$ functions whose partial derivatives of order k or less all have continuous extensions to zero at the boundary $\partial\Omega$.

Theorem 1 ([27], [11])**.** *If $V\hookrightarrow C_0^k(\Omega,\mathbb{R}^d)$, then for any $v\in V^{[0,1]}$ there exists a unique class of $C^k(\Omega,\mathbb{R}^d)$ diffeomorphisms $\{\phi_t^v\}_{t\in[0,1]}$ which satisfy (1) and $\phi_0^v(x)=x$ for all $x\in\Omega$.*

To derive our finite dimensional characterization we will make the additional assumption that V is a reproducing kernel Hilbert space of vector fields. This will guarantee the existence of a reproducing kernel $K(x,y)\colon\Omega\times\Omega\to\mathbb{R}^{n\times n}$ which can be used to compute the evaluation functional. In particular, K has the property that for any $x\in\Omega$ and $f\in V$ the following identity holds $\langle K(x,\cdot)p,f\rangle_V=p^Tf(x)$ for all column vectors $p\in\mathbb{R}^d$. To simplify the following computations we will only work with kernels of the form $K(x,y)=R(x,y)I_{d\times d}$ where $R:\Omega\times\Omega\to\mathbb{R}$ is a positive definite function and $I_{d\times d}$ is the d-by-d identity matrix.

Now the class of diffeomorphisms we consider in this paper corresponds to the set of all time varying vector field flows evaluated at $t=1$: ϕ_1^v, where v ranges through $V^{[0,1]}$. The class $V^{[0,1]}$ will be completely specified by the reproducing kernel $R(x,y)I_{d\times d}$ which has the flexibility to control the smoothness of the resulting maps ϕ_1^v through Theorem 1.

3. Penalized maximum likelihood estimation

Formally, we model our data $X_1,\dots,X_n$ as independent samples of an *unknown* diffeomorphism ϕ_1^v of a *known* probably distribution $\mathbb{P}$ on $\mathbb{R}^d$. In particular,

$$X_1,\dots,X_n\overset{iid}{\sim}\mathbb{P}\circ\phi_1^v \tag{2}$$

where ϕ_1^v is generated by a time varying vector field flow $\{\phi_t^v\}_{t\in[0,1]}$ which satisfies (1) and $v\in V^{[0,1]}$. As remarked in the introduction, by simply choosing any absolutely continuous measure $\mathbb{P}$ the model (2) is still completely nonparametric. The advantage is that when partial information exists on the sampling distribution of the data, it can potentially be encoded in the choice of $\mathbb{P}$. The notation $\mathbb{P}\circ\phi_1^v$, in (2), is taken to mean that the probability of $X\in A$ is given by $\mathbb{P}(\phi_1^v(A))$ where $\phi_1^v(A)=\{\phi(x)\colon x\in A\}$. An important observation is that the model (2) implies that $\phi_1^v(X)\sim\mathbb{P}$. Therefore, one can imagine estimating ϕ_1^v by attempting to "deform" the data $X_1,\dots,X_n$ by a transformation which satisfies

$$\phi_1^v(X_1),\dots,\phi_1^v(X_n)\overset{iid}{\sim}\mathbb{P}.$$

In this section, we construct a penalized maximum likelihood estimate (PMLE) of $\hat v$ given the data (2), whereby obtaining an estimate $\mathbb{P}\circ\phi_1^{\hat v}$ of the true sampling distribution.

The target probability measure $\mathbb{P}$ is assumed to have a bounded density with respect to Lebesque measure on $\mathbb{R}^d$. Therefore, by writing the density of $\mathbb{P}$ as $\exp H$ for some function $H:\mathbb{R}^d\to\mathbb{R}\cup\{-\infty\}$,

the probability measure $\mathbb{P} \circ \phi_1^v$ has density given by

$$d\,\mathbb{P} \circ \phi_1^v(x) = \det(D\phi_1^v(x)) \exp H \circ \phi_1^v(x) dx$$

where $\det(D\phi_1^v(x))$ is defined as the determinant of the Jacobian of ϕ_1^v evaluated at $x \in \Omega$ (always positive by the orientation preserving nature of ϕ_1^v). Since ϕ_1^v ranges over an infinite dimensional space of diffeomorphisms, the likelihood for v given the data will typically be unbounded as v ranges in $V^{[0,1]}$. The natural solution is to regularize the log likelihood using the corresponding Hilbert space norm on $V^{[0,1]}$ with a multiplicative tuning factor $\lambda/2$. The penalized log-likelihood (scaled by $1/n$) for the unknown vector field v flow given data $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P} \circ \phi_1^v$ is then given by

$$E_\lambda(v) \equiv \frac{1}{n} \sum_{k=1}^{n} \log \det[D\phi(X_k)] + H \circ \phi(X_k) - \frac{\lambda}{2} \int_0^1 \|v_t\|_V^2 dt. \tag{3}$$

The estimated vector field $\hat{v}$ is chosen to be any element of $V^{[0,1]}$ which maximizes E_λ over $V^{[0,1]}$. The following theorem establishes that such a $\hat{v}$ exists.

Claim 1. *Let V be a Hilbert space which is continuously embedded in $C_0^2(\Omega, \mathbb{R}^d)$ where Ω is a bounded open subset of $\mathbb{R}^d$. Suppose $e^{H(\cdot)}$ is a bounded and continuous density on Ω. Then there exists a time varying vector field $\hat{v} \in V^{[0,1]}$ such that*

$$E_\lambda(\hat{v}) = \sup_{v \in V^{[0,1]}} E_\lambda(v). \tag{4}$$

Proof. We first establish $\sup_{v \in V^{[0,1]}} E_\lambda(v) < \infty$ by splitting the energy E_λ into three parts

$$E_\lambda(v) = \underbrace{\frac{1}{n} \sum_{k=1}^{n} \log \det D\phi_1^v(X_k)}_{=:E_1(v)} + \underbrace{\frac{1}{n} \sum_{k=1}^{n} H \circ \phi_1^v(X_k)}_{=:E_2(v)} \underbrace{- \frac{\lambda}{2} \int_0^1 \|v_t\|_V^2 dt}_{=:E_3(v)}. \tag{5}$$

Notice that each term is well defined and finite whenever $v \in V^{[0,1]}$, since the assumption $V \hookrightarrow C_0^2(\Omega, \mathbb{R}^d)$ is sufficient for Theorem 8.7 in [27] to apply to the class $V^{[0,1]}$. In particular, for any $v \in V^{[0,1]}$ there exists a unique class of C^1 diffeomorphisms of Ω, $\{\phi_t^v\}_{t \in [0,1]}$, which satisfies (1) (also see Theorem 2.5 in [11]). The term $E_2(v)$ is clearly bounded from above since $\sup_{x \in \Omega} H(x) < \infty$ by assumption. For the remaining two terms notice that the determinant of the Jacobian is given by $\log \det D\phi_1^v(x) = \int_0^1 \operatorname{div} v_t(\phi_t^v(x)) dt$ (by equation (26) in the Appendix). Therefore

$$\begin{aligned}
E_1(v) + E_3(v) &= \frac{1}{n} \sum_{k=1}^{n} \int_0^1 \left(\operatorname{div} v_t(\phi_t^v(X_k)) - \frac{\lambda}{2} \|v_t\|_V^2 \right) dt \\
&\leq \frac{1}{n} \sum_{k=1}^{n} \int_0^1 \left(\sup_{x \in \Omega} |\operatorname{div} v_t(x)| - \frac{\lambda}{2} \|v_t\|_V^2 \right) dt \\
&\leq \int_0^1 \left(c\|v_t\|_V - \frac{\lambda}{2} \|v_t\|_V^2 \right) dt, \text{ by the assumption } V \hookrightarrow C_0^2(\Omega, \mathbb{R}^d) \\
&\leq \frac{c^2}{2\lambda} < \infty.
\end{aligned}$$

Now let $v^1, v^2, \ldots$ be any maximizing sequence that satisfies $\lim_{m \to \infty} E(v^m) = \sup_{v \in V^{[0,1]}} E_\lambda(v)$. Since $\sup_{v \in V^{[0,1]}} E_\lambda(v) < \infty$ we can construct the sequence v^m so that there exists an $M < \infty$ such that $\|v^m\|_{V^{[0,1]}} \leq M$ for all m. Since Ω is bounded, closed finite balls in $V^{[0,1]} = L^2([0,1], V)$ are weakly compact (by [11]). Therefore we may extract a subsequence from v^m (relabeled by m) which weakly converges to a $\hat{v} \in V^{[0,1]}$. In particular, $\langle v^m, w \rangle_{V^{[0,1]}} \to \langle \hat{v}, w \rangle_{V^{[0,1]}}$ for all $w \in V^{[0,1]}$. Furthermore we have lower semicontinuity of the norm

$$\liminf_{m \to \infty} \|v^m\|_{V^{[0,1]}}^2 \geq \|\hat{v}\|_{V^{[0,1]}}^2. \tag{6}$$

Now by Theorem 3.1 in [11] we have that $\phi_t^{v^m}(x) \to \phi_t^{\hat{v}}(x)$ uniformly in $t \in [0,1]$ as $m \to \infty$. This allows us to show that $\log\det D\phi_1^{v^m}(x) \overset{m\to\infty}{\longrightarrow} \log\det D\phi_1^{\hat{v}}(x)$ for every $x \in \Omega$. To see why, one can use similar reasoning as in [8]. First write

$$|\log\det D\phi_1^{v^m}(x) - \log\det D\phi_1^{\hat{v}}(x)| = \left|\int_0^1 \operatorname{div} v_t^m(\phi_t^{v^m}(x)) - \operatorname{div}\hat{v}_t(\phi_t^{\hat{v}}(x))dt\right| = I + II$$

where the first term I satisfies

$$\begin{aligned}
I &\equiv \left|\int_0^1 \operatorname{div} v_t^m(\phi_t^{v^m}(x)) - \operatorname{div} v_t^m(\phi_t^{\hat{v}}(x))dt\right| \\
&\le \int_0^1 \|\operatorname{div} v_t^m\|_{1,\infty}\left|\phi_t^{v^m}(x) - \phi_t^{\hat{v}}(x)\right|dt \\
&\le \int_0^1 c\|v_t^m\|_V\left|\phi_t^{v^m}(x) - \phi_t^{\hat{v}}(x)\right|dt, \text{ since } V \hookrightarrow C_0^2(\Omega,\mathbb{R}^d) \\
&\le c\|v^m\|_{V^{[0,1]}}\Big[\int_0^1 \underbrace{\left|\phi_t^{v^m}(x) - \phi_t^{\hat{v}}(x)\right|^2}_{=o(1)\text{ uniformly in } t} dt\Big]^{1/2}, \text{ by Hölder.} \\
&\to 0, \text{ since } \|v^m\|_{V^{[0,1]}} \le M \text{ for all } m.
\end{aligned}$$

For the second term II notice that the map sending $v \mapsto \int_0^1 \operatorname{div} v_t(y_t)dt$ is a bounded linear functional on $V^{[0,1]}$ (using the fact that $V \hookrightarrow C_0^1(\Omega,\mathbb{R}^d)$) where $y_t \equiv \phi_t^{\hat{v}}(x)$. By the Riesz representation theorem there exists a $w^{\hat{v}} \in V^{[0,1]}$ such that $\int_0^1 \operatorname{div} v_t(y_t)dt = \langle v, w^{\hat{v}}\rangle_{V^{[0,1]}}$. Therefore

$$\begin{aligned}
II &\equiv \left|\int_0^1 \operatorname{div} v_t^m(\phi_t^{\hat{v}}(x)) - \operatorname{div}\hat{v}_t(\phi_t^{\hat{v}}(x))dt\right| \\
&= \left|\langle v^m - \hat{v}, w^{\hat{v}}\rangle_{V^{[0,1]}}\right| \to 0, \text{ by weak convergence.}
\end{aligned}$$

Combining the results for I and II we can conclude that $\log\det D\phi_1^{v^m}(x) \overset{m\to\infty}{\longrightarrow} \log\det D\phi_1^{\hat{v}}(x)$ for every $x \in \Omega$.

