SIAM J. Imaging Sciences
Vol. 10, No. 2, pp. 782–807

Langevin Equations for Landmark Image Registration with Uncertainty*

Stephen Marsland† and Tony Shardlow‡

Abstract. Registration of images parameterized by landmarks provides a useful method of describing shape variations by computing the minimum-energy time-dependent deformation field that flows from one landmark set to the other. This is sometimes known as the geodesic interpolating spline and can be solved via a Hamiltonian boundary-value problem to give a diffeomorphic registration between images. However, small changes in the positions of the landmarks can produce large changes in the resulting diffeomorphism. We formulate a Langevin equation for looking at small random perturbations of this registration. The Langevin equation and three computationally convenient approximations are introduced and used as prior distributions. A Bayesian framework is then used to compute a posterior distribution for the registration, and also to formulate an average of multiple sets of landmarks.

Key words. image registration, landmarks, shape, Bayesian statistics, SDEs, Langevin equation

AMS subject classifications. 92C55, 82C31, 34A55

DOI. 10.1137/16M1079282

1. Introduction.

The mathematical description of shape and shape change has become an area of significant research interest in recent years, not least because of its applications in computational anatomy, where variations in the appearance of objects in medical images are described mathematically in the hope that their change can be linked to disease progression. When two images are topologically equivalent, they can be brought into alignment (registered) by deforming one of the images without tearing or folding, so that their appearance matches as closely as possible. This can be formulated mathematically by taking two images $I, J\colon B \to \mathbb{R}$ (for some physical domain $B \subset \mathbb{R}^d$) that act as reference and target, respectively. (In medical imaging, these are typically greyscale images.) Image I is then deformed by some diffeomorphism $\mathbf{\Phi}\colon B \to B$ such that $I \circ \mathbf{\Phi}^{-1}$ and J are as close as possible according to some model of similarity. In addition to defining similarity, we must also select the metric on the diffeomorphism group; the typical setting is to use the right-invariant H^1_α metric, which leads to the so-called EPDiff equation [13]. We can also define a "bending energy" of $\mathbf{\Phi}$ in analogy to the thin-plate spline [3, 6]. For a general treatment and an overview of the subject, see the monograph [30] and references therein.

Similarity can be understood as a norm on the images $\|I \circ \mathbf{\Phi}^{-1} - J\|$, in which case a common choice is the sum-of-squares of pixel values, although there are plenty of other options (see, e.g., [24]). Alternatively, similarity can be expressed by a set of landmarks that

*Received by the editors June 9, 2016; accepted for publication (in revised form) January 9, 2017; published electronically May 25, 2017.
http://www.siam.org/journals/siims/10-2/M107928.html
Funding: This work was partially supported by the LMS Scheme 7 grant SC7-1415-09.
†School of Engineering and Advanced Technology, Massey University, Palmerston North, 4442, New Zealand (s.r.marsland@massey.ac.nz).
‡Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (t.shardlow@bath.ac.uk).

identify corresponding points on each image. Our focus is on the second of these two methods. Specifically, we consider a set of landmarks on the reference and target images, $\boldsymbol{q}_i^r$ and $\boldsymbol{q}_i^t$ for $i = 1, \dots, N$, in B and we aim to find $\boldsymbol{\Phi}$ such that $\boldsymbol{\Phi}(\boldsymbol{q}_i^r) = \boldsymbol{q}_i^t$. Obviously, landmarks need to correspond between the images, and this is a difficulty with landmark-based methods whether the landmarks are selected manually or automatically (for example, by an algorithm that looks for points in the images that should be well defined such as points of maximum curvature or minimum intensity). In either case, it is easy for errors to be made, so that points that should be in correspondence are not, or where there is some random error in the positioning of the landmark with respect to the point that it is intended to mark. For humans, marking up points on objects consistently is particularly difficult, and there is experimental evidence that the lack of correspondence between pairs of landmarks can substantially affect the diffeomorphisms that are identified in order to match the images; see, for example, [20]. We provide a solution to this problem based on a Bayesian formulation of the landmark-matching problem.

