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

Ethan Anderes\textsuperscript{1} and Marc Coram\textsuperscript{2}

\textsuperscript{1}Department of Statistics, University of California, Davis, CA 95616, USA  
\textit{e-mail: anderes@stat.ucdavis.edu}

\textsuperscript{2}Health Research and Policy Department, Stanford University, Palo Alto, CA 94304, USA

\textsuperscript{1}Supported in part by National Science Foundation grant DMS-1007480
\textsuperscript{2}Supported in part by National Institute of Health grant 5UL1 RR02574404
Abstract

A theorem of McCann [13] 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 $P$, and then reparameterize the whole class of absolutely continuous probability measures as monotone transformations from $P$. In this paper we utilize this reparameterization of densities, as monotone transformations from some $P$, 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 $P$ 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.

Key words and phrases: Density estimation, Euler-Lagrange, penalized maximum likelihood, diffeomorphism.

MSC 2010 Subject Classification: 62G07
1 Introduction

A theorem of McCann [13] 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 \( \mathbb{P} \). The advantage of this new viewpoint is the flexibility in choosing the target measure \( \mathbb{P} \) which can allow 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.

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 transformations which are smooth, invertible and which have attractive computational properties. Formally, we model our data \( X_1, \ldots, X_n \) as being generated by some diffeomorphism \( \phi: \mathbb{R}^d \rightarrow \mathbb{R}^d \) of some fixed but known probably distribution \( \mathbb{P} \) on \( \mathbb{R}^d \):

\[
X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P} \circ \phi.
\]  

(1)

The notation \( \mathbb{P} \circ \phi \) is taken to mean that the probability of \( X \in A \) is given by \( \mathbb{P}(\phi(A)) \) where \( \phi(A) = \{\phi(x): x \in A\} \). An important observation is that the model (1) implies that \( \phi(X) \sim \mathbb{P} \). Therefore, one can imagine estimating \( \phi \) by attempting to “deform” the data \( X_1, \ldots, X_n \) by a transformation which satisfies

\[
\phi(X_1), \ldots, \phi(X_n) \overset{iid}{\sim} \mathbb{P}.
\]

One of the main difficulties when working with such a model is the invertibility condition on the maps \( \phi \). The nonlinearity of this condition in \( \mathbb{R}^d \) when \( d > 1 \) makes constructing rich classes and optimizing over such classes difficult. One of the early attempts at circumventing such difficulties, found in [2], utilized the class of quasi-conformal maps to generate penalized maximum likelihood estimates of \( \phi \). However, these tools were only developed for \( \mathbb{R}^2 \) with no clear generalization for higher dimension. 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 [24] and the references therein) and apply them to the estimation of \( \phi \) from data \( X_1, \ldots, X_n \). Indeed, these techniques allow us to show that a penalized maximum likelihood estimate of \( \phi \) exists and that the solution has a finite dimensional characterization, similar to those results found in the spline literature (see [24]).

We start the paper in Section 2 with an overview of using the dynamics of time varying vector field flows to generate rich classes of diffeomorphisms. Then in sections 3 and 4 we define our penalized maximum likelihood estimate (PMLE) of \( \phi \) 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 \( \phi \) and a generalization of Stein’s lemma for characterizing the normal distribution (see [19]). 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.
A rich class of diffeomorphisms

In this section we give a brief overview of the technique of using the dynamics of time varying vector field flows to generate rich classes of diffeomorphisms. These time varying flows have been utilized with spectacular success in the fields of computational anatomy and image processing (see [1] [4] [5] [7] [11] [14] [15] [16] [20] [21] [22] [24] [25], and references therein). The tools developed in this body of work 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 $\phi: \Omega \rightarrow \mathbb{R}^d$ is said to be a $C^k(\Omega, \mathbb{R}^d)$ diffeomorphism of the open set $\Omega \subset \mathbb{R}^d$ if $\phi$ is one-to-one, maps onto $\Omega$, 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 $\Omega$ into $\mathbb{R}^d$. Under mild smoothness conditions there exists a unique class of diffeomorphisms of $\Omega$, denoted $\{\phi_t^v\}_{t \in [0,1]}$, which satisfy the following 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 $x \in \Omega$ (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 to 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 $\phi_t^s(x) = \phi_s(\phi_t^{-1}(x))$.

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 $\Omega$ 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 $v_t(x): [0,1] \times \Omega \rightarrow \mathbb{R}^d$ such that $v_t \in V$ for all $t \in [0,1]$ and $\int_0^1 \|v_t\|_V^2 \, dt < \infty$.

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 $\langle v, h \rangle_{V^{[0,1]}} = \int_0^1 \langle v_t, h_t \rangle_V \, dt$ (see Proposition 1 in the Appendix or Proposition 8.17 in [24]). 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 [24] 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

$$\|v\|_H \leq c \|v\|_V$$

for all $v \in V$ where $\| \cdot \|_H$ denotes the norm in $H$. Also we let $C^k_0(\Omega, \mathbb{R}^d)$ 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 ([24], [10]). If $V \hookrightarrow C^k_0(\Omega, \mathbb{R}^d)$, 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 (2) 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 $K(x, y): \Omega \times \Omega \rightarrow \mathbb{R}^{n \times n}$ 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 $\langle K(x, \cdot) p, f \rangle_V = f(x) \cdot p$.
for all $p \in \mathbb{R}^d$. 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 $t = 1$: $\phi^v_t$, 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 $\phi^v_t$ through Theorem 2.5 in [10].