To finish the proof notice that

$$\begin{aligned}
\sup_{v\in V^{[0,1]}} E_\lambda(v) &= \lim_{m\to\infty} E(v^m) = \limsup_{m\to\infty} E(v^m) \\
&= \frac{1}{n}\sum_{k=1}^n \log\det D\phi_1^{\hat{v}}(X_k) + \frac{1}{n}\sum_{k=1}^n H\circ\phi_1^{\hat{v}}(X_k) - \frac{\lambda}{2}\liminf_{m\to\infty}\int_0^1 \|v_t^m\|_V^2 dt \\
&\le \frac{1}{n}\sum_{k=1}^n \log\det D\phi_1^{\hat{v}}(X_k) + \frac{1}{n}\sum_{k=1}^n H\circ\phi_1^{\hat{v}}(X_k) - \frac{\lambda}{2}\int_0^1 \|\hat{v}_t\|_V^2 dt, \quad \text{by (6)} \\
&= E_\lambda(\hat{v})
\end{aligned}$$

□

One of the important facts about any vector field flow $\hat{v} \in V^{[0,1]}$ which maximizes E_λ is that the resulting estimated transformation $\phi_1^{\hat{v}}$ is a geodesic (or minimum energy) flow with respect to the vector field norm $\int_0^1 \|\hat{v}_t\|_V^2 dt$. To see this is first notice that the parameterization of time $t = 1$ maps, ϕ_1^v, by vector fields $v \in V^{[0,1]}$ is a many-to-one parameterization. In other words there exist multiple pairs of vector fields $v, w \in V^{[0,1]}$ such that $\phi_1^v = \phi_1^w$ but $v \neq w$. Notice, however, that the log-likelihood term in E_λ only depends on $\phi_1^{\hat{v}}$. This implies that any maximizer $\hat{v}$ of E_λ must simultaneously minimize the penalty $\int_0^1 \|\hat{v}_t\|_V^2 dt$ over the class of all $w \in V^{[0,1]}$ which has the same terminal value,

i.e. $\phi_1^{\hat{v}} = \phi_1^w$. Consequently, the PMLE estimate $\hat{v}$ must be a geodesic flow. An important consequence is that geodesic flows $\{\hat{v}_t\}_{t\in[0,1]}$ are completely determined by the initial vector field $\hat{v}_0$. This will become particularly important in the next section where the initial velocity field will be completely parameterized by n coefficient vectors.

4. Spline representation from Euler-Lagrange

In this section we work under the additional assumption that V is a reproducing kernel Hilbert space. This assumption allows one to derive the Euler-Lagrange equation for any maximizer $\hat{v}$ of which satisfies (4). This leads to a finite dimensional characterization of $\hat{v}$ which parallel those results found in the spline literature for function estimation.

Claim 2. *Let V be a reproducing kernel Hilbert space, with kernel $R(x,y)I_{d\times d}$, continuously embedded in $C_0^3(\Omega,\mathbb{R}^d)$ where Ω is bounded open subset of $\mathbb{R}^d$. Suppose $e^{H(\cdot)}$ is a $C^1(\bar{\Omega},\mathbb{R})$ density on Ω. Then any time varying vector field $\hat{v}\in V^{[0,1]}$ which satisfies (4) also satisfies the following Euler-Lagrange equation:*

$$\hat{v}_t(x) = \frac{1}{\lambda n}\sum_{k=1}^n \beta_{k,t}^T R(x, X_{k,t}) + \frac{1}{\lambda n}\sum_{k=1}^n \nabla_y^T R(x,y)\Big|_{y=X_{k,t}} \tag{7}$$

where $X_{k,t}\equiv \phi_t^{\hat{v}}(X_k)$, $\beta_{k,t}\equiv \nabla H(X_{k,1})D\phi_{t1}^{\hat{v}}(X_{k,t}) + \nabla \log\det D\phi_{t1}^{\hat{v}}(X_{k,t})$ and $\nabla_y^T = (\partial_{y_1},\ldots,\partial_{y_d})^T$ is the transpose of the gradient operator applied to the y variable.

Proof. Let E_1, E_2 and E_3 decompose E_λ as in (5). Notice first that if $h\in V^{[0,1]}$ and $\epsilon\in\mathbb{R}$ then $2E_3(\hat{v}+\epsilon h) = \lambda\|\hat{v}\|^2_{V^{[0,1]}} + \epsilon 2\lambda\langle v,h\rangle_{V^{[0,1]}} + \epsilon^2\lambda\|h\|^2_{V^{[0,1]}}$. Therefore $E_3(\hat{v}+\epsilon h)$ is differentiable with respect to ϵ with derivative given by

$$\partial_\epsilon E_3(\hat{v}+\epsilon h)\big|_{\epsilon=0} = \int_0^1 \langle h_t, \lambda\hat{v}_t\rangle_V dt. \tag{8}$$

In addition, Theorem 8.10 of [27] implies that $\phi_1^{\hat{v}+\epsilon h}(x)$ is differentiable at $\epsilon=0$. Now, the assumption $H\in C^1(\bar{\Omega})$ combined with equation (31), in the Appendix, gives

$$\begin{aligned}
\partial_\epsilon E_2(\hat{v}+\epsilon h)\big|_{\epsilon=0} &= -\frac{1}{n}\sum_{k=1}^n \nabla H(\phi_1^{\hat{v}}(X_k))\cdot \partial_\epsilon \phi_1^{\hat{v}+\epsilon h}(X_k)\big|_{\epsilon=0} \\
&= -\frac{1}{n}\sum_{k=1}^n \nabla H(\phi_1^{\hat{v}}(X_k))\cdot \int_0^1 \left\{D\phi_{u1}^{\hat{v}} h_u\right\}\circ\phi_u^{\hat{v}}(X_k)\,du \\
&= -\frac{1}{n}\sum_{k=1}^n \int_0^1 \left\{\nabla H(X_{k,1})D\phi_{u1}^{\hat{v}}(X_{k,u})\right\}\cdot h_u(X_{k,u})\,du \\
&= \int_0^1 \left\langle h_u(\cdot), -\frac{1}{n}\sum_{k=1}^n \left\{\nabla H(X_{k,1})D\phi_{u1}^{\hat{v}}(X_{k,u})\right\}^T R(\cdot, X_{k,u})\right\rangle_V du
\end{aligned} \tag{9}$$

Finally, Proposition 3, from the Appendix, implies $E_3(\hat{v}+\epsilon h)$ is differentiable at $\epsilon=0$ with derivative given by

$$\begin{aligned}
\partial_\epsilon E_1(\hat{v}+\epsilon h)\big|_{\epsilon=0} &= -\frac{1}{n}\sum_{k=1}^n \partial_\epsilon \log\det D\phi_1^{\hat{v}+\epsilon h}(X_k)\big|_{\epsilon=0} \\
&= -\frac{1}{n}\sum_{k=1}^n \int_0^1 \left[h_u\cdot\nabla\log\det D\phi_{u1}^{\hat{v}} + \operatorname{div} h_u\right]\circ\phi_u^{\hat{v}}(X_k)\,du
\end{aligned}$$

$$= \int_0^1 \Big\langle h_u(\cdot), -\frac{1}{n}\sum_{k=1}^n \big\{\nabla \log\det D\phi^{\hat v}_{u1}(X_{k,u})\big\}^T R(\cdot, X_{k,u}) + \nabla_y^T R(\cdot,y)\big|_{y=X_{k,u}} \Big\rangle_V \, du \tag{10}$$

Remark: the above equation requires $\partial_{x_i}(e_i \cdot h_u(x)) = \partial_{x_i}\langle e_i R(\cdot,x), h_u\rangle_V = \langle e_i \partial_{x_i} R(\cdot,x), h_u\rangle_V$ which follows since $\operatorname{div} h_u \in V$ by the assumption $V \hookrightarrow C_0^2(\Omega, \mathbb{R}^d)$ (see [4]). Now from (8), (9) and (10), the energy $E_\lambda(\hat v + \epsilon h)$ is differentiable with respect to ϵ at 0 and

$$0 = \partial_\epsilon E_\lambda(\hat v + \epsilon h)\big|_{\epsilon=0} = \langle \mathcal{E}^{\hat v}, h\rangle_{V^{[0,1]}} \tag{11}$$

where

$$\mathcal{E}^{\hat v}_t = \lambda \hat v_t - \frac{1}{n}\sum_{k=1}^n \beta_{k,t}^T R(\cdot, X_{k,t}) - \frac{1}{n}\sum_{k=1}^n \nabla_y^T R(\cdot,y)\big|_{y=X_{k,t}} \tag{12}$$

with $\beta_{k,t} \equiv \nabla H(X_{k,1}) D\phi^{\hat v}_{t1}(X_{k,t}) + \nabla \log\det D\phi^{\hat v}_{t1}(X_{k,t})$. Since $h \in V^{[0,1]}$ was arbitrary, equation (11) implies $\mathcal{E}^{\hat v} = 0$, which then gives (7). *Remark:* we are using the fact that the zero function in a reproducing kernel space is point-wise zero since the evaluation functionals are bounded. □