In this paper, we parameterize the diffeomorphisms by time-dependent deformation fields $\boldsymbol{v}\colon [0,1] \times B \to \mathbb{R}^d$ and define $\boldsymbol{\Phi}(\boldsymbol{Q}) = \boldsymbol{q}(1)$ for $\boldsymbol{Q} \in B$, where $\boldsymbol{q}(t)$ for $t \in [0,1]$ satisfies the initial-value problem

$$\frac{d\boldsymbol{q}}{dt} = \boldsymbol{v}(t, \boldsymbol{q}(t)), \qquad \boldsymbol{q}(0) = \boldsymbol{Q}. \tag{1.1}$$

The bending energy of $\boldsymbol{\Phi}$ is defined as follows via a norm on the deformation field:

$$\text{Energy}(\boldsymbol{\Phi}) := \frac{1}{2}\|\boldsymbol{v}\|^2, \qquad \|\boldsymbol{v}\| := \left(\int_0^1 \|\mathcal{L}\boldsymbol{v}(t,\cdot)\|^2_{L^2(B,\mathbb{R}^d)}\, dt \right)^{1/2}, \tag{1.2}$$

for a differential operator $\mathcal{L}$ (for example, $\mathcal{L}$ equals the Laplacian Δ with clamped-plate boundary conditions [19]).

The case where landmarks are fully observed is well studied, and the solution is given by the following boundary-value problem: let G be the Green's function associated to the operator $\mathcal{L}^2$, and let $\boldsymbol{p}_i(t), \boldsymbol{q}_i(t)$ satisfy the Hamiltonian boundary-value problem

$$\frac{d\boldsymbol{p}_i}{dt} = -\nabla_{\boldsymbol{q}_i} H, \qquad \frac{d\boldsymbol{q}_i}{dt} = \nabla_{\boldsymbol{p}_i} H, \tag{1.3}$$

subject to $\boldsymbol{q}_i(0) = \boldsymbol{q}_i^r$ and $\boldsymbol{q}_i(1) = \boldsymbol{q}_i^t$ for the Hamiltonian $H := \frac{1}{2}\sum_{i,j=1}^N \boldsymbol{p}_i^{\mathsf{T}} \boldsymbol{p}_j G(\boldsymbol{q}_i, \boldsymbol{q}_j)$. Here $\boldsymbol{p}_i$ are known as generalized momenta. The diffeomorphism $\boldsymbol{\Phi}$ is now defined by (1.1) with

$$\boldsymbol{v}(t, \boldsymbol{q}) = \sum_{i=1}^N \boldsymbol{p}_i(t) G(\boldsymbol{q}, \boldsymbol{q}_i(t)). \tag{1.4}$$

In general, G is defined directly rather than by specifying the Green's functions of a known $\mathcal{L}$. In our experiments, we take the Gaussian function $G(\boldsymbol{q}_1, \boldsymbol{q}_2) = \exp(-(\|\boldsymbol{q}_1 - \boldsymbol{q}_2\|/\ell)^2)$ for a length scale ℓ. For smooth choices of G such as this, $\boldsymbol{\Phi}$ is a continuously differentiable function. It is invertible by reversing the direction of the flow, and hence $\boldsymbol{\Phi}\colon B \to B$ is a diffeomorphism. See, for example, [23] and, in more general situations, [14, 21].

Our focus in this paper is on treating uncertainty around landmark positions and sensitivity of the diffeomorphism to noise. To study this problem, we introduce a Bayesian formulation

and define prior distributions on the set of diffeomorphisms. We then condition the prior on noisy observations of the landmarks to define a posterior distribution.

The choice of prior distribution is an important consideration, and we make a practical choice ensuring that diffeomorphisms having less bending energy are preferred. This is the Gibbs canonical distribution, the benefits of which are that both ends of the path are treated equally and that it has a time reversal symmetry (i.e., the Gibbs distribution is invariant under change of variable $t \mapsto 1-t$).