3 Penalized maximum likelihood estimation

In this section we construct a penalized maximum likelihood estimate (PMLE) of $\hat{v}$ given the data $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P} \circ \phi^v_1$ where $\{\phi^v_t\}_{t \in [0,1]}$ satisfies (2) and $v \in V^{[0,1]}$. Under mild assumptions on $V$ and the density of $\mathbb{P}$ we prove the existence of a PMLE estimate of $\hat{v}$, whereby obtaining an estimate $\mathbb{P} \circ \phi^v_1$ of the true sampling distribution.

The target probability measure $\mathbb{P}$ is assumed known with 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 $\mathbb{P} \circ \phi^v_1$ has density given by

$$d\mathbb{P} \circ \phi^v_1(x) = \det(D\phi^v_1(x)) \exp H \circ \phi^v_1(x) dx$$

where $\det(D\phi^v_1(x))$ is defined as the determinant of the Jacobian of $\phi^v_1$ evaluated at $x \in \Omega$ (always positive by the orientation preserving nature of $\phi^v_t$). Since $\phi^v_t$ 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 $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P} \circ \phi^v_1$ is then given by

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

The estimated vector field $\hat{v}$ is chosen to be any element of $V^{[0,1]}$ which maximizes $E_\lambda$ 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 $C^2_0(\Omega, \mathbb{R}^d)$ where $\Omega$ is a bounded open subset of $\mathbb{R}^d$. Suppose $e^{H(\cdot)}$ is a bounded and continuous density on $\Omega$. Then there exists a time varying vector field $\hat{v} \in V^{[0,1]}$ such that

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

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

$$E_\lambda(v) = \underbrace{\frac{1}{n} \sum_{k=1}^n \log \det|D\phi^v_1(X_k)|}_{:= E_1(v)} + \underbrace{\frac{1}{n} \sum_{k=1}^n H \circ \phi^v_1(X_k)}_{:= E_2(v)} - \frac{\lambda}{2} \int_0^1 \|v_t\|_V^2 dt. \quad (5)$$

Notice that each term is well defined and finite whenever $v \in V^{[0,1]}$, since the assumption $V \to C^2_0(\Omega, \mathbb{R}^d)$ is sufficient for Theorem 8.7 in [24] 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 $\Omega$, $\{\phi^v_t\}_{t \in [0,1]}$, which satisfies (2) (also see Theorem 2.5 in [10]). 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
\( \log \det D\phi_t^v(x) = \int_0^1 \text{div } v_t(\phi_t^v(x))dt \) (by equation (26) in the Appendix). Therefore
\[
E_1(v) + E_3(v) = \frac{1}{n} \sum_{k=1}^n \int_0^1 \left( \text{div } v_t(\phi_t^v(X_k)) - \frac{\lambda}{2} \|v_t\|^2_V \right) dt
\]
\[
\leq \frac{1}{n} \sum_{k=1}^n \int_0^1 \left( \sup_{x \in \Omega} |\text{div } v_t(x)| - \frac{\lambda}{2} \|v_t\|^2_V \right) dt
\]
\[
\leq \int_0^1 \left( c\|v_t\|_V - \frac{\lambda}{2} \|v_t\|^2_V \right) dt, \text{ by the assumption } V \hookrightarrow C_0^2(\Omega, \mathbb{R}^d)
\]
\[
\leq \frac{c^2}{2\lambda} < \infty.
\]

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 \( \Omega \) is bounded, closed finite balls in \( V^{[0,1]} = L^2([0,1], V) \) are weakly compact (by [10]). Therefore we may extract a subsequence from \( v^m \) (relabelled 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
\[
\liminf_{m \to \infty} \|v^m\|_{V^{[0,1]}}^2 \geq \|\hat{v}\|_{V^{[0,1]}}^2.
\]

Now by Theorem 3.1 in [10] 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 \( \log \det D\phi_t^{v^m}(x) \to \log \det D\phi_t^\hat{v}(x) \) for every \( x \in \Omega \). To see why, one can use similar reasoning as in [7]. First write
\[
|\log \det D\phi_t^{v^m}(x) - \log \det D\phi_t^\hat{v}(x)| = \left| \int_0^1 \text{div } v_t^m(\phi_t^{v^m}(x)) - \text{div } \hat{v}_t(\phi_t^\hat{v}(x))dt \right| = I + II
\]
where the first term \( I \) satisfies
\[
I \equiv \left| \int_0^1 \text{div } v_t^m(\phi_t^{v^m}(x)) - \text{div } v_t^m(\phi_t^\hat{v}(x))dt \right|
\]
\[
\leq \int_0^1 \|\text{div } v_t^m\|_{1, \infty} |\phi_t^{v^m}(x) - \phi_t^\hat{v}(x)|dt
\]
\[
\leq \int_0^1 c\|v_t^m\|_V |\phi_t^{v^m}(x) - \phi_t^\hat{v}(x)|dt, \text{ since } V \hookrightarrow C_0^2(\Omega, \mathbb{R}^d)
\]
\[
\leq c\|v^m\|_{V^{[0,1]}} \left[ \int_0^1 |\phi_t^{v^m}(x) - \phi_t^\hat{v}(x)|^2 dt \right]^{1/2}, \text{ by Hölder.}
\]
\[
= o(1) \text{ uniformly in } t
\]
\[
\to 0, \text{ since } \|v^m\|_{V^{[0,1]}} \leq M \text{ for all } m.
\]

For the second term \( II \) notice that the map sending \( v \mapsto \int_0^1 \text{div } v_t(y_t)dt \) 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 \( \int_0^1 \text{div } v_t(y_t)dt = \langle v, w^\hat{v} \rangle_{V^{[0,1]}} \). Therefore
\[
II \equiv \left| \int_0^1 \text{div } v_t^m(\phi_t^\hat{v}(x)) - \text{div } \hat{v}_t(\phi_t^\hat{v}(x))dt \right|
\]
\[
= \left| \langle v^m - \hat{v}, w^\hat{v} \rangle_{V^{[0,1]} \to V^{[0,1]}} \right| \to 0, \text{ by weak convergence.}
\]
Combining the results for $I$ and $II$ we can conclude that $\log \det D\phi_t^m(x) \xrightarrow{m \to \infty} \log \det D\phi_t^0(x)$ for every $x \in \Omega$.