There are a few things things to note here. First, the Euler-Lagrange equation (7) only implicitly characterizes $\hat v$ since it appears on both sides of the equality ($\beta_{k,t}$ and $X_{k,t}$ also depend on $\hat v$). Regardless, (7) is useful since it implies that $\hat v$ must lie within a known $n \times d$ dimensional sub-space of $V^{[0,1]}$. In particular, as discussed at the end of Section 3, the estimate $\{\hat v_t\}_{t\in[0,1]}$ is completely characterized by it's value at time $t = 0$, i.e. $\hat v_0$ (by the geodesic nature of $\hat v$). Restricting equation (7) to $t = 0$ one obtains

$$\hat v_0(x) = \frac{1}{\lambda n}\sum_{k=1}^n \beta_{k,0}^T R(x, X_k) + \frac{1}{\lambda n}\sum_{k=1}^n \nabla_y^T R(x,y)\Big|_{y=X_k}. \tag{13}$$

Simply stated, $\hat v_0$ has a finite dimensional spline characterization with spline knots set at the observations $X_1, \ldots, X_n$. Therefore to recover $\{\hat v_t\}_{t\in[0,1]}$ one simply needs to find the n row vectors $\beta_{1,0}, \ldots, \beta_{n,0}$ which satisfy the following fixed point equation

$$\beta_{k,0} = \nabla H(\phi_1^{\hat v}(X_k)) D\phi_1^{\hat v}(X_k) + \nabla \log\det D\phi_1^{\hat v}(X_k) \tag{14}$$

for all $k = 1, \ldots, n$.

5. Connection to Stein's Method

The Euler-Lagrange equation given in (7) has a surprising connection with a generalization of Stein's lemma for characterizing the normal distribution (see [22]). The main connection is that the Euler Lagrange equation for the PMLE estimate $\hat v_t$, simplified at initial time $t = 0$ and terminal time $t = 1$, can be reinterpreted as an empirical version of a generalization of Stein's lemma. This is interesting in it's own right, however, the connection may also bear theoretical fruit for deriving asymptotic estimation bounds on the nonparametric and semiparametric estimates derived from $\hat v$. In this section we make this connection explicit with the goal of of motivating and explaining the Euler-Lagrange equation for $\hat v$ derived above.

To relate $\hat v_t$ at $t = 0$ with Stein's lemma, and more generally Stein's method for distributional approximation, first notice that (14) implies the coefficients $\beta_{k,0}$, from the implicit equation (13) for $\hat v$, satisfy $\beta_{k,0} = \nabla \log \hat f(X_k)$ where $\hat f = e^{H\circ\phi_1^{\hat v}} |D\phi_1^{\hat v}|$ is the estimated density of X using the pullback of the target measure with the estimated diffeomorphisms $\phi_1^{\hat v}$. Now by computing the inner product of both sides of the Euler-Lagrange equation (13) with any vector field $u \in V$ and applying the reproducing property of the kernel $R(\cdot,\cdot)$ one derives

$$\lambda \langle \hat v_0, u\rangle_V = \mathbb{E}_n\big\{\nabla \log \hat f(X) \cdot u(X) + \operatorname{div} u(X)\big\}. \tag{15}$$

where $\mathbb{E}_n$ denotes expectation with respect to the empirical measure generated by the data: $\frac{1}{n}\sum_{k=1}^n \delta_{X_k}$. To relate with Stein first let $\mathbb{E}$ denote expectation with respect to the population density $f = e^{H\circ\phi}|D\phi|$ given in our basic model (2). Notice that a generalization of Stein's lemma shows that if the densities f and $\hat{f}$ give rise to the same probability measure then

$$0 = \mathbb{E}\big\{\nabla \log \hat{f}(X)\cdot u(X) + \operatorname{div} u(X)\big\} \tag{16}$$

for all u in a large class of test functions $\mathcal{U}$ (see Proposition 4 in [22]). For example, a simple consequence of Lemma 2 in [21] implies that when $\hat{f}$ is the density of a d dimensional Gaussian distribution $\mathcal{N}_d(\hat{\mu},1)$ and $X\sim\mathcal{N}_d(\mu,1)$ then $\hat{\mu}=\mu$ implies $\mathbb{E}\big\{-(X-\hat{\mu})\cdot u(X)+\operatorname{div}u(X)\big\}=0$ for any bounded function $u:\mathbb{R}^d\to\mathbb{R}^d$ with bounded gradient. Stein's method, on the other hand, generally refers to a technique for bounding the distance between two probability measures f and $\hat{f}$ using bounds on departures from a characterizing equation, such as (16) for example (see [9] for an exposition). The bounds typically take the form

$$\sup_{h\in\mathcal{H}}\left|\int (hf - h\hat{f})\right| \le \sup_{u\in\mathcal{U}}\big|\mathbb{E}\big\{\nabla\log\hat{f}(X)\cdot u(X)+\operatorname{div}u(X)\big\}\big| \tag{17}$$

where $\mathcal{H}$ and $\mathcal{U}$ are two class of functions related through a set of differential equations. In our case, applying a Hölder's inequality to the Euler-Lagrange equation (15) gives a bound on right hand side of (17) in terms of a regularization measurement on the PMLE $\hat{v}$ and an empirical process error:

$$\sup_{u\in\mathcal{U}}\big|\mathbb{E}\big\{\nabla\log\hat{f}(X)\cdot u(X)+\operatorname{div}u(X)\big\}\big| \le \underbrace{\lambda\|\hat{v}_0\|_V \sup_{u\in\mathcal{U}}\|u\|_V}_{\text{regularization at } t=0} + \underbrace{\sup_{u\in\mathcal{U}}\big|(\mathbb{E}-\mathbb{E}_n)\nu_{\hat{f},u}\big|}_{\text{empirical process error}} \tag{18}$$

where $\nu_{\hat{f},u} = \nabla\log\hat{f}(X)u(X)+\operatorname{div}u(X)$. This makes it clear that theoretical control of the PMLE estimate $\hat{v}_t$ at time $t=0$, using the Euler-Lagrange equation characterization (13), allows asymptotic control of the distance between the estimated density $\hat{f}$ and the true density f.

At terminal time $t=1$, there is a similar connection with Stein's lemma. In contrast to time $t=0$, which quantifies the distance between the estimated and population densities $\hat{f}$ and f, time $t=1$ quantifies the distance between $\phi(X)$ (the target measure) with $\phi_1^{\hat{v}}(X)$ (the push forward of the true population distribution though the estimated map). To make the connection, one follows the same line of argument as above to find that for any $u\in V$

$$\lambda\langle\hat{v}_1,u\rangle_V = \mathbb{E}_n^{\hat{v}}\big\{\nabla H(X)\cdot u(X)+\operatorname{div}u(X)\big\} \tag{19}$$

where $\mathbb{E}_n^{\hat{v}}$ denotes expectation with respect to the empirical measure $\frac{1}{n}\sum_{i=1}^n\delta_{\phi_1^{\hat{v}}(X_k)}$, which is simply the push forward of the empirical measure $\frac{1}{n}\sum_{i=1}^n\delta_{X_k}$ through the estimated map $\phi_1^{\hat{v}}$. Now the analog to (18) becomes

$$\sup_{u\in\mathcal{U}}\big|\mathbb{E}^{\hat{v}}\big\{\nabla H(X)\cdot u(X)+\operatorname{div}u(X)\big\}\big| \le \underbrace{\lambda\|\hat{v}_1\|_V\sup_{u\in\mathcal{U}}\|u\|_V}_{\text{regularization at } t=1} + \sup_{u\in\mathcal{U}}\big|(\mathbb{E}^{\hat{v}}-\mathbb{E}_n^{\hat{v}})\gamma_u\big| \tag{20}$$

where $\gamma_u = \nabla H(X)u(X)+\operatorname{div}u(X)$ and $\mathbb{E}^{\hat{v}}$ denotes expectation with respect to the push forward of the population density $f=e^{H\circ\phi}|D\phi|$ though the estimated map $\phi_1^{\hat{v}}$. Since the target measure $\mathbb{P}$ is assumed to have density e^H, this bounds the distributional distance between $\phi_1^{\hat{v}}(X)$ and $\mathbb{P}$ when $X\sim\mathbb{P}\circ\phi$.

6. Nonparametric example

In this section we utilize the finite dimensional characterization of the PMLE $\hat{v}$ at time $t=0$, given in (13), to construct nonparametric density estimates of the form $\hat{f}=e^{H\circ\phi_1^{\hat{v}}}|D\phi_1^{\hat{v}}|$ from *iid* samples

$X_1, \ldots, X_n$. As was discussed in the introduction, so long as the target measure $\mathbb{P}$ is absolutely continuous, the assumption that $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P} \circ \phi$ encompasses all absolutely continuous measures. Since the class of diffeomorphisms $\{\phi_1^v \colon v \in V^{[0,1]}\}$ is nonparametric, the estimate $\hat{f} = e^{H \circ \phi_1^{\hat{v}}} |D\phi_1^{\hat{v}}|$ is inherently nonparametric regardless of the choice of target probability measure $\mathbb{P}$ (with density e^H). In effect, the choice of target $\mathbb{P}$ specifies a shrinkage direction for the nonparametric estimate: larger values of λ shrink $\hat{f}$ further toward the target $\mathbb{P}$. In this section we illustrate the nonparametric nature of the density estimate $\hat{f}$, whereas the next section explores semiparametric estimation with parametric models on the target $\mathbb{P}$. One key feature of our methodology is the use of the Euler-Lagrange equation (7) as a stopping criterion for a gradient based optimization algorithm for constructing $\hat{v}$. In fact, to avoid computational challenges associated with generating geodesics with initial velocities given by (13), we consider a finite dimensional subclass of $V^{[0,1]}$ which have geodesics that are amenable to computation (and for which gradients are easy to compute). The key is that we use the Euler-Lagrange identity (7) to measure the richness of the subclass, within the larger infinite dimensional Hilbert space $V^{[0,1]}$, whereby allowing a dynamic choice of the approximating dimension for a target resolution level.