We consider Langevin-type perturbations of (1.3), which have the Gibbs distribution $\exp(-\beta H)$ (with inverse temperature $\beta > 0$) as an invariant measure. The advantage now is that, with suitable initial data, the solutions of the Langevin equation $[\boldsymbol{p}_i(t), \boldsymbol{q}_i(t)]$ all follow the same distribution $\exp(-\beta H)$ for $t \in [0,1]$.

It can be seen that diffeomorphisms with lower bending energy are preferred by considering the Hamiltonian using (1.4):

$$H(\boldsymbol{p}_i(t), \boldsymbol{q}_i(t)) = \frac{1}{2}\sum_{j=1}^{N} \boldsymbol{p}_j(t)^{\mathsf{T}} \boldsymbol{v}(t, \boldsymbol{q}_j(t)) = \frac{1}{2}\sum_{j=1}^{N} \int_B \boldsymbol{p}_j(t)^{\mathsf{T}} \delta_{\boldsymbol{q}_j(t)}(\boldsymbol{x}) \boldsymbol{v}(t,\boldsymbol{x})\, d\boldsymbol{x}$$

(if $\mathcal{L}^2 G = \delta$ and $\mathcal{L}$ is self adjoint)

$$= \frac{1}{2}\langle \mathcal{L}^2 \boldsymbol{v}(t,\cdot), \boldsymbol{v}(t,\cdot)\rangle_{L^2(B,\mathbb{R}^d)} = \frac{1}{2}\big\|\mathcal{L}\boldsymbol{v}(t,\cdot)\big\|^2_{L^2(B,\mathbb{R}^d)}.$$

Hence, $\int_0^1 H(\boldsymbol{p}_i(t), \boldsymbol{q}_i(t))\, dt = \text{Energy}(\boldsymbol{\Phi})$, and we see that diffeomorphisms $\boldsymbol{\Phi}$ with less bending energy are associated to paths $[\boldsymbol{p}_i(t), \boldsymbol{q}_i(t)]$ that have a larger density under the Gibbs measure $\exp(-\beta H)$.

1.1. Previous work. We are aware of three papers that have looked at image registration in the presence of noise. Most similar to ours is [29], where the trajectories $\boldsymbol{q}_i(t)$, for $t \in [0,1]$ and $i = 1, \dots, N$, are imagined to be noisy observations of some true trajectories $\boldsymbol{Q}_i(t)$. Specifically, they wish to minimize

$$\int_0^1 \big\|\mathcal{L}\boldsymbol{v}(t,\cdot)\big\|^2_{L^2(B,\mathbb{R}^d)}\, dt + \sigma \sum_{i=1}^{N} \int_0^1 \big\|\boldsymbol{q}_i(t) - \boldsymbol{Q}_i(t)\big\|^2\, dt$$

for a parameter $\sigma > 0$. The first term corresponds to a bending energy, and the second penalizes deviations from $\boldsymbol{Q}_i(t)$. This leads to a controlled Hamiltonian system,

$$\frac{d\boldsymbol{p}_i}{dt} = -\nabla_{\boldsymbol{q}_i} H + \sigma(\boldsymbol{q}_i - \boldsymbol{Q}_i(t)), \qquad \frac{d\boldsymbol{q}_i}{dt} = \nabla_{\boldsymbol{p}_i} H.$$

If a white-noise model is assumed for the observation error $\boldsymbol{q}_i(t) - \boldsymbol{Q}_i(t)$, this gives the SDE

$$d\boldsymbol{p}_i = -\nabla_{\boldsymbol{q}_i} H\, dt + \sigma\, d\boldsymbol{W}_i(t), \qquad \frac{d\boldsymbol{q}_i}{dt} = \nabla_{\boldsymbol{p}_i} H. \tag{1.5}$$

This system is identical to (2.1), except that no dissipation is included, and therefore it will not have a Gibbs distribution as invariant measure.