To finish the proof notice that

$$\sup_{v \in V^{[0,1]}} E_\lambda(v) = \lim_{m \to \infty} E(v^m) = \limsup_{m \to \infty} E(v^m)$$

$$= \frac{1}{n} \sum_{k=1}^{n} \log \det D\phi_t^0(X_k) + \frac{1}{n} \sum_{k=1}^{n} H \circ \phi_t^0(X_k) - \frac{\lambda}{2} \lim_{m \to \infty} \int_0^1 \|v_t^m\|_V^2 \, dt$$

$$\leq \frac{1}{n} \sum_{k=1}^{n} \log \det D\phi_t^0(X_k) + \frac{1}{n} \sum_{k=1}^{n} H \circ \phi_t^0(X_k) - \frac{\lambda}{2} \int_0^1 \|\hat{v}_t\|_V^2 \, dt, \quad \text{by } \Box$$

$$= E_\lambda(\hat{v})$$

One of the important facts about any vector field flow $\hat{v} \in V^{[0,1]}$ which maximizes $E_\lambda$ is that the resulting estimated transformation $\phi_t^\hat{v}$ 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, $\phi_t^v$, 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 $\phi_t^v = \phi_t^w$ but $v \neq w$. Notice, however, that the log-likelihood term in $E_\lambda$ only depends on $\phi_t^v$. This implies that any maximizer $\hat{v}$ of $E_\lambda$ 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. $\phi_1^v = \phi_1^w$. 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 $C_0^0(\Omega, \mathbb{R}^d)$ where $\Omega$ is bounded open subset of $\mathbb{R}^d$. Suppose $e^H(\cdot)$ is a $C^1(\Omega, \mathbb{R})$ density on $\Omega$. Then any time varying vector field $\hat{v} \in V^{[0,1]}$ which satisfies $[4]$ also satisfies the following Euler-Lagrange equation:

$$\hat{v}_t(x) = \frac{1}{\lambda n} \sum_{k=1}^{n} \beta^T_{k,t} R(x, X_{k,t}) + \frac{1}{\lambda n} \sum_{k=1}^{n} \nabla_y R(x, y) \big|_{y = X_{k,t}}$$

where $X_{k,t} \equiv \phi_t^\hat{v}(X_k)$ and $\beta_{k,t} \equiv \nabla H(X_{k,1})D\phi_t^\hat{v}(X_{k,t}) + \nabla \log \det D\phi_t^\hat{v}(X_{k,t})$.

**Proof.** Let $E_1, E_2$ and $E_3$ decompose $E_\lambda$ as in $[4]$. Notice first that if $h \in V^{[0,1]}$ and $\epsilon \in \mathbb{R}$ then $2E_3(\hat{v} \epsilon h) = \lambda \|\hat{v}\|_V^2 + \epsilon 2\lambda \|v\|_V^{[0,1]} + \epsilon^2 \lambda \|h\|_V^{[0,1]}$. Therefore $E_3(\hat{v} \epsilon h)$ is differentiable with respect to $\epsilon$ with derivative given by

$$\partial_\epsilon E_3(\hat{v} \epsilon h) \big|_{\epsilon = 0} = \int_0^1 \langle h_t, \lambda \hat{v}_t \rangle_V \, dt.$$
In addition, Theorem 8.10 of [24] implies that $\phi_{i}^{\epsilon+\epsilon h}(x)$ is differentiable at $\epsilon = 0$. Now, the assumption $H \in C^{1}(\Omega)$ combined with equation (31), in the Appendix, gives

\[
\partial_{\epsilon} E_2(\hat{v} + \epsilon h)|_{\epsilon = 0} = -\frac{1}{n} \sum_{k=1}^{n} \nabla H(\phi_{1}^{\epsilon}(X_k)) \cdot \partial_{\epsilon} \phi_{1}^{\epsilon+\epsilon h}(X_k)|_{\epsilon = 0}
\]

\[
= -\frac{1}{n} \sum_{k=1}^{n} \nabla H(\phi_{1}^{\epsilon}(X_k)) \cdot \int_{0}^{1} \{ D\phi_{u1}^{\epsilon} h_u \} \circ \phi_{u}^{\epsilon}(X_k) \, du
\]

\[
= -\frac{1}{n} \sum_{k=1}^{n} \int_{0}^{1} \{ \nabla H(X_k,1)D\phi_{u1}^{\epsilon}(X_k) \cdot h_u(X_k,u) \}
\]

\[
= \int_{0}^{1} \langle h_u(\cdot), -\frac{1}{n} \sum_{k=1}^{n} \{ \nabla H(X_k,1)D\phi_{u1}^{\epsilon}(X_k) \}^T R(\cdot, X_k,u) \rangle_{V \to \mathbb{R}^d} \, du
\]

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

\[
\partial_{\epsilon} E_1(\hat{v} + \epsilon h)|_{\epsilon = 0} = -\frac{1}{n} \sum_{k=1}^{n} \partial_{\epsilon} \log \det D\phi_{1}^{\epsilon+\epsilon h}(X_k)|_{\epsilon = 0}
\]

\[
= -\frac{1}{n} \sum_{k=1}^{n} \int_{0}^{1} \{ h_u \cdot \nabla \log \det D\phi_{u1}^{\epsilon} + \text{div } h_u \} \circ \phi_{u}^{\epsilon}(X_k) \, du
\]

\[
= \int_{0}^{1} \langle h_u(\cdot), -\frac{1}{n} \sum_{k=1}^{n} \{ \nabla \log \det D\phi_{u1}^{\epsilon}(X_k) \}^T R(\cdot, X_k,u) + \nabla_y R(\cdot,y) |_{y=X_k,u} \rangle_{V \to \mathbb{R}^d} \, du
\]