Claim 2 shows that the PMLE vector field $\hat{v} \in V^{[0,1]}$ obeys a parametric form determined up to the identification of the n functions $t \mapsto \beta_{k,t}$ as t ranges in $[0,1]$. Moreover, the whole path of coefficients $\beta_{k,t}$ is determined from the initial values $\beta_{k,0}$, by the geodesic nature of $\hat{v}$. In this way, we are free to optimize, over the vectors $\{\beta_{1,0}, \ldots, \beta_{n,0}\} \subset \mathbb{R}^d$ using equation (13) and are guaranteed that the global maximum, over the full infinite dimensional space $\{\phi_1^v \colon v \in V^{[0,1]}\}$, has this form. Unfortunately, deriving geodesic maps with this type of initial velocity field is challenging. To circumvent this difficulty we choose an approximating subclass of vector fields at time $t = 0$ which are parametrized by the selection of N knots $\{\kappa_1, \ldots, \kappa_N\} \subset \Omega$ and N initial momentum row vectors $\{\eta_1, \ldots, \eta_N\} \subset \mathbb{R}^d$ and have the form:

$$v_0(x) = \sum_{k=1}^{N} \eta_k^T R(x, \kappa_k). \tag{21}$$

The knots $\{\kappa_1, \ldots, \kappa_N\}$ need not be located at the data points $\{X_1, \ldots, X_n\}$. Indeed, we will see that alternative configurations of knots can be numerically beneficial. The key point is that vector fields at time $t = 0$, which satisfy (21), generate geodesics with respect to norm $\left[\int_0^1 \|v^t\|_V^2 dt\right]^{1/2}$ that are easy to compute. Moreover, the variational derivatives of the terminal map ϕ_1^v with respect to the initial η coefficients and the knots κ are easily computed when utilizing similar techniques as those developed in [25] and [1]. This enables efficient gradient based algorithms for optimizing the PMLE criterion over the class generated by (21).

As a first illustration, we show that the naïve choice of initial knots obtained by setting $\{\kappa_1, \ldots, \kappa_N\} = \{X_1, \ldots, X_n\}$ in (21) is *not* sufficient to solve (7); then show how it can be easily fixed using the Euler-Lagrange methodology. Our data set, shown with blue sticks in Figure 1, consists of $n = 10$ independent samples from a mixture of two normals, truncated so the support is $[0,1]$. Our target probability measure $\mathbb{P}$ is set to the uniform distribution on $[0,1]$ (smoothly tapering to zero 0 outside of $[0,1]$ for numerical convenience). For simplicity we choose the Gaussian kernel $R(x,y) = \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$, with $\sigma = 0.1$, to generate the RKHS V and use the penalty parameter λ set to 10. The top left plot in Figure 1 shows the non-parametric density estimate $\hat{f} = e^{H \circ \phi_1^{\hat{v}}} |D\phi_1^{\hat{v}}|$ in red, generated by applying a gradient based optimization algorithm applied to the subclass (21) where the knots $\{\kappa_1, \ldots, \kappa_N\} = \{X_1, \ldots, X_n\}$ are kept fixed and the coefficients $\eta_1, \ldots, \eta_N$ are optimized by minimizing the penalized log likelihood function $E_\lambda(v)$ given in (3). To diagnose the richness of subclass (21) within the full Hilbert space we define the function $\mathcal{D}_t^v(x)$ for any $v \in V^{[0,1]}$ and any $t \in [0,1]$ as follows

$$\mathcal{D}_t^v(x) \equiv \frac{1}{n} \sum_{k=1}^{n} \left[\beta_{k,t}^v\right]^T R(x, X_{k,t}^v) + \frac{1}{n} \sum_{k=1}^{n} \nabla_y^T R(x,y) \Big|_{y = X_{k,t}^v} \tag{22}$$

where $X_{k,t}^v \equiv \phi_t^v(X_k)$ and $\beta_{k,t}^v \equiv \nabla H(X_{k,1}^v) D\phi_{t1}^v(X_{k,t}^v) + \nabla \log\det D\phi_{t1}^v(X_{k,t}^v)$. The function $\lambda v_t - \mathcal{D}_t^v$ serves as a diagnostic criterion in the sense that the Hilbert norm of $\lambda v_t - \mathcal{D}_t^v$ gives the maximal

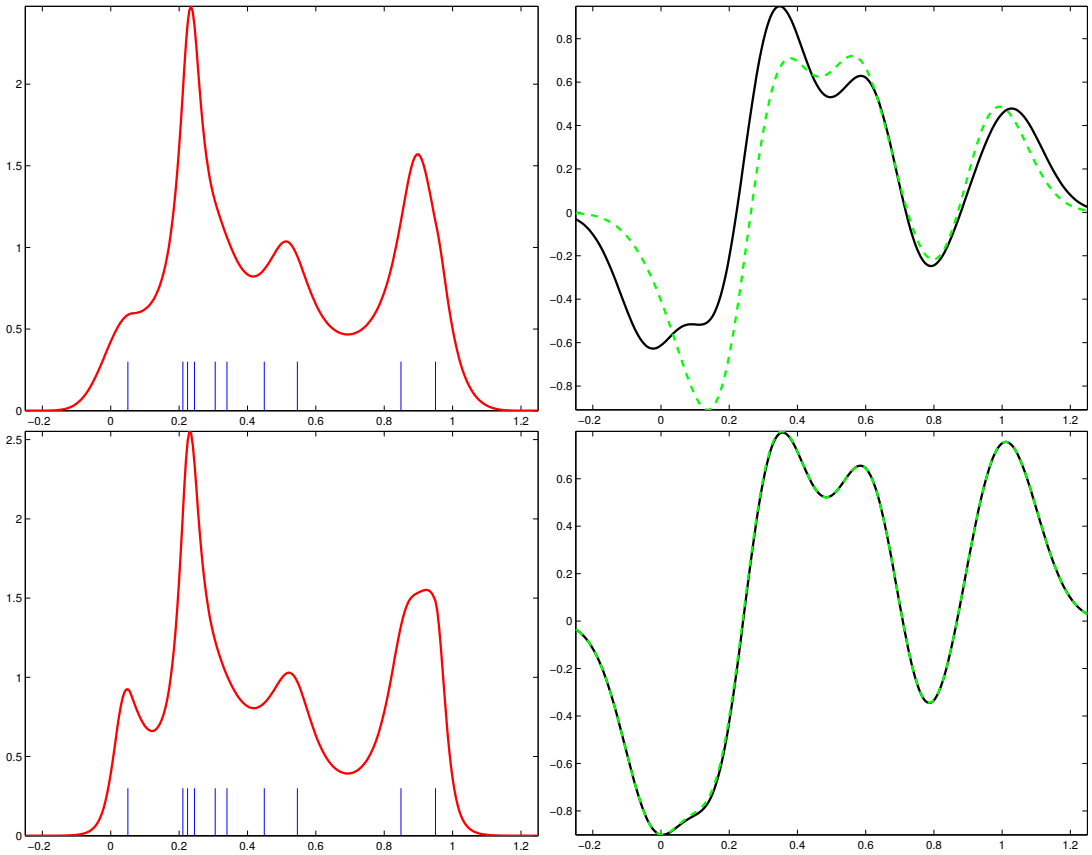

FIG 1. *In this example we compare two different knot configurations, in (21), for generating nonparametric density estimates using approximate solutions to the Euler-Lagrange equation (7). The left column of images shows two different density estimates (red), based on the same data set (blue), using two different knot configurations (top-left uses* 10 *knots, bottom-left uses* 30 *knots). The right column of images show the corresponding diagnostic curves which characterize the richness of the approximating subclass generated by the knots. The fact that the two diagnostic curves shown bottom-right are similar suggests that the* 30 *knots used generate the approximating subclass by (21) is sufficiently rich to reach the stationary points of the penalized log likelihood E_λ given in (3). See Section 6 for details.*

rate of change of the penalized log-likelihood $E_\lambda(v)$, within the full infinite dimensional Hilbert space $V^{[0,1]}$. In particular,

$$\left[\int_0^1 \| \underbrace{\lambda v_t - \mathcal{D}_t^v}_{\text{diagnostic}} \|_V^2 dt\right]^{1/2} = \sup_{\{u\colon \|u\|_{V^{[0,1]}}=1\}} \left[\frac{d}{d\epsilon} E_\lambda(v+\epsilon u)\right]_{\epsilon=0}.$$

Therefore if $\lambda v_t(x) - \mathcal{D}_t^v(x) = 0$ for all $t \in [0,1]$ and $x \in \mathbb{R}^d$, then v satisfies the Euler-Lagrange equation. Discrepancies between $\lambda v_t(x)$ and $\mathcal{D}_t^v$ when optimizing over the subclass (21) indicates the subclass that is insufficient rich to reach the stationary points of $E_\lambda(v)$. The diagnostic plots in this example, which correspond to our density estimate shown in the upper-left image of Figure 1, are shown in the upper-right plot of Figure 1 where $\lambda v_0(x)$ is plotted in black and $\mathcal{D}_0^v(x)$ is plotted as a dashed green line. The large amount of discrepancy between $\lambda v_0(x)$ and $\mathcal{D}_0^v(x)$ indicates that the knots $\{\kappa_1, \ldots, \kappa_N\} = \{X_1, \ldots, X_n\}$ are insufficient.

To generate knots which are sufficiently rich, in this first example, we apply a discrete approximation at initial time $t = 0$ to the gradient term $\nabla_y^T R(x,y)$ appearing in the Euler-Lagrange equation (7).