In [4], registrations where two curves are matched (in two dimensions) are studied. A set of discrete points is defined on one curve, and noisy observations are made on the second. Registrations are defined by an initial momentum and, to match curves rather than points, reparameterizations of the curve are also included. A Gaussian prior distribution is defined on the joint space of initial momentum and reparameterizations. Observations are made with independent Gaussian noise. The authors provide a Monte Carlo Markov chain (MCMC) method for sampling the posterior distribution. Hamiltonian equations are used to define the diffeomorphism, and no noise is introduced along the trajectories. In the case of landmark matching, there is no advantage to introducing a prior distribution on the initial momentum, as the data specifies the initial momentum completely. For noisy landmark matching, the approach has value, being simpler than the Langevin equations, but the results will depend on the end on which the prior distribution is specified.

A method to include stochasticity into the large deformation diffeomorphic metric mapping (LDDMM) framework of image registration (see [30] for details) is presented in [1]. In this approach, noise is introduced into the time-dependent deformation field from the start point to the end point, leading to a stochastic version of the Euler–Poincaré (EPDiff) equations. The authors also introduce an expectation-maximization (EM) algorithm for estimating the noise parameters based on data. The approach is based on two other papers of relevance, which add cylindrical noise to the variational principles of systems of evolutionary PDEs. By taking the system in variational form, this introduces noise perturbations into the advection equation (which corresponds to (1.4)). To preserve the conservation laws encoded in the PDEs, the momentum equations are left unchanged. The resulting trajectories in $\boldsymbol{q}_i(t)$ have the same regularity as Brownian motion and satisfy Stratonovich SDEs, which are invariant to the relabeling Lie group. The approach was originally developed for the Euler equations for an ideal fluid in [12] and was extended to the EPDiff equations in [15]. While their examples are for soliton dynamics in one spatial dimension, under particular choices of metric on the diffeomorphism group, the equations of image deformation are also EPDiff equations; hence the work in [1].

1.2. Organization. This paper is organized as follows. Our Langevin equations are described in section 2 and some basic theory established. Unfortunately, these Langevin equations are hypoelliptic, and the Hamiltonian is not separable, making the equations difficult to work with numerically. Therefore, in section 3 we introduce three numerically convenient prior distributions based on the Langevin equation. Section 4 formulates inverse problems based on the prior distributions. Two are image registrations given noisy observations of the landmarks; a third asks for the average position of a family of landmark sets. This section includes numerical experiments demonstrating our method on a variety of simple curve registrations. Further simulations and examples are given in the supplementary material (M107928_01.pdf [local/web 3.06MB]).

1.3. Notation. We denote the Euclidean norm on $\mathbb{R}^d$ by $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}}$ and the $d\times d$ identity matrix by I_d. For a subset B of $\mathbb{R}^d$, $L^2(B,\mathbb{R}^d)$ is the usual Hilbert space of square-integrable functions from $B \to \mathbb{R}^d$ with inner product $\langle \boldsymbol{f},\boldsymbol{g}\rangle_{L^2(B,\mathbb{R}^d)} = \int_B \boldsymbol{f}(\boldsymbol{x})^\mathsf{T}\boldsymbol{g}(\boldsymbol{x})\,d\boldsymbol{x}$. We often work with position vectors $\boldsymbol{q}_i \in B \subset \mathbb{R}^d$ and conjugate momenta $\boldsymbol{p}_i \in \mathbb{R}^d$ for $i = 1,\dots,N$. We denote the joint vector $[\boldsymbol{p}_1,\dots,\boldsymbol{p}_N]$ by $\boldsymbol{p} \in \mathbb{R}^{dN}$ and similarly for $\boldsymbol{q} \in \mathbb{R}^{dN}$. The combined vector

$[\boldsymbol{p}, \boldsymbol{q}]$ is denoted $\boldsymbol{z} \in \mathbb{R}^{2dN}$. For a symmetric and positive-definite function $G\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, let $\mathcal{G}(\boldsymbol{q})$ denote the $N \times N$ matrix with entries $G(\boldsymbol{q}_i, \boldsymbol{q}_j)$.