Remark: the above equation requires \( \partial_{x_i} (e_i \cdot h_u(x)) = \partial_{x_i} (e_i R(\cdot, x), h_u) \rangle_{V} = \langle e_i \partial_{x_i} R(\cdot, x), h_u \rangle_{V} \) which follows since \( \text{div } h_u \in V \) by the assumption $V \hookrightarrow C_0^2(\Omega, \mathbb{R}^d)$ (see [24]). Now from (8), (9) and (10), the energy $E_3(\hat{v} + \epsilon h)$ is differentiable with respect to $\epsilon$ at 0 and

\[0 = \partial_{\epsilon} E_3(\hat{v} + \epsilon h)|_{\epsilon = 0} = \langle \mathcal{E}_{0}^{\hat{v}}, \epsilon h \rangle_{V^{[0,1]}}\]

where

\[
\mathcal{E}_{0}^{\hat{v}} = \lambda \hat{v} - \frac{1}{n} \sum_{k=1}^{n} \beta_{k,t} R(\cdot, X_k,t) - \frac{1}{n} \sum_{k=1}^{n} \nabla_y R(\cdot,y) |_{y=X_k,u}
\]

with $\beta_{k,t} = \nabla H(X_k,1)D\phi_{u1}^{\epsilon}(X_k) + \nabla \log \det D\phi_{u1}^{\epsilon}(X_k)$. Since $h \in V^{[0,1]}$ was arbitrary, equation (11) implies $\mathcal{E}_{0}^{\hat{v}} = 0$, 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 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

\[
\hat{v}_0(x) = \frac{1}{\lambda n} \sum_{k=1}^{n} \beta_{k,0} R(x, X_k) + \frac{1}{\lambda n} \sum_{k=1}^{n} \nabla_y R(x,y) |_{y=X_k,u}.
\]

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 $n$ vectors $\beta_{1,0}, \ldots, \beta_{n,0}$ which satisfy the following fixed point equation

\[
\beta_{k,0} = \nabla H(\phi_{1}^{\epsilon}(X_k)) D\phi_{1}^{\epsilon}(X_k) + \nabla \log \det D\phi_{1}^{\epsilon}(X_k)
\]
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 (see 19). The main connection is that the Euler Lagrange equation for the PMLE estimate $\hat{v}$, 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 its 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 motivating and explaining the Euler-Lagrange equation for $\hat{v}$ derived above.

To relate $\hat{v}$ 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 $\hat{f} = e^{H \phi_k} |D\phi_k|$ is the estimated density of $X$ using the pullback of the target measure with the estimated diffeomorphisms $\phi_k$. 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

$$\lambda(\hat{v}_0, u)_V = \mathbb{E}_n \{ \nabla \log \hat{f}(X) \cdot u(X) + \text{div} \ u(X) \}.$$  

(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 $f = e^{H \phi} |D\phi|$ given in our basic model (1). 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

$$0 = \mathbb{E} \{ \nabla \log \hat{f}(X) \cdot u(X) + \text{div} \ u(X) \}$$  

(16)

for all $u$ in a large class of test functions $U$ (see Proposition 4 in 19). For example, a simple consequence of Lemma 2 in 18 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 $\mathbb{E} \{-X \cdot u(X) + \text{div} \ u(X)\} = 0$ for any bounded function $u : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 8 for an exposition). The bounds typically take the form

$$\sup_{h \in H} \left| \int (hf - h\hat{f}) \right| \leq \sup_{u \in U} \mathbb{E} \{ \nabla \log \hat{f}(X) \cdot u(X) + \text{div} u(X) \}$$  

(17)

where $H$ and $U$ 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 (13) 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:

$$\sup_{u \in U} \mathbb{E} \{ \nabla \log \hat{f}(X) \cdot u(X) + \text{div} u(X) \} \leq \lambda \|\hat{v}_0\|_V \sup_{u \in U} \|u\|_V + \sup_{u \in U} \left( \mathbb{E} - \mathbb{E}_n \right) \nu_{f,u}$$  

(18)

where $\nu_{f,u} = \nabla \log \hat{f}(X)u(X) + \text{div} u(X)$. This makes it clear that theoretical control of the PMLE estimate $\hat{v}$ 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 $\phi^1(X)$ (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 $u \in V$

$$\lambda(\hat{v}, u)_V = \mathbb{E}_n^\hat{\phi} \{ \nabla H(X) \cdot u(X) + \operatorname{div} u(X) \}$$

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

$$\sup_{u \in \mathcal{U}} \left| \mathbb{E}^\hat{\phi} \{ \nabla H(X) \cdot u(X) + \operatorname{div} u(X) \} \right| \leq \lambda \| \hat{v} \|_V \sup_{u \in \mathcal{U}} \| u \|_V + \sup_{u \in \mathcal{U}} \left| (\mathbb{E}^\hat{\phi} - \mathbb{E}_n^\hat{\phi}) \gamma_u \right|$$

(20)

where $\gamma_u = \nabla H(X)u(X) + \operatorname{div} u(X)$ and $\mathbb{E}^\hat{\phi}$ denotes expectation with respect to the push forward of the population density $f = e^{H \circ \phi} |D\phi|$ though the estimated map $\phi_t^\hat{\phi}$. Since the target measure $\mathbb{P}$ is assumed to have density $e^H$, this bounds the distributional distance between $\phi_t^\hat{\phi}(X)$ 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 $\hat{f} = e^{H \circ \hat{\phi}^t} |D\hat{\phi}^t|$ 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 $\{ \phi_t^v : v \in V^{[0,1]} \}$ is nonparametric, the estimate $\hat{f} = e^{H \circ \phi^\hat{\phi}} |D\phi^\hat{\phi}|$ 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 $\lambda$ 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 of 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 $\{ \phi_t^v : v \in V^{[0,1]} \}$, 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:

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

(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
Figure 1: In this example we compare two different knot configurations, in (21), for generating non-parametric 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 $E_\lambda$ given in (3). See Section 6 for details.