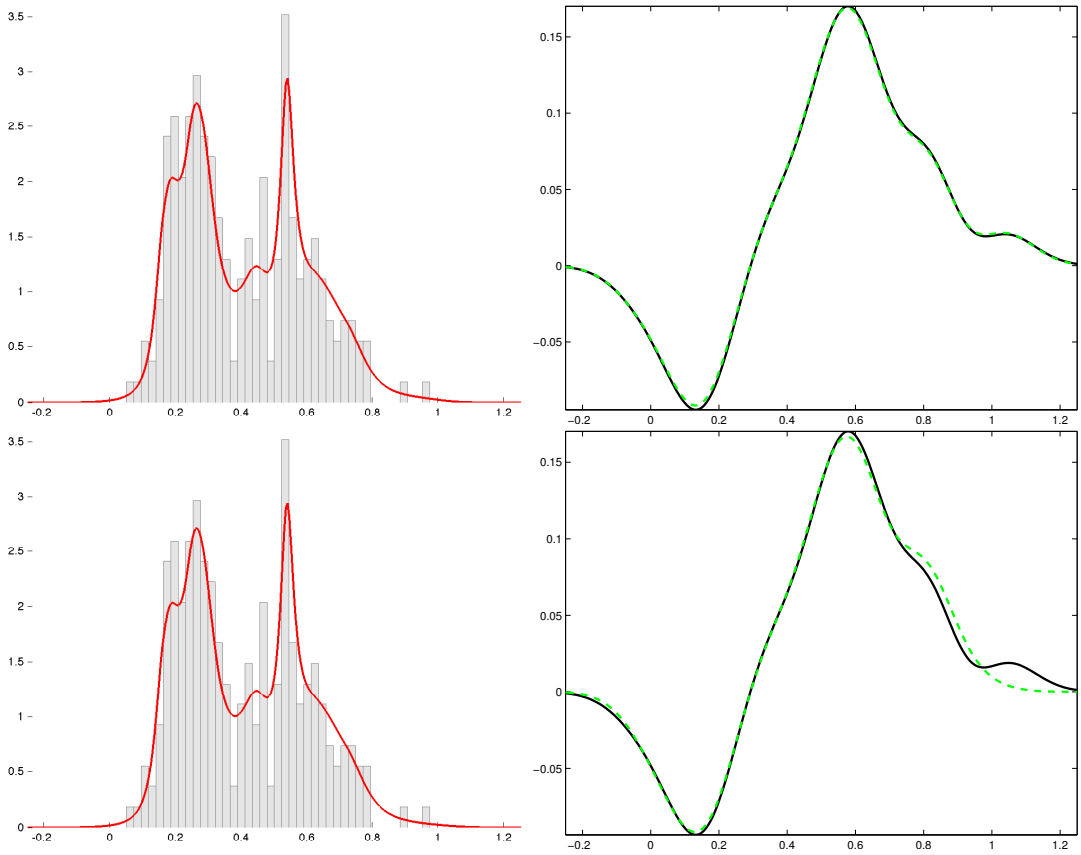

FIG 2. *In this example we demonstrate that a small number of knots, in (21), can be enough to approximate solutions to the Euler-Lagrange equation (7). The left column of images shows two different density estimates (red), based on the same data set of size* $n = 240$ *(grey histogram), using two different knot configurations (top-left uses* 240 *knots, bottom-left uses* 20 *knots). The right column of images show the corresponding diagnostic curves which characterize the richness of the approximating subclass generated by the knots. The fact that the two diagnostic curves shown bottom-right are nearly identical suggests that the* 20 *knots used generate the approximating subclass by (21) is sufficiently rich to reach the stationary points of the penalized log-likelihood* E_λ *given in (3). See Section 6 for details.*

For this approximation we use $N = 3n$ knots in the pattern given by the following approximation

$$\hat{v}_0(x) = \frac{1}{\lambda n}\sum_{k=1}^{n}\beta_{k,0}^T R(x, X_k) + \frac{1}{\lambda n}\sum_{k=1}^{n}\nabla_y^T R(x,y)\Big|_{y=X_k} \tag{23}$$

$$\approx \frac{1}{\lambda n}\sum_{k=1}^{n}\beta_{k,0}^T R(x, \kappa_k) + \frac{1}{\delta\lambda n}\sum_{k=1}^{n} R(x,\kappa_{n+k}) - R(x,\kappa_{2n+k}) \tag{24}$$

$$= \sum_{k=1}^{N}\eta_k^T R(x,\kappa_k) \tag{25}$$

where $\kappa_k \equiv \begin{cases} X_k & \text{if } k \in 1\dots n \\ X_k + \frac{\delta}{2} & \text{if } k \in n+1\dots 2n \\ X_k - \frac{\delta}{2} & \text{if } k \in 2n+1\dots 3n \end{cases}$ and $\eta_k \equiv \begin{cases} \frac{1}{\lambda n}\beta_{k,0} & \text{if } k \in 1\dots n \\ \frac{1}{\delta\lambda n} & \text{if } k \in n+1\dots 2n \\ -\frac{1}{\delta\lambda n} & \text{if } k \in 2n+1\dots 3n \end{cases}$ with $\delta = 10^{-4}$.

With this new set of knots, the resulting PMLE over the new class is show at bottom left in Figure 1. Notice that now the diagnostic function $\lambda v_0 - \mathcal{D}_0^v$ (the difference between the black and green line in the bottom-right plot of Figure 1) is much closer to zero. Indeed, for every $t \in [0,1]$ the diagnostic

function $\lambda v_t - \mathcal{D}_t^v$ is similarly close to zero (not pictured). This implies that the maximal rate of change of the penalized log-likelihood within the infinite dimensional Hilbert space $V^{[0,1]}$, at our estimate, is very small and hence our knots are sufficiently rich.

In the previous example we used $N = 3n$ knots in (21) to construct a sufficiently rich class for solving the Euler-Lagrange equation (7). Now we demonstrate that with larger data sets and smaller smoothness penalties one can actually use a smaller set of knots, $N \ll n$, to approximate the solutions to Euler-Lagrange equation (7). The histograms in the left column of Figure 2 show $n = 240$ *iid* samples from the same truncated mixture of normals used in the previous example. The resulting density estimates, shown in red, use a smoothness penalty set to $\lambda = 1/4$. The estimate shown top-left utilizes $N = n = 240$ knots set at the data points whereas the estimate shown bottom-left uses $N = 20$ knots randomly selected from the data. The right column shows the corresponding diagnostic plots (λv_0 shown in black and $\mathcal{D}_0^v$ shown in green). The relative agreement of the diagnostic curves in the bottom-right plot suggests that 20 knots are reasonably adequate for finding approximate solutions to the Euler-Lagrange equation. We expect this situation to improve as the number of data points increase. This has the potential to dramatically decrease the computational load when applying this estimate to extremely large data sets.

7. Semiparametric example

In this section we demonstrate how the PMLE $\phi_1^{\hat{v}}$ can be used to generate semiparametric estimation procedures obtained by assuming a parametric model on the target distribution $\mathbb{P}$ then introduce a nonparametric diffeomorphism to the target model. Indeed, any parametric model $\{\mathbb{P}_\theta \colon \theta \in \Theta \subset \mathbb{R}^m\}$ can be extended to a semiparametric class by considering diffeomorphisms of the data to the parametric target as follows: $\{\mathbb{P}_\theta \circ \phi_1^v \colon \theta \in \Theta, v \in V^{[0,1]}\}$. Since the model $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_\theta \circ \phi_1^v$ implies $\phi_1^v(X_1), \ldots, \phi_1^v(X_n) \overset{iid}{\sim} \mathbb{P}_\theta$ it is natural to alternate the optimization of θ and ϕ to compute the estimates $\hat{\theta}$ and $\hat{\phi}$ under this semiparametric model. This optimization routine is outlined explicitly in Algorithm 1.

Algorithm 1 Compute the semiparametric estimates $\hat{\theta}, \hat{\phi}$

1: Set $i = 0$ and initialize (θ^0, ϕ^0).

2: Set ϕ^{i+1} to the PMLE of ϕ defined in Section 3 under the model $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_{\theta^i} \circ \phi$.

3: Set θ^{i+1} to the maximum likelihood estimate of $\theta \in \Theta$ under the following model for the transformed data points:

$$\phi^{i+1}(X_1), \ldots, \phi^{i+1}(X_n) \overset{iid}{\sim} \mathbb{P}_\theta$$

4: If $\theta^i \approx \theta^{i+1}$ and $\phi^i \approx \phi^{i+1}$ then return $(\hat{\theta}, \hat{\phi}) \leftarrow (\theta^{i+1}, \phi^{i+1})$; else set $i \leftarrow i + 1$ and return to step 2.

To illustrate the semiparametric nature of our estimate we sample from a population density which is a mixture of a χ^2 density (with 20 degrees of freedom) and a Gaussian density ($\mu = 55$ and $\sigma = 3$) shown in black on the left column of plots in Figure 3. The data comprises $n = 200$ independent samples from this mixture, the histogram of which is shown on the left column of plots in Figure 3. We consider two different semiparametric estimates of the population density. The first uses a basic location-scale Gaussian family to the parametric target model $\{\mathbb{P}_\theta \colon \theta \in \Theta\} \equiv \{\mathbb{G}_{\mu,\sigma} \colon \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ where $\mathbb{G}_{\mu,\sigma}$ denotes the Gaussian measure centered at μ with variance σ^2. The second example uses a mixture of two Gaussian measures $\{\mathbb{P}_\theta \colon \theta \in \Theta\} \equiv \{\alpha\, \mathbb{G}_{\mu_1,\sigma_1} + (1-\alpha)\, \mathbb{G}_{\mu_2,\sigma_2} \colon \mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 \in \mathbb{R}^+, 0 < \alpha < 1\}$. As in Section 6 we use a Gaussian reproducing kernel to generate the Hilbert space V. In this example, however, we use a wider kernel, with standard deviation set to half the sample standard deviation of the data. This is done to illustrate the flexibility in the estimated density obtained by simply changing the kernel width and the penalty parameter λ (which is decreased to 1/2500 in this example). Wider kernels tend to produce estimates which have restricted local variability but can still have sufficient flexibility to model large amplitude variations over large spatial scales.

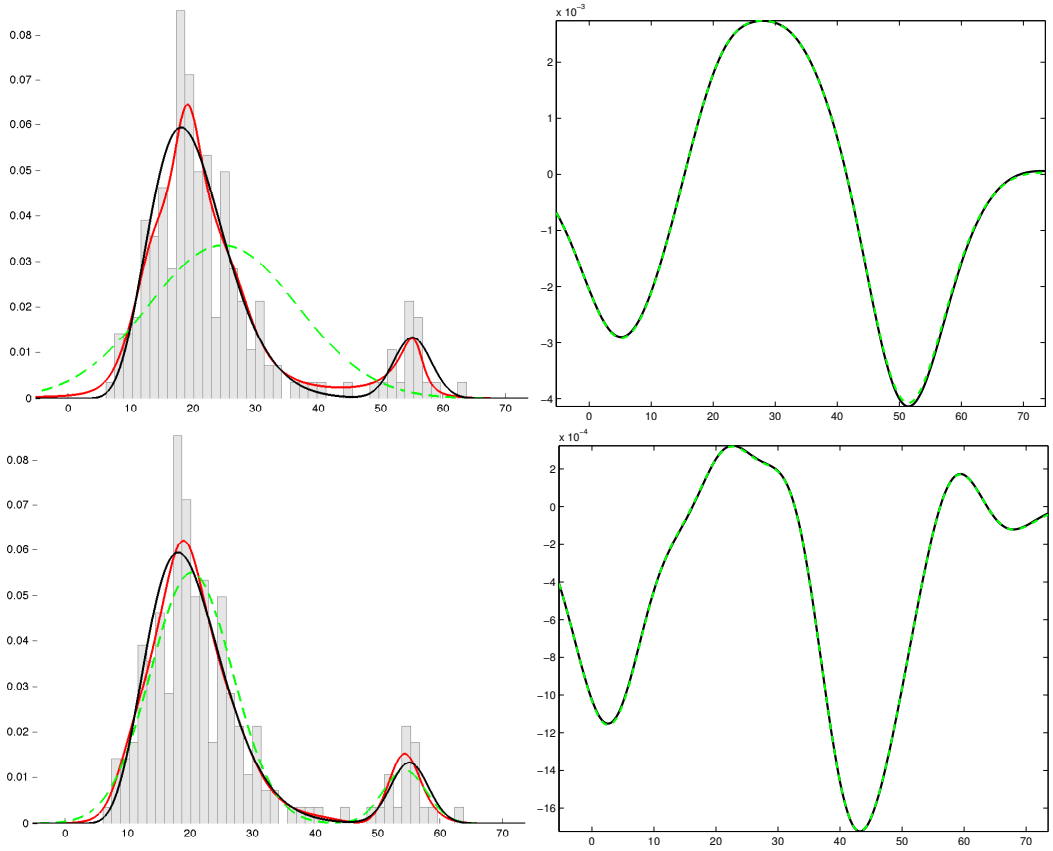

FIG 3. *In this example we demonstrate that by parametrically modeling the target distribution one can produce flexible semiparametric density estimates. The left column of histograms show the data (the same histogram plotted twice) sampled from the population density shown in black. Two semiparametric estimates are shown in red which correspond to different parametric targets. The estimated target distribution is shown in green on the left column of images. The right column of images show the corresponding diagnostic curves which characterize the richness of the approximating subclass generated by the knots (λv_0 is plotted in black and $\mathcal{D}_0^v$ is plotted in dashed-green). See Section 7 for details.*