2. Generalized Langevin equations. The classical landmark-matching problem can be solved as a Hamiltonian boundary-value problem. The dynamics in a Hamiltonian model have constant energy as measured by H. Instead, we connect the system to a heat bath and look at constant-temperature dynamics. We consider a heat bath with inverse temperature β. One method of constant-temperature particle dynamics is the Langevin equation. That is, we consider the system of stochastic ODEs on $\mathbb{R}^{2dN}$ given by

$$d\boldsymbol{p}_i = \Big[-\lambda \nabla_{\boldsymbol{p}_i} H - \nabla_{\boldsymbol{q}_i} H \Big]\, dt + \sigma\, d\boldsymbol{W}_i(t), \qquad \frac{d\boldsymbol{q}_i}{dt} = \nabla_{\boldsymbol{p}_i} H \tag{2.1}$$

for a dissipation $\lambda > 0$ and diffusion $\sigma > 0$. Here $\boldsymbol{W}_i(t)$ are i.i.d. (independently and identically distributed) $\mathbb{R}^d$ Brownian motions. For $\beta = 2\lambda/\sigma^2$, a potential $V\colon \mathbb{R}^{dN} \to \mathbb{R}$, and $H = \frac{1}{2}\boldsymbol{p}^{\mathsf{T}}\boldsymbol{p} + V(\boldsymbol{q})$, (2.1) is the classical Langevin equation where the marginal invariant distribution for $\boldsymbol{p}$ is $\mathrm{N}(\boldsymbol{0}, \beta^{-1} I_{dN})$, and hence the average temperature $\frac{1}{d}\mathbb{E}\big[\boldsymbol{p}_i^{\mathsf{T}}\boldsymbol{p}_i\big]$ per degree of freedom is the constant β^{-1}. Let $[\boldsymbol{p}_i(t), \boldsymbol{q}_i(t)]$ for $t \in [0,1]$ satisfy (2.1), and define $\boldsymbol{\Phi}(\boldsymbol{Q})$ as in (1.1) and (1.4). Notice that $\boldsymbol{\Phi}(\boldsymbol{q}_i(0)) = \boldsymbol{q}_i(1)$. In perturbing (2.1) from (1.3), only the momentum equation is changed, so the equations for $\boldsymbol{q}$ are untouched and consistent with the definition of $\boldsymbol{v}(t, \boldsymbol{q})$ and hence $\boldsymbol{\Phi}$.

The solution of (2.1) is related to (1.5) by a Girsanov transformation. Let π and ν be the distribution on the path space $C([0,1], \mathbb{R}^{2dN})$ of (2.1) and (1.5), respectively. Then for $\boldsymbol{z} = [\boldsymbol{p}, \boldsymbol{q}]$,

$$d\pi(\boldsymbol{z}) = \frac{1}{\phi(\boldsymbol{z})} d\nu(\boldsymbol{z}),$$

where

$$\log(\phi(\boldsymbol{z})) = \sum_{i=1}^{N} \left[\frac{\lambda}{\sigma} \int_0^1 \boldsymbol{p}_i(t)^{\mathsf{T}} d\boldsymbol{W}_i(t) - \frac{\lambda^2}{2\sigma^2} \int_0^1 \big\|\boldsymbol{p}_i(t)\big\|^2 dt \right];$$

see [10, Lemma 5.2].

To define a distribution on the family of diffeomorphisms, it remains to choose initial data. If we specify a distribution on $[\boldsymbol{p}, \boldsymbol{q}]$ at $t = 0$, (2.1) implies a distribution on the paths and hence on $\boldsymbol{\Phi}$ via (1.1) and (1.4). The obvious choice is the Gibbs distribution $\exp(-\beta H)$. If $\sigma^2\beta = 2\lambda$ (the fluctuation–dissipation relation), then the Gibbs distribution is an invariant measure of (2.1). To see this, note that the generator of (2.1) is

$$L = \nabla_{\boldsymbol{p}} H \cdot \nabla_{\boldsymbol{q}} + (-\lambda \nabla_{\boldsymbol{p}} H - \nabla_{\boldsymbol{q}} H) \cdot \nabla_{\boldsymbol{p}} + \frac{1}{2}\sigma^2 \nabla_{\boldsymbol{p}}^2,$$

and its adjoint is

$$L^* \rho = -\nabla_{\boldsymbol{q}} \cdot ((\nabla_{\boldsymbol{p}} H)\rho) - \nabla_{\boldsymbol{p}} \cdot ((-\lambda \nabla_{\boldsymbol{p}} H - \nabla_{\boldsymbol{q}} H)\rho) + \frac{1}{2}\sigma^2 \nabla_{\boldsymbol{p}}^2 \rho.$$