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 $P$ 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 $R(x, y) = \exp(-\frac{(x-y)^2}{2\sigma^2})$, with $\sigma = 0.1$, to generate the RKHS $V$ and use the penalty parameter $\lambda$ set to 10. The top left plot in Figure 1 shows the non-parametric density estimate $\hat{f} = e^{H\phi^i} |D\phi^i|$ 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 \(D^\nu_t(x)\) for any \(v \in V^{[0,1]}\) and any \(t \in [0,1]\) as follows
\[
D^\nu_t(x) = \frac{1}{n} \sum_{k=1}^{n} \left[ \beta^\nu_{k,t} R(x, X^\nu_{k,t}) + \frac{1}{n} \sum_{k=1}^{n} \nabla_y R(x, y) \right]_{y=X^\nu_{k,t}}
\]
where \(X^\nu_{k,t} \equiv \phi^\nu_t(X_k)\) and \(\beta^\nu_{k,t} \equiv \nabla H(X^\nu_{k,t})D\phi^\nu_t(X^\nu_{k,t}) + \nabla \log \det D\phi^\nu_t(X^\nu_{k,t})\). The function \(\lambda v_t - D^\nu_t\) serves as a diagnostic criterion in the sense that the Hilbert norm of \(\lambda v_t - D^\nu_t\) gives the maximal rate of change of the penalized log-likelihood \(E_\lambda(v)\), within the full infinite dimensional Hilbert space \(V^{[0,1]}\). In particular,
\[
\left[ \int_0^1 \left\| \lambda v_t - D^\nu_t \right\|^2 dt \right]^{1/2} = \sup_{\{u: \|u\|_{V^{[0,1]}} = 1\}} \left[ \frac{d}{dt} E_\lambda(v + \epsilon u) \right]_{\epsilon=0}.
\]
Therefore if \(\lambda v_t - D^\nu_t(x) = 0\) 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 \(D^\nu_t\) 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 \(D^\nu_0(x)\) is plotted as a dashed green line. The large amount of discrepancy between \(\lambda v_0(x)\) and \(D^\nu_0(x)\) 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 \(\nabla_y R(x, y)\) appearing in the Euler-Lagrange equation (7). For this approximation we use \(N = 3n\) knots in the pattern given by the following approximation
\[
\hat{v}_0(x) = \frac{1}{\lambda n} \sum_{k=1}^{n} \beta^T_{k,0} R(x, X_k) + \frac{1}{\lambda n} \sum_{k=1}^{n} \nabla_y R(x, y) \big|_{y=X^\nu_{k,t}}
\]
\[
\approx \frac{1}{\lambda n} \sum_{k=1}^{n} \beta^T_{k,0} R(x, \kappa_k) + \frac{1}{\delta \lambda n} \sum_{k=1}^{n} R(x, \kappa_{n+k}) - R(x, \kappa_{2n+k})
\]
\[
= \sum_{k=1}^{N} \eta_k T R(x, \kappa_k)
\]
where \(\kappa_k \equiv \begin{cases} X_k & \text{if } k \in 1 \ldots n \\ X_k + \frac{\delta}{2} & \text{if } k \in n + 1 \ldots 2n \text{ and } \eta_k \equiv \begin{cases} \frac{1}{\lambda n} \beta_{k,0} & \text{if } k \in 1 \ldots n \\ \frac{1}{\delta \lambda n} & \text{if } k \in n + 1 \ldots 2n \text{ with } \delta = 10^{-4}. \end{cases} \end{cases} \]

With this new set of knots, the resulting PMLE over the new class is shown at bottom left in Figure 1. Notice that now the diagnostic function \(\lambda v_0 - D^\nu_0\) (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 \(\lambda v_t - D^\nu_t\) 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 \(n = 240\) \textit{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.
Figure 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 $E_{\lambda}$ given in (3). See Section 6 for details.

(\lambda v_0$ shown in black and $D_v^0$ 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 $\phi^\tilde{v}_1$ can be used to generate semiparametric estimation procedures obtained by assuming a parametric model on the target distribution $P$ then introduce a nonparametric diffeomorphism to the target model. Indeed, any parametric model $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^m\}$ can be extended to a semiparametric class by considering diffeomorphisms of the data to the parametric target as follows: $\{P_\theta \circ \phi^\tilde{v}_1 : \theta \in \Theta, v \in V^{[0,1]}\}$. Since the model $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta \circ \phi^\tilde{v}_1$ implies
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 (for example, Proposition 1 reflects most of Proposition 2 and equation 31 can be found in [24]). However, the main goal of this section is to
Figure 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 ($\lambda v_0$ is plotted in black and $D_v$ is plotted in dashed-green). See Section 7 for details.

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, Trouve, Miller and co-authors (see [24] 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 $\Omega$ mapping into $\mathbb{R}^d$). Also let $C^k(\Omega, \mathbb{R}^d)$ 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 4 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 reprenter of the equivalence class to which it belongs.
Therefore, \( V \) free to set Proof.

\[ \langle \cdot \rangle \]

all \( x \) satisfy the ordinary differential equation \( \phi \) with boundary condition \( \langle \cdot \rangle \).

Lemma 1. Let Proposition 2.