The estimate obtained from the basic location-scale Gaussian family is shown in red in the top-left plot of Figure 3. Conversely, the estimated density which uses the mixture target model is shown in red in the bottom-left plot of Figure 3. The corresponding estimated target density $d\mathbb{P}_{\hat{\theta}}/dx$ is shown in green on the left two plots. To numerically approximate the PMLE initial velocity field $\hat{v}_0$, needed in step 2 of Algorithm 1, we used the approximating subclass for the initial velocity field given in the form (21) with 200 knots located at the data values. The corresponding time zero diagnostic plots are shown in the right column of Figure 3 (green for λv_0 and black for $\mathcal{D}_0^v$). Notice that in both cases the semiparametric estimates do a good job at estimating the population density. In the case of the location-scale Gaussian target the estimated target density does a poor job of explaining the true density. However, the presence of a nonparametric diffeomorphism allows this model to fit nearly as well as a fit from a mixture model. Notice also that the semiparametric estimate based on the location-scale Gaussian target overestimates the true sampling density between the two modes. This seems due to the fact that the estimation procedure prefers an overly dispersed target density which allows the estimated diffeomorphism to effectively add mass around the smaller mode. The situation seem to be corrected when using a mixture.

8. Discussion

In this paper, we adapt the powerful tools developed by Grenander, Miller, Younes, Trouvé and co-authors in the image processing and computational anatomy literature (see [27] and the references therein) to generate a PMLE of a deformation to a target measure with all the required properties: smoothness, invertability and computational tractability. The two main theoretical contributions of this paper are found in Claim 1 and Claim 2. Claim 1 establishes the existence of the diffeomorphism PMLE over a Hilbert space generating the initial velocity fields which give rise to the geodesic diffeomorphisms. Claim 2 shows that the PMLE has a finite dimensional spline characterization when the initial velocities are restricted to reproducing kernel Hilbert spaces. This finite dimensional characterization, although similar in spirit to spline function estimation, is completely different in that it holds for the initial velocity fields which generate the geodesic diffeomorphisms. A secondary contribution of this paper is the realization that the Euler-Lagrange equation for the PMLE also allows one to construct a diagnostic for approximating sub-models of the initial velocity field which are more amenable to computation. This diagnostic can be used to test whether a sub-model is sufficiently rich to reach the stationary points of the penalized maximum likelihood over the full infinite dimensional Hilbert space. This has the potential for significant computational savings when applying the estimate to large data sets in high dimension. In Section 5 we make an explicit connection between the Euler-Lagrange equation for the PMLE and Stein's method for bounding distances between two probability measures. This connection is used to motivate and explain the Euler-Lagrange equation and also hints at a new approach for deriving a theoretical understanding of the PMLE. The paper concludes with two illustrative examples which are not intended to be a compete simulation analysis but instead to illustrate the estimate and give a hint at the potential applicability of this new methodology. Indeed, the flexibility provided by both the diffeomorphism and the target measure make the methodology potentially applicable to a wide range of problems: from manifold estimation to density estimation on manifolds and from nonparametric measures of goodness of fit to heteroscedastic regression.

Appendix A: Technical Details

This section serves to present some technical details which are used in the proofs of Claim 1 and Claim 2. Some of these results can be found in the current literature. In particular, Proposition 1, most of Proposition 2 and equation (31) can be found in [27]. However, the main goal of this section is to establish equation (32) in Proposition 3 which is key to establishing Claim 2. We mention that all of the derivations presented in this section rely heavily on techniques developed by Younes, Tróuve, Miller and co-authors (see [27] and references therein).

To set notation let $C^k(\Omega, \mathbb{R}^d)$ denote the set of functions, mapping an open set $\Omega \subset \mathbb{R}^d$ into $\mathbb{R}^d$, which have continuous derivatives of order $\leq k$ (so that $C^0(\Omega, \mathbb{R}^d)$ is the continuous functions on Ω mapping into $\mathbb{R}^d$). Also let $C^k(\bar{\Omega}, \mathbb{R}^d)$ denote the set of functions in $C^k(\Omega, \mathbb{R}^d)$ whose derivatives of order $\leq k$ have continuous extensions to $\bar{\Omega}$. Finally, $C^k_0(\Omega, \mathbb{R}^d)$ is the set of functions in $C^k(\bar{\Omega}, \mathbb{R}^d)$ whose derivatives of order $\leq k$ take the value 0 on the boundary $\partial\Omega$. It is a well known fact that $C^k(\bar{\Omega}, \mathbb{R}^d)$ is a Banach space with respect to the norm: $\|f\|_{C^k(\bar{\Omega})} \equiv \|f\|_{k,\infty} \equiv \sum_{j=0}^k \sup_{|\beta|=j} \sup_\Omega |D^\beta f|$.

Remark: The norm $\|v\|_V \equiv \int_0^1 \|v_t\|_V^2 dt$ given in Definition 1 is technically only a semi-norm since one is free to change v_t on a set of $t \in [0,1]$ with Lebesque measure zero (and not effect the norm on $V^{[0,1]}$). This is easily fixed by identifying $V^{[0,1]}$ with the set of equivalence classes of measurable functions where $\{v_t\}_{t\in[0,1]}$ and $\{w_t\}_{t\in[0,1]}$ are said to be in the same equivalence class if $\|v_t - w_t\|_V = 0$ for almost every $t \in [0,1]$. For the remainder of the paper we treat this identification as implicit with the understanding that $\{v_t\}_{t\in[0,1]}$ denotes a representer of the equivalence class to which is belongs. The following proposition establishes the Hilbert space structure of $V^{[0,1]}$ (stated without proof in 8.17 of [27]).

Proposition 1. *If V is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$, then $V^{[0,1]}$ is a Hilbert space with inner product defined by $\langle v, h \rangle_{V^{[0,1]}} \equiv \int_0^1 \langle v_t, h_t \rangle_V dt$.*

Proof. Notice first that $\|v_t\|_V$ and $\langle v_t, h_t\rangle_V$ are measurable functions of t (this can be taken to be implicit in definitional requirement for membership in $V^{[0,1]}$: that $\int_0^1 \|v_t\|_V^2 dt < \infty$). Now, with the exception of completeness, all the properties of a Hilbert space inner product are inherited from $\langle\cdot,\cdot\rangle_V$ and the linear properties of Lebseque integration over $[0,1]$. To show completeness let $v^n \in V^{[0,1]}$ be a Cauchy sequence so that $\int_0^1 \|v_t^m - v_t^n\|_V^2 dt \to 0$. By the completeness of V there exists a Borel set $B \subset [0,1]$ such that for all $t \in B$ there exists a $v_t \in V$ such that $\|v_t^n - v_t\|_V \to 0$. On $t \in [0,1]\setminus B$ we are free to set $v_t \equiv 0$ (the zero element of V). For this v_t we have that $\|v_t^n - v_t\|_{V^{[0,1]}}^2 \equiv \int_0^1 \|v_t^n - v_t\|_V^2 dt \to 0$. Therefore $V^{[0,1]}$ is complete. □

Proposition 2. *Let Ω be an open bounded subset of $\mathbb{R}^d$ and V be a Hilbert space such that $V \hookrightarrow C_0^1(\Omega, \mathbb{R}^d)$. If $v \in V^{[0,1]}$ then there exists a unique class of C^1 diffeomorphisms of Ω, $\{\phi_t^v\}_{t\in[0,1]}$, such that $\phi_t^v(x) \in C^0([0,1]\times\overline{\Omega}, \mathbb{R}^d)$ and which satisfy the ordinary differential equation $\partial_t \phi_t^v(x) = v_t(\phi_t^v(x))$ with boundary condition $\phi_0^v(x) = x$, for all $x \in \Omega$. Moreover,*

$$\log\det D\phi_{st}^v(x) = \int_s^t \operatorname{div} v_u(\phi_{su}^v(x))du. \tag{26}$$

Proof. First note that if $v \in V^{[0,1]}$ and $V \hookrightarrow C_0^1(\Omega, \mathbb{R}^d)$ then $\|v_t\|_{1,\infty} \le c\|v_t\|_V$. Now by Hölder, $\int_0^1 \|v_t\|_V dt \le \|v\|_{V^{[0,1]}} < \infty$ so that the arguments for Theorem 8.7 in [27] to apply to the class $V^{[0,1]}$. In particular, there exists a unique class of C^1 diffeomorphisms of Ω, $\phi_t^v(x) \in C^0([0,1]\times\overline{\Omega}, \mathbb{R}^d)$, which satisfy the ordinary differential equation $\partial_t\phi_t^v(x) = v_t(\phi_t^v(x))$ with boundary condition $\phi_0^v(x) = x$, for all $x \in \Omega$. Moreover, by Proposition 8.8 in [27] we have that

$$\partial_t D\phi_{st}^v(x) = Dv_t(\phi_{st}^v(x))D\phi_{st}^v(x) \tag{27}$$

where $\det D\phi_{ss}^v(x) = Id_d$. Since $D\phi_{st}^v(x)$ is nonsingular and differentiable in t we have that (see (6.5.53) of [14], for example)

$$\begin{aligned}
\partial_t \log\det D\phi_{st}^v(x) &= \operatorname{trace}\big\{[D\phi_{st}^v(x)]^{-1}\partial_t D\phi_{st}^v(x)\big\} \\
&= \operatorname{trace}\big\{[D\phi_{st}^v(x)]^{-1}Dv_t(\phi_{st}^v(x))D\phi_{st}^v(x)\big\}, \text{ by (27)} \\
&= \operatorname{trace}\big\{Dv_t(\phi_{st}^v(x))\big\} \\
&= \operatorname{div} v_t(\phi_{st}^v(x)).
\end{aligned}$$

Therefore $\log\det D\phi_{st}^v(x)$ is differentiable everywhere on $t \in [0,1]$ with derivative given by $\operatorname{div} v_t(\phi_{st}^v(x))$.