The Fokker–Planck equation for the pdf $\rho(\boldsymbol{p}, \boldsymbol{q}, t)$ is

$$\frac{\partial \rho}{\partial t} = -\nabla_{\boldsymbol{q}} \rho \cdot \nabla_{\boldsymbol{p}} H + \big(\lambda \nabla_{\boldsymbol{p}} H \cdot \nabla_{\boldsymbol{p}} + \nabla_{\boldsymbol{p}} \cdot \lambda \nabla_{\boldsymbol{p}} H\big)\rho + \nabla_{\boldsymbol{p}} \rho \cdot \nabla_{\boldsymbol{q}} H + \frac{1}{2}\sigma^2 \nabla_{\boldsymbol{p}}^2 \rho.$$

Put $\rho = e^{-\beta H}$ to see that

$$\frac{\partial \rho}{\partial t} = (\beta \nabla_{\boldsymbol{q}} H) \cdot \nabla_{\boldsymbol{p}} H \, \rho + (-\lambda \nabla_{\boldsymbol{p}} H \cdot \beta \nabla_{\boldsymbol{p}} H + \nabla_{\boldsymbol{p}} \cdot \lambda \nabla_{\boldsymbol{p}} H)\rho - \beta \nabla_{\boldsymbol{p}} H \cdot \nabla_{\boldsymbol{q}} H \rho + \frac{1}{2}\sigma^2(-\beta \nabla_{\boldsymbol{p}}^2 H + \beta^2 \nabla_{\boldsymbol{p}} H \cdot \nabla_{\boldsymbol{p}} H)\rho.$$

Then, $\partial\rho/\partial t = 0$ if $\sigma^2\beta = 2\lambda$, and ρ is an invariant measure. In some cases, it can be shown additionally that ρ is a probability distribution. When B is bounded (as is usually the case for images), the phase space is compact in position space, and, if G is a uniformly positive-definite function, $\exp(-\beta H)$ can be rescaled to be a probability measure. This happens for the clamped-plate Green's function [19]. Furthermore, in some cases, the system is ergodic; precise conditions are given in [28], which studies generalized Langevin equations such as (2.1) and provides conditions on H to achieve a unique invariant measure.

While invariant measures are appealing, we view the trajectories as convenient parameterizations of the diffeomorphism, and they are not themselves of interest. Furthermore, in some cases (see section 9.2 of [30]), the domain B is taken to be $\mathbb{R}^d$, and G is translation invariant (e.g., $G(\boldsymbol{q}_1, \boldsymbol{q}_2) = \exp(-(\|\boldsymbol{q}_1 - \boldsymbol{q}_2\|/\ell)^2)$ for a length scale ℓ), and this means $\exp(-\beta H)$ cannot be a probability measure on $\mathbb{R}^{2dN}$. It is simpler to ask for a distribution on the diffeomorphism that is invariant under taking the inverse; that is, $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^{-1}$ have the same distribution. To achieve this, $[\boldsymbol{p}(t), \boldsymbol{q}(t)]$ should have the same distribution under the time reversal $t \mapsto 1 - t$. This can be achieved simply by setting initial data at $t = 1/2$ and flowing forward and backward using the same dynamics. Precisely, choose an initial probability distribution μ^* on $\mathbb{R}^{2dN}$. Given $[\boldsymbol{p}(1/2), \boldsymbol{q}(1/2)] \sim \mu^*$, compute $\boldsymbol{p}(t)$ and $\boldsymbol{q}(t)$ for $t > 1/2$ by solving (2.1). For $t < 1/2$, solve

$$d\boldsymbol{p}_i = \left[\lambda \nabla_{\boldsymbol{p}_i} H - \nabla_{\boldsymbol{q}_i} H\right] dt + \sigma \, d\boldsymbol{W}_i(t), \qquad \frac{d\boldsymbol{q}_i}{dt} = \nabla_{\boldsymbol{p}_i} H. \tag{2.2}$$

Here the sign of the dissipation is changed as we evolve the system forward by decreasing t. The distribution of $[\boldsymbol{p}(t), \boldsymbol{q}(t)]$ is unchanged by $t \mapsto 1 - t$, as can be verified using the Fokker–Planck equation.