Proof. Let Proposition 1.

Proof.\[ \int_{0}^{1} \| v_{t} \| d t \leq \| v \|_{V[0,1]} < \infty \] so that the arguments for Theorem 8.7 in \[ 24 \] to apply to the class \( V[0,1] \). In particular, there exists a unique class of \( C^{1} \) diffeomorphisms of \( \Omega \), \( \{ \phi_{t} \}_{t \in [0,1]} \), such that \( \phi_{t}^{\circ} (x) \in C^{0}([0,1] \times \overline{\Omega}, \mathbb{R}^{d}) \) and which satisfy the ordinary differential equation \( \partial_{t} \phi_{t}^{\circ} (x) = v_{t} (\phi_{t}^{\circ} (x)) \) with boundary condition \( \phi_{0}^{\circ} (x) = x \), for all \( x \in \Omega \). Moreover,\[ \log \det D \phi_{t}^{\circ} (x) = \int_{0}^{1} \text{div} v_{u} (\phi_{su}^{\circ} (x)) du. \] (26)

Proof. First note that if \( v \in V[0,1] \) and \( V \hookrightarrow C_{0}^{0} (\Omega, \mathbb{R}^{d}) \) then \( \| v_{t} \|_{1, \infty} \leq c \| v \|_{V} \). Now by Hölder, \( \int_{0}^{1} \| v_{t} \|_{V} d t \leq \| v \|_{V[0,1]} < \infty \) so that the arguments for Theorem 8.7 in \[ 24 \] to apply to the class \( V[0,1] \). In particular, there exists a unique class of \( C^{1} \) diffeomorphisms of \( \Omega \), \( \phi_{t}^{\circ} (x) \in C^{0}([0,1] \times \overline{\Omega}, \mathbb{R}^{d}) \), which satisfy the ordinary differential equation \( \partial_{t} \phi_{t}^{\circ} (x) = v_{t} (\phi_{t}^{\circ} (x)) \) with boundary condition \( \phi_{0}^{\circ} (x) = x \), for all \( x \in \Omega \). Moreover, by Proposition 8.8 in \[ 24 \] we have that\[ \partial_{t} D \phi_{t}^{\circ} = D v_{t} (\phi_{t}^{\circ} (x)) D \phi_{t}^{\circ}. \] (27)

where \( \text{det} D \phi_{t}^{\circ} = \text{Id}_{d} \). Since \( D \phi_{t}^{\circ} \) is nonsingular and differentiable in \( t \) we have that (see (6.5.53) of \[ 12 \], for example)\[ \partial_{t} \log \det D \phi_{t}^{\circ} (x) = \text{trace} \left\{ \left[ D \phi_{t}^{\circ} (x) \right]^{-1} \partial_{t} D \phi_{t}^{\circ} (x) \right\} \]

\[ = \text{trace} \left\{ \left[ D \phi_{t}^{\circ} (x) \right]^{-1} D v_{t} (\phi_{t}^{\circ} (x)) D \phi_{t}^{\circ} (x) \right\}, \] by (27)

\[ = \text{trace} \left\{ D v_{t} (\phi_{t}^{\circ} (x)) \right\} \]

\[ = \text{div} v_{t} (\phi_{t}^{\circ} (x)). \]

Therefore \( \log \det D \phi_{t}^{\circ} (x) \) is differentiable everywhere on \( t \in [0, 1] \) with derivative given by \( \text{div} v_{t} (\phi_{t}^{\circ} (x)) \).

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 \( \text{div} v_{t} (x) \) is also measurable. Since \( \phi_{t}^{\circ} (x) \) is continuous with respect to both \( t \) and \( x \), \( \text{div} v_{t} (\phi_{t}^{\circ} (x)) \) is also measurable. Notice that \( \text{div} v_{t} (\phi_{t}^{\circ} (x)) \) is also Lebesque integrable since \( \| \text{div} v_{t} (\phi_{t}^{\circ} (x)) \| \leq c \| v_{t} \|_{V} \) by the embedding \( V \hookrightarrow C_{0}^{0} (\Omega, \mathbb{R}^{d}) \) and the fact that that \( \int_{0}^{1} \| v_{t} \|_{V} d t \leq \| v \|_{V[0,1]} < \infty \). Therefore by Theorem 7.21 of \[ 17 \] we have that\[ \log \det D \phi_{t}^{\circ} (x) = \int_{0}^{1} \text{div} v_{u} (\phi_{su}^{\circ} (x)) du \]

since \( \log \det D \phi_{t}^{\circ} (x) = 0 \).

\[ \text{Lemma 1.} \] If \( V \hookrightarrow C_{0}^{0} (\Omega, \mathbb{R}^{d}) \) and \( v, w \in V[0,1] \), then\[ \| \phi_{t}^{\circ} - \phi_{t}^{\circ} \| \leq c \| v - w \|_{V[0,1]} \exp \left( c \| v \|_{V[0,1]} \right) \] (28)
where \( c \) is a constant which does not depend on \( v, w, s \) or \( t \). Moreover, if we additionally suppose \( V \hookrightarrow C^1_0(\Omega, \mathbb{R}^d) \) then

\[
\|\phi^v_{st} - \phi^w_{su}\|_{1,\infty} \leq \|v - w\|_{v^{[0,1]}} F(\|v\|_{v^{[0,1]}}, \|w\|_{v^{[0,1]}}) \tag{29}
\]

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

\[
\frac{d}{dt} \phi^w_{su}(x) = -\phi^v_{st}(x) - \phi^w_{su}(x) + \phi^w_{su}(x) - w(u) \tag{30}
\]

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 \( \partial_t \phi^v_{st}(x)h = Du(v_{su}(x))D\phi^v_{st}(x)h \) where \( D\phi^v_{st}(x)h = h \) (by Proposition 8.8 in [24] and also [10]). Therefore