Since $v_t(x)$ is measurable with respect to both arguments t and x (by definition) and limits of measurable functions are measurable, the function $\operatorname{div} v_t(x)$ is also measurable. Since $\phi_{st}^v(x)$ is continuous with respect to both t and x, $\operatorname{div} v_t(\phi_{st}^v(x))$ is also measurable. Notice that $\operatorname{div} v_t(\phi_{st}^v(x))$ is also Lebesque integrable since $|\operatorname{div} v_t(\phi_{st}^v(x))| \le c\|v_t\|_V$ by the embedding $V \hookrightarrow C_0^1(\Omega, \mathbb{R}^d)$ and the fact that that $\int_0^1 \|v_t\|_V dt \le \|v\|_{V^{[0,1]}} < \infty$. Therefore by Theorem 7.21 of [20] we have that

$$\log\det D\phi_{st}^v(x) = \int_s^t \operatorname{div} v_u(\phi_{su}^v(x))du$$

since $\log\det D\phi_{ss}^v(x) = 0$. □

Lemma 1. *If $V \hookrightarrow C_0^1(\Omega, \mathbb{R}^d)$ and $v, w \in V^{[0,1]}$, then*

$$\|\phi_{st}^v - \phi_{st}^w\|_\infty \le c\|v - w\|_{V^{[0,1]}} \exp\left(c\|v\|_{V^{[0,1]}}\right) \tag{28}$$

where c is a constant which does not depend on v, w, s or t. Moreover, if we additionally suppose $V \hookrightarrow C_0^2(\Omega, \mathbb{R}^d)$ then

$$\|\phi_{st}^v - \phi_{st}^w\|_{1,\infty} \le \|v - w\|_{V^{[0,1]}} F\big(\|v\|_{V^{[0,1]}}, \|w\|_{V^{[0,1]}}\big) \tag{29}$$

where $F(\cdot,\cdot)$ is a finite function on $\mathbb{R}\times\mathbb{R}$, monotonically increasing in both arguments, which does not depend on v, w, s or t.

Proof. The inequality (28) follows directly from Grownwell's lemma applied to the following inequality

$$\begin{aligned}|\phi^v_{st}(x) - \phi^w_{st}(x)| &= \left| \int_s^t v_u(\phi^v_{su}(x)) - w_u(\phi^w_{su}(x)) du \right| \\ &\leq \int_s^t |v_u(\phi^v_{su}(x)) - v_u(\phi^w_{su}(x))| du + \int_s^t |v_u(\phi^w_{su}(x)) - w_u(\phi^w_{su}(x))| du \\ &\leq \int_s^t c\|v_u\|_V |\phi^v_{su}(x) - \phi^w_{su}(x)| du + c\|v - w\|_{V^{[0,1]}}\end{aligned}$$

where the last inequality follows from the assumption $V \hookrightarrow C^1_0(\Omega, \mathbb{R}^d)$.

To prove (29) notice that for any vector $h \in \mathbb{R}^d$ we have that $\partial_t D\phi^v_{st}(x)h = Dv_t(\phi^v_{st}(x))D\phi^v_{st}(x)h$ where $D\phi^v_{ss}(x)h = h$ (by Proposition 8.8 in [27] and also [11]). Therefore

$$D\phi^v_{st}(x)h - D\phi^w_{st}(x)h = \int_s^t \left[Dv_u(\phi^v_{su}(x))D\phi^v_{su}(x)h - Dw_u(\phi^w_{su}(x))D\phi^w_{su}(x)h \right] du \tag{30}$$

where we are using the fact that $D\phi^v_{st}(x)h$ is differentiable with respect to t everywhere in $[0,1]$ and with Lebseque integrable derivative (and using Theorem 8.21 of [20]). Now notice that the integrand of (30) satisfies

$$\left| Dv_u(\phi^v_{su}(x))D\phi^v_{su}(x)h - Dw_u(\phi^w_{su}(x))D\phi^w_{su}(x)h \right| \leq I + II$$

where

$$I = \left| Dv_u(\phi^v_{su}(x)) \left\{ D\phi^v_{su}(x)h - D\phi^w_{su}(x)h \right\} \right| \leq c\left\|v_u\right\|_V \left| D\phi^v_{su}(x)h - D\phi^w_{su}(x)h \right|$$

and

$$\begin{aligned} II &= \left| \left\{ Dv_u(\phi^v_{su}(x)) - Dw_u(\phi^w_{su}(x)) \right\} D\phi^w_{su}(x)h \right| \\ &\leq \left\{ \|v_u\|_{2,\infty} \|\phi^v_{su} - \phi^w_{su}\|_\infty + \left\|v_u - w_u\right\|_{1,\infty} \right\} \left\|\phi^w_{su}\right\|_{1,\infty} |h| \\ &\leq \left\{ c\|v_u\|_V \, \|v - w\|_{V^{[0,1]}} \exp\left(c\|w\|_{V^{[0,1]}}\right) + c\left\|v_u - w_u\right\|_V \right\} \left\|\phi^w_{su}\right\|_{1,\infty} |h| \end{aligned}$$

where the last inequality follows by (28). To bound II further notice $\|\phi^v_{st}\|_{1,\infty} \leq c_1 \exp\left(c\|v\|_{V^{[0,1]}}\right)$. To see why, apply Gronwall's lemma to the following inequality

$$\begin{aligned}|\phi^v_{st}(x) - \phi^v_{st}(y)| &= \left| x - y + \int_s^t v_u(\phi^v_{su}(x)) - v_u(\phi^v_{su}(y)) du \right| \\ &\leq |x - y| + \int_s^t c \, \|v_u\|_V \, |\phi^v_{su}(x) - \phi^v_{su}(y)| du\end{aligned}$$

which yields $|\phi^v_{st}(x) - \phi^v_{st}(y)| \leq |x - y| \exp\left(c\|v\|_{V^{[0,1]}}\right)$. Since $\phi^v_{st}(x)$ is differentiable with respect to x everywhere in Ω, for each multi-index β such that $|\beta| = 1$ there exists a direction $h^\beta \in \mathbb{R}^d$ (with $|h^\beta| = 1$) such that

$$|D^\beta \phi^v_{st}(x)| = \lim_{\epsilon \downarrow 0} \frac{|\phi^v_{st}(x + \epsilon h^\beta) - \phi^v_{st}(x)|}{\epsilon} \leq \exp\left(c\|v\|_{V^{[0,1]}}\right).$$

Combining the above inequality with the fact that $\|\phi^v_{st}\|_\infty \leq \sup_{x \in \Omega} |x|$ gives the desired inequality $\|\phi^v_{st}\|_{1,\infty} \leq c_1 \exp\left(c \, \|v\|_{V^{[0,1]}}\right)$. Applying this to II gives

$$II \leq \left\{ c\|v_u\|_V \|v - w\|_{V^{[0,1]}} \exp\left(c\|w\|_{V^{[0,1]}}\right) + c\left\|v_u - w_u\right\|_V \right\} c_1 \exp\left(c\|w\|_{V^{[0,1]}}\right) |h|$$

Therefore

$$\int_s^t II\, du \le c_1 c|h|\|v-w\|_{V^{[0,1]}}\Big\{\|v\|_{V^{[0,1]}}\exp\big(2c\|w\|_{V^{[0,1]}}\big)+\exp\big(c\|w\|_{V^{[0,1]}}\big)\Big\}$$
$$= |h|\|v-w\|_{V^{[0,1]}}F\left(\|v\|_{V^{[0,1]}},\|w\|_{V^{[0,1]}}\right)$$

where $F(x,y)$ is monotone and finite in both x and y. Now by equation (30) we have that

$$\begin{aligned}\left|D\phi^v_{st}(x)h - D\phi^w_{st}(x)h\right| &\le \int_s^t I du + \int_s^t II du\\ &\le \int_s^t c\left\|v_u\right\|_V\left|D\phi^v_{su}(x)h - D\phi^w_{su}(x)h\right|du\\ &\qquad + |h|\|v-w\|_{V^{[0,1]}}F\big(\|v\|_{V^{[0,1]}},\|w\|_{V^{[0,1]}}\big).\end{aligned}$$

By Gronwell's lemma we have that

$$\left|D\phi^v_{st}(x)h - D\phi^w_{st}(x)h\right| \le |h|\|v-w\|_{V^{[0,1]}}F\big(\|v\|_{V^{[0,1]}},\|w\|_{V^{[0,1]}}\big)\exp\big(c\left\|v\right\|_{V^{[0,1]}}\big).$$

Now by taking a supremum over $x\in\Omega$, $|h|=1$ and combining with (28) gives (29), after redefining F to accommodate the extra term $\exp(c\left\|v\right\|_{V^{[0,1]}})$.

□

Proposition 3. *If $v,h\in V^{[0,1]}$ and $V\hookrightarrow C^1_0(\Omega,\mathbb{R}^d)$ then for all $x\in\Omega$ and $s,t\in[0,1]$*

$$\partial_\epsilon\phi^{v+\epsilon h}_{st}(x) = \int_s^t \left\{D\phi^{v+\epsilon h}_{ut}h_u\right\}\circ\phi^{v+\epsilon h}_{su}(x)\,du. \tag{31}$$

If, in addition, $V\hookrightarrow C^3_0(\Omega,\mathbb{R}^d)$ then

$$\partial_\epsilon \log\det D\phi^{v+\epsilon h}_1(x)\big|_{\epsilon=0} = \int_0^1\Big[h_u\cdot\nabla\log\det D\phi^v_{u1}+\operatorname{div}h_u\Big]\circ\phi^v_u(x)\,du \tag{32}$$

Proof. The assumption that $v,h\in V^{[0,1]}$ and $V\hookrightarrow C^1_0(\Omega,\mathbb{R}^d)$ are sufficient to apply Theorem 8.10 of [27] which gives

$$\partial_\epsilon\phi^{v+\epsilon h}_{st}(x)\big|_{\epsilon=0} = \int_s^t\left\{D\phi^v_{ut}h_u\right\}\circ\phi^v_{su}(x)\,du.$$

Now since $\partial_\epsilon\phi^{v+\epsilon h}_{st}(x) = \partial_\xi\phi^{v+\epsilon h+\xi h}_{st}(x)\big|_{\xi=0}$ one immediately obtains (31).