One choice for μ^* comes by choosing distinguished landmark positions $\boldsymbol{q}_i^*$ and conditioning the Gibbs distribution on $\boldsymbol{q}_i^*$. Define the covariance matrix C by $C^{-1} = \beta \mathcal{G}(\boldsymbol{q}^*) \otimes I_d$ (the matrix C is positive definite if G is a positive-definite function and the points are distinct; see subsection 1.3 for a definition of $\mathcal{G}$). With $\boldsymbol{q}^* := [\boldsymbol{q}_1^*, \ldots, \boldsymbol{q}_N^*]$, we could choose $\mu^* = \mathrm{N}(\boldsymbol{0}, C) \times \delta_{\boldsymbol{q}^*} \asymp \exp(-\beta H(\cdot, \boldsymbol{q}^*)) \times \delta_{\boldsymbol{q}^*}$, which is the Gibbs distribution conditioned on positions $\boldsymbol{q}^*$. We prefer to allow deviation in the position also, and set $\mu^* = \mathrm{N}(\boldsymbol{0}, C) \times \mathrm{N}(\boldsymbol{q}^*, \delta^2 I_{dN})$ for some variance $\delta^2 > 0$. Then μ^* is the product of Gaussian distributions, where positions are easily sampled independently from $\mathrm{N}(\boldsymbol{q}_i^*, \delta^2 I_d)$ and momenta $\boldsymbol{p}$ are sampled from $\mathrm{N}(\boldsymbol{0}, C)$. The matrix C is a $dN \times dN$-covariance matrix. Despite the size, standard techniques such as the Cholesky or spectral factorization can be used to sample $\boldsymbol{p}$.

To summarize, we have defined two prior distributions, both based on the generalized Langevin system (2.2). Ideally, we take the Gibbs distribution for initial data and flow forward by (2.2) to define a distribution on $\boldsymbol{\Phi}\colon B \to B$. This approach is not always convenient, as the Gibbs distribution may not be a probability distribution and may also be difficult to sample

and calculate with. An alternative is to choose a convenient distribution at $t = 1/2$ and flow forward by (2.1) and backward by (2.2) to define a distribution on paths and hence on $\boldsymbol{\Phi}$.

2.1. Push-forward example. The generalized Langevin equation defines a distribution on the family of diffeomorphisms $\boldsymbol{\Phi}\colon B \to B$. We choose landmarks $\boldsymbol{q}_1, \dots, \boldsymbol{q}_N$, the inverse temperature β, the dissipation coefficient λ, and an initial probability distribution μ^* on $\mathbb{R}^{2dN}$ at some time $t^* \in [0,1]$. Then the Langevin equation can be solved to find paths $\boldsymbol{q}_i(t), \boldsymbol{p}_i(t)$ for $t \in [0,1]$, and this defines $\boldsymbol{\Phi}\colon B \to B$ by (1.1) and (1.4).

To numerically simulate (2.1) with a time step $\Delta t = 1/N_{\Delta t}$ for $N_{\Delta t} \in \mathbb{N}$, consider times $t_n = n\Delta t$ and the approximations $\boldsymbol{P}_n \approx [\boldsymbol{p}_1(t_n), \dots, \boldsymbol{p}_N(t_n)]$ and $\boldsymbol{Q}_n \approx [\boldsymbol{q}_1(t_n), \dots, \boldsymbol{q}_N(t_n)]$ given by the Euler–Maruyama method,