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

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 Lebesgue integrable derivative (and using Theorem 8.21 of [17]). Now notice that the integrand of (30) satisfies

\[
\left| Du(v_{su}(x))D\phi^v_{st}(x)h - Dw_{su}(x)D\phi^w_{su}(x)h \right| \leq I + II
\]

where

\[
I = \left| Du(v_{su}(x)) \right| \left\{ D\phi^v_{st}(x)h - D\phi^w_{su}(x)h \right\} \leq c \|v_{su}\|_{V} \left| D\phi^v_{st}(x)h - D\phi^w_{su}(x)h \right|
\]

and

\[
II = \left\{ Du(v_{su}(x)) - Dw_{su}(x) \right\} \left| D\phi^w_{su}(x)h \right|
\]

\[
\leq \left\{ \|v_{su}\|_{2,\infty} \|\phi^v_{st} - \phi^w_{su}\|_{\infty} + \|v_{su} - w_{su}\|_{1,\infty} \right\} \|\phi^w_{su}\|_{1,\infty} |h|
\]

\[
\leq \left\{ c \|v_{su}\|_{V} \|v - w\|_{v^{[0,1]}} \exp (c \|v\|_{v^{[0,1]}}) + c \|v_{su} - w_{su}\|_{V} \right\} \|\phi^w_{su}\|_{1,\infty} |h|
\]

where the last inequality follows by (28). To bound \( II \) further notice \( \|\phi^v_{st}\|_{1,\infty} \leq c_1 \exp (c \|v\|_{v^{[0,1]}}) \).

To see why, apply Gronwall’s lemma to the following inequality

\[
|\phi^v_{st}(x) - \phi^w_{st}(y)| = |x - y| + \int_y^t v_{su}(\phi^v_{su}(x)) - v_{su}(\phi^w_{su}(y)) du \leq |x - y| + \int_y^t c \|v_{su}\|_{V} |\phi^v_{su}(x) - \phi^w_{su}(y)| du
\]

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

\[
|D^\beta \phi^v_{st}(x)| = \lim_{\epsilon \downarrow 0} \frac{|\phi^v_{st}(x + \epsilon h^\beta) - \phi^v_{st}(x)|}{\epsilon} \leq \exp (c \|v\|_{v^{[0,1]}}).
\]
Combining the above inequality with the fact that \( \| \phi_{st}^v \|_\infty \leq c_1 \exp (c \| v \|_{V[0,1]}) \). Applying this to \( II \) gives

\[
II \leq \left\{ c \| u \|_V \| v - w \|_{V[0,1]} \exp (c \| w \|_{V[0,1])} + c \| u - w \|_V \right\} \]

Therefore

\[
\int_s^t II \, du \leq c_1 \| \phi \| V[0,1] \left\{ \| v \|_{V[0,1]} \exp (2c \| w \|_{V[0,1])} + \exp (c \| w \|_{V[0,1])} \right\}
\]

\[
= |h| \| v - w \|_{V[0,1]} F (\| v \|_{V[0,1]}, \| w \|_{V[0,1])}
\]

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

\[
|D\phi_{st}^v(x)h - D\phi_{st}^w(x)h| \leq \int_s^t Idu + \int_s^t II du
\]

\[
\leq \int_s^t c \| u \|_V \| D\phi_{st}^v(x)h - D\phi_{st}^w(x)h \| du
\]

\[
+ |h| \| v - w \|_{V[0,1]} F (\| v \|_{V[0,1]}, \| w \|_{V[0,1])} \).
\]

By Gronwall’s lemma we have that

\[
|D\phi_{st}^v(x)h - D\phi_{st}^w(x)h| \leq |h| \| v - w \|_{V[0,1]} F (\| v \|_{V[0,1]}, \| w \|_{V[0,1]} \}) \exp (c \| v \|_{V[0,1])}.
\]

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 \( \exp (c \| v \|_{V[0,1])} \).

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 \( s, t \in [0,1] \)

\[
\partial_s \phi_{st}^{v+ch} (x) = \int_s^t \{ D\phi_{st}^{v+ch} h_u \} \circ \phi_{su}^{v+ch} (x) du.
\]

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

\[
\partial_s \log det D\phi_{st}^{v+ch} (x) \bigg|_{t=0} = \int_0^1 \left[ h_u \cdot \nabla \log det D\phi_{su}^v + \text{div} h_u \right] \circ \phi_{su}^v (x) du
\]

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 (24) which gives

\[
\partial_s \phi_{st}^{v+ch} (x) \bigg|_{t=0} = \int_s^t \{ D\phi_{su}^v h_u \} \circ \phi_{su}^v (x) du.
\]

Now since \( \partial_s \phi_{st}^{v+ch} (x) = \partial_t \phi_{st}^{v+ch+th} (x) \bigg|_{t=0} \) one immediately obtains (31).