To show (32) notice that partial derivatives on x can pass under the integral in (31) to compute $D\partial_\epsilon\phi^{v+\epsilon h}_1$. This follows by first noticing that $V\hookrightarrow C^2_0(\Omega,\mathbb{R}^d)$ implies

$$\sup_{u\in[0,1]}\left\|\phi^{v+\epsilon h}_{ut}\right\|_{2,\infty}\le c_1\exp\left(c_2\|v\|_{V^{[0,1]}}+M\|h\|_{V^{[0,1]}}\right) \tag{33}$$

for all $|\epsilon|<M$, by equation (8.11) of [27]. Therefore when fixing $v,h\in V^{[0,1]}$ the function $\|D\phi^{v+\epsilon h}_{ut}h_u\|_{1,\infty}$ is bounded above by a finite constant over $(u,\epsilon)\in[0,1]\times(-M,M)$. With an additional application of Proposition 8.4 of [27] we also have that $\left\|\{D\phi^{v+\epsilon h}_{ut}h_u\}\circ\phi^{v+\epsilon h}_{su}\right\|_{1,\infty}<\infty$ uniformly over $(u,\epsilon)\in[0,1]\times(-M,M)$. Therefore, indeed, partial derivatives on x can pass under the integral in (31) to obtain

$$D\partial_\epsilon\phi^{v+\epsilon h}_1(x) = \int_0^1 D\big[\{D\phi^{v+\epsilon h}_{u1}h_u\}\circ\phi^{v+\epsilon h}_u(x)\big]du. \tag{34}$$

Now we show that $D\partial_\epsilon\phi^{v+\epsilon h}_1(x)$ is continuous over $(x,\epsilon)\in\Omega\times(-M,M)$. This will allow us to switch the order of D and ∂_ϵ and establish (32). The same reasoning which allows D to pass under the integral

in (31) also allows us to pass limits on x and ϵ under the integral in (34). Therefore it will be sufficient to show the integrand, $D\big[\{D\phi_{u1}^{v+\epsilon h}h_u\}\circ\phi_u^{v+\epsilon h}(x)\big]$, in (34) is continuous over $(x,\epsilon)\in\Omega\times(-M,M)$. To see why the integrand in (34) is continuous first note that $\phi_{st}^{v+\epsilon h}$ is a C^2 diffeomorphism (by a similar proof Theorem 8.7 in [27]). Secondly, under the assumption $V\hookrightarrow C_0^3(\Omega,\mathbb{R}^d)$ one can extend (29) to bound $\|\phi_{st}^{v+\epsilon h}-\phi_{st}^{v+\xi h}\|_{2,\infty}$ by $c|\xi-\epsilon|$, where c is a finite constant which may depend on v,h but not on ξ,ϵ. These two facts imply that $D\big[\{D\phi_{u1}^{v+\epsilon h}h_u\}\circ\phi_u^{v+\epsilon h}(x)\big]$ is indeed continuous in $(x,\epsilon)\in\Omega\times(-M,M)$ which implies that $D\partial_\epsilon\phi_1^{v+\epsilon h}(x)$ is also.

The continuity of $D\partial_\epsilon\phi_1^{v+\epsilon h}(x)$ over $(x,\epsilon)\in\Omega\times(-M,M)$ implies $\partial_\epsilon D\phi_{st}^{v+\epsilon h}$ exists and $D\partial_\epsilon\phi_{st}^{v+\epsilon h}=\partial_\epsilon D\phi_{st}^{v+\epsilon h}$ (see [10], page 56). Then since $D\phi_{st}^{v+\epsilon h}$ is nonsingular (by the diffeomorphic property) and differentiable with respect to ϵ we have that

$$\begin{aligned}\partial_\epsilon\log\det D\phi_{st}^{v+\epsilon h}(x)&=\operatorname{trace}\big\{[D\phi_{st}^{v+\epsilon h}(x)]^{-1}\partial_\epsilon D\phi_{st}^{v+\epsilon h}(x)\big\}\\&=\operatorname{trace}\big\{[D\phi_{st}^{v+\epsilon h}(x)]^{-1}D\partial_\epsilon\phi_{st}^{v+\epsilon h}(x)\big\}.\end{aligned}$$

Therefore, by (34),

$$\begin{aligned}\partial_\epsilon\log\det D\phi_1^{v+\epsilon h}(x)\big|_{\epsilon=0}&=\operatorname{trace}\left\{\big[D\phi_1^v(x)\big]^{-1}\int_0^1 D\big[\{D\phi_{u1}^v h_u\}\circ\phi_{1u}^v\circ\phi_1^v(x)\big]du\right\}\\&=\operatorname{trace}\left\{\big[D\phi_1^v(x)\big]^{-1}\int_0^1 D\big[\{D\phi_{u1}^v h_u\}\circ\phi_{1u}^v(y)\big]\Big|_{y=\phi_1^v(x)}du\,D\phi_1^v(x)\right\}\\&=\int_0^1\operatorname{trace}\left\{D\big[\{D\phi_{u1}^v h_u\}\circ\phi_{1u}^v(y)\big]\Big|_{y=\phi_1^v(x)}\right\}du.\end{aligned}$$

Now notice that

$$\begin{aligned}\operatorname{trace}D\big[\{D\phi_{u1}^v h_u\}\circ\phi_{1u}^v\big]&=\operatorname{trace}\big[\{D(D\phi_{u1}^v h_u)\}\circ\phi_{1u}^v D(\phi_{1u}^v)\big]\\&=\operatorname{trace}\big[\{D(D\phi_{u1}^v h_u)\}(D\phi_{u1}^v)^{-1}\big]\circ\phi_{1u}^v\\&=\big\langle h_u\circ\phi_{1u}^v,(\nabla\log\det D\phi_{u1}^v)\circ\phi_{1u}^v\big\rangle_d+(\operatorname{div}h_u)\circ\phi_{1u}^v.\end{aligned}$$

The last line follows from the identity: $\operatorname{trace}\big[\{D[D\phi_{u1}^v h_u]\}(D\phi_{u1}^v)^{-1}\big]=\big\langle h_u,\nabla\log\det D\phi_{u1}^v\big\rangle_d+\operatorname{trace}(Dh_u)$. Therefore

$$\partial_\epsilon\log\det D\phi_1^{v+\epsilon h}(x)\big|_{\epsilon=0}=\int_0^1\Big[h_u\cdot\nabla\log\det D\phi_{u1}^v+\operatorname{div}h_u\Big]\circ\phi_u^v(x)du.$$

□

References

[1] S. Allassonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

[2] E. Anderes and M. Coram. Two-dimensional density estimation using smooth invertible transformations. *Journal of Statistical Planning and Inference*, 141(3):1183 – 1193, 2011.

[3] E. Anderes, R. Huser, D. Nychka, and M. Coram. Nonstationary positive definite tapering on the plane. *Journal of Computational and Graphical Statistics*, to appear.

[4] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[5] M. Beg and A. Khan. Computing anb average anatomical atlas using lddmm and geodesic shooting. *Medical Image Analysis*, pages 1116–1119, 2006.

[6] M. Beg, M. Miller, A. Trouvé, and L. Younes. The euler-lagrange equation for interpolating sequence of landmark datasets. In Randy E. Ellis and Terry M. Peters, editors, *MICCAI (2)*, volume 2879 of *Lecture Notes in Computer Science*, pages 918–925. Springer, 2003.

[7] M. Beg, M. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005.

[8] Y. Cao, M. Miller, R. Winslow, and L. Younes. Large deformation diffeomorphic metric mapping of vector fields. *IEEE Transactions on Medical Imaging*, 24(9):1216–1230, 2005.

[9] L. Chen and Q. Shao. Stein's method for normal approximation. In *An introduction to Stein's method. . Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.*, volume 4. Singapore Univ. Press, Singapore, 2005.

[10] R. Courant. *Differential and Integral Calculus*, volume 2. Wiley, 1936.

[11] P. Dupuis, U. Grenander, and M. Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly of Applied Mathematics*, 56(3):587–600, 1998.

[12] Das et al. Detection of the power spectrum of cosmic microwave background lensing by the atacama cosmology telescope. *Physical Review Letters*, 107(2), 2011.

[13] U. Grenander and M. Miller. Computational anatomy: an emerging discipline. *Q. Appl. Math.*, LVI(4):617–694, December 1998.

[14] R. Horn and C. Johnson. *Topics in matrix analysis*. Cambridge University Press, New York, 1991.

[15] R. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2):309–323, 1995.

[16] M Miller, S Joshi, and G Christensen. Large deformation fluid diffeomorphisms for landmark and image matching. *A. Toga*, Brain Warping:115–132, 1999.

[17] M. Miller, A. Trouvé, and L. Younes. Geodesic shooting for computational anatomy. *Journal of Mathematical Imaging and Vision*, 24(2):209–228, 2006.

[18] M. Miller and L. Younes. Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41:61–84, 2001.

[19] T. El Moselhy and Y. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815 – 7850, 2012.

[20] W. Rudin. *Real and complex analysis*. McGraw-Hill, 1966.

[21] C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

[22] C. Stein, P. Diaconis, S. Holmes, and G. Reinert. Use of exchangeable pairs in the analysis of simulations. In *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser.*, pages 1–26. Inst. Math. Statist., Beachwood, OH, 2004.

[23] A Trouvé. Diffeomorphisms groups and pattern matching in image analysis. *International Journal of Computer Vision*, 28(3):213–221, 1998.

[24] A. Trouvé and Younes L. Local geometry of deformable templates. *SIAM J. on Mathematical Analysis*, 37(1):17–59, 2005.

[25] M. Vaillant, M. Miller, L. Younes, and A. Trouvé. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23, Supplement 1(0):S161 – S169, 2004.

[26] G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, PA., 1990.

[27] L. Younes. *Shapes and diffeomorphisms*. Springer, Heidelberg, 2010.

[28] L. Younes, A. Qiu, R. Winslow, and M. Miller. Transport of relational structures in groups of diffeomorphisms. *J. Math. Imaging Vis.*, 32(1):41–56, 2008.