To show (32) notice that partial derivatives on \( x \) can pass under the integral in (31) to compute \( D\partial_t \phi_{st}^{v+ch} \). This follows by first noticing that \( V \hookrightarrow C^3_0(\Omega, \mathbb{R}^d) \) implies

\[
\sup_{u \in [0,1]} \| \phi_{st}^{v+ch} \|_V \leq c_1 \exp (c_2 \| v \|_{V[0,1]} + M \| h \|_{V[0,1]})
\]

for all \( |c| < M \), by equation (8.11) of (24). Therefore when fixing \( v, h \in V[0,1] \) the function \( \| D\phi_{st}^{v+ch} h_u \|_{V[0,1]} \) is bounded above by a finite constant over \( (u, c) \in [0,1] \times (-M, M) \). With an additional application of Proposition 8.4 of (24) we also have that \( \| \{ D\phi_{st}^{v+ch} h_u \} \circ \phi_{su}^{v+ch} \|_{V[0,1]} < \infty \) uniformly over
Now notice that the last line follows from the identity: 
\[ \text{trace}(D\phi_{x,\epsilon}^\nu(x)) = \int_0^1 D\{[D\phi_{x,\epsilon}^\nu(x)] \circ \phi_{x,\epsilon}^\nu(x)\} du. \] (34) 

Now we show that \( D\partial_\epsilon \phi_{x,\epsilon}^\nu(x) \) is continuous over \((x,\epsilon) \in \Omega \times (-M,M)\). This will allow us to switch the order of \( D \) and \( \partial_\epsilon \) 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 \( \epsilon \) under the integral in (34). Therefore it will be sufficient to show the integrand, \( D\{[D\phi_{x,\epsilon}^\nu(x)] \circ \phi_{x,\epsilon}^\nu(x)\} \), in (34) is continuous over \((x,\epsilon) \in \Omega \times (-M,M)\). To see why the integrand in (34) is continuous first note that \( \phi_{x,\epsilon}^\nu(x) \) is a \( C^2 \) diffeomorphism (by a similar proof Theorem 8.7 in [24]). Secondly, under the assumption \( V \to C^0_0(\Omega, \mathbb{R}^d) \) one can extend (29) to bound \( \|\phi_{x,\epsilon}^\nu(x) - \phi_{x,\epsilon}^\nu(y)\|_{2,\infty} \) by \( c|\xi - \epsilon| \), where \( c \) is a finite constant which may depend on \( v, h \) but not on \( \xi, \epsilon \). These two facts imply that \( D\{[D\phi_{x,\epsilon}^\nu(x)] \circ \phi_{x,\epsilon}^\nu(x)\} \) is indeed continuous in \((x,\epsilon) \in \Omega \times (-M,M)\) which implies that \( D\partial_\epsilon \phi_{x,\epsilon}^\nu(x) \) is also.

The continuity of \( D\partial_\epsilon \phi_{x,\epsilon}^\nu(x) \) over \((x,\epsilon) \in \Omega \times (-M,M)\) implies \( \partial_\epsilon \phi_{x,\epsilon}^\nu(x) \) exists and \( D\partial_\epsilon \phi_{x,\epsilon}^\nu(x) = \partial_\epsilon D\phi_{x,\epsilon}^\nu(x) \) (see [9], page 56). Then since \( D\phi_{x,\epsilon}^\nu(x) \) is nonsingular (by the diffeomorphic property) and differentiable with respect to \( \epsilon \) we have that

\[ \partial_\epsilon \log \det D\phi_{x,\epsilon}^\nu(x) = \text{trace}\{[D\phi_{x,\epsilon}^\nu(x)]^{-1} \partial_\epsilon \phi_{x,\epsilon}^\nu(x)\} = \text{trace}\{[D\phi_{x,\epsilon}^\nu(x)]^{-1} \partial_\epsilon \phi_{x,\epsilon}^\nu(x)\}. \]

Therefore, by (34),

\[
\partial_\epsilon \log \det D\phi_{x,\epsilon}^\nu(x) \big|_{\epsilon=0} = \text{trace}\left\{ [D\phi_{x,u}^\nu(x)]^{-1} \int_0^1 D\{[D\phi_{x,u}^\nu(x)] \circ \phi_{x,u}^\nu(x)\} du \right\}
\]

\[
= \text{trace}\left\{ [D\phi_{x,u}^\nu(x)]^{-1} \int_0^1 D\{[D\phi_{x,u}^\nu(x)] \circ \phi_{x,u}^\nu(y)\} \big|_{y=\phi_{x,u}^\nu(x)} du \right\}
\]

\[
= \int_0^1 \text{trace}\left\{ D\{[D\phi_{x,u}^\nu(x)] \circ \phi_{x,u}^\nu(y)\} \big|_{y=\phi_{x,u}^\nu(x)} \right\} du.
\]

Now notice that

\[
\text{trace} D\{[D\phi_{x,u}^\nu(x)] \circ \phi_{x,u}^\nu(x)\} = \text{trace} \{D(D\phi_{x,u}^\nu(x) \circ \phi_{x,u}^\nu(x))\}
\]

\[
= \text{trace} \{D(D\phi_{x,u}^\nu(x)) \circ (D\phi_{x,u}^\nu(x))^{-1} \circ \phi_{x,u}^\nu(x)\}
\]

\[
= \langle h_u \circ \phi_{x,u}^\nu, \nabla \log \det D\phi_{x,u}^\nu \rangle + (\text{div} h_u) \circ \phi_{x,u}^\nu.
\]

The last line follows from the identity: \( \text{trace} \{[D(D\phi_{x,u}^\nu(x)) \circ (D\phi_{x,u}^\nu(x))^{-1}] = \langle h_u, \nabla \log \det D\phi_{x,u}^\nu \rangle + \text{trace}(Dh_u) \). Therefore

\[
\partial_\epsilon \log \det D\phi_{x,\epsilon}^\nu(x) \big|_{\epsilon=0} = \int_0^1 \left[ h_u \cdot \nabla \log \det D\phi_{x,u}^\nu + \text{div} h_u \right] \circ \phi_{x,u}^\nu(x) du.
\]

\[ \square \]

References

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

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

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

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

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

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

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

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

[19] 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.

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

[21] A. Trouvé and Younes L. Local geometry of deformable templates. *SIAM J. on Mathematical Analysis*, 37(1):17–59, 2005.
[22] M. Vaillant, M. Miller, L. Younes, and A. Trouvé. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23, Supplement 1(0):S161 – S169, 2004.

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

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

[25] 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.