$$\begin{pmatrix} \boldsymbol{P}_{n+1} \\ \boldsymbol{Q}_{n+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{P}_n \\ \boldsymbol{Q}_n \end{pmatrix} + \begin{pmatrix} -\lambda \nabla_{\boldsymbol{p}} H \Delta t - \nabla_{\boldsymbol{q}} H \Delta t + \sigma \Delta \boldsymbol{W}_n \\ \nabla_{\boldsymbol{p}} H \Delta t \end{pmatrix}, \tag{2.3}$$

where H on the right-hand side is evaluated at $(\boldsymbol{P}_n, \boldsymbol{Q}_n)$ and $\Delta \boldsymbol{W}_n$ are i.i.d. $N(\boldsymbol{0}, I_{dN}\Delta t)$ random variables. This method converges in the root-mean-square sense with first order (subject to smoothness and growth conditions on H) [17].

We give numerical examples of the push-forward map $\boldsymbol{\Phi}$ for the Green's function $G(\boldsymbol{q}_1, \boldsymbol{q}_2) = \exp(-(\|\boldsymbol{q}_1 - \boldsymbol{q}_2\|/\ell)^2)$ with $\ell = 0.5$ in two dimensions ($d = 2$). Consider $B = [-1,1]^2$ and 20 regularly spaced reference points $\boldsymbol{q}_i^r$ on the unit circle. For the initial distribution, we take $\boldsymbol{q}_i(0) = \boldsymbol{q}_i^r$ and generate reference momenta $\boldsymbol{p}_i(0)$ from the conditional Gibbs distribution, so that $\boldsymbol{p}(0) \sim \mathrm{N}(\boldsymbol{0}, C)$ for $C^{-1} = \beta \mathcal{G}(\boldsymbol{q}^r) \otimes I_2$. Then approximate $\boldsymbol{p}_i(t_n), \boldsymbol{q}_i(t_n)$ by (2.3). We can now apply the explicit Euler method to (1.1) and (1.4) to define a mapping $\boldsymbol{\Phi}\colon B \to B$. It can be shown [22] that the approximate $\boldsymbol{\Phi}$ is also a diffeomorphism when Δt is sufficiently small. We show samples of the action of $\boldsymbol{\Phi}$ on a rectangular grid in Figure 1 for different values of the inverse temperature β.

3. Approximation of generalized Langevin equations. Suppose that reference and target landmarks $\boldsymbol{q}_i^r$ and $\boldsymbol{q}_i^t$ are known exactly. In Bayesian statistics, the prior distribution is conditioned on the data (landmarks in our case) to define a posterior distribution (on the paths $\boldsymbol{p}(t), \boldsymbol{q}(t)$ and hence on diffeomorphisms $\boldsymbol{\Phi}$). For the generalized Langevin prior with Gibbs initial data and exact landmark data, the posterior distribution on $[\boldsymbol{p}(t), \boldsymbol{q}(t)]$ is generated by taking solutions of (2.1) with initial data $\boldsymbol{q}(0) = \boldsymbol{q}^r$ and $\boldsymbol{p}(0) \sim \exp(-\beta H(\boldsymbol{q}^r, \cdot))$ and conditioning on $\boldsymbol{q}(1) = \boldsymbol{q}^t$. This is a type of diffusion bridge, which is important in parameter-estimation algorithms for SDEs; see [2, 11, 25].

In our case, the SDE gives a hypoelliptic diffusion, and we condition only on the position variables. The problem is similar to [11], which develops a stochastic PDE for sampling Langevin diffusion bridges with the separable Hamiltonian $H = \frac{1}{2}p^2 + V(q)$ for a potential V. It is not clear how their approach generalizes to the present situation with a nonseparable H. The method of analysis uses the Girsanov theorem to replace (2.1) by a diffusion bridge for a linear SDE [5]. The linear SDE has a Gaussian distribution, and standard formulas for conditioning Gaussian distributions are available. This technique underlies several approaches to sampling diffusion bridges such as [9, 11]. In the present situation, Girsanov is much less effective, as the nonlinearities in the position equation due to $\nabla_{\boldsymbol{p}_i} H = \sum_{j=1}^N \boldsymbol{p}_i G(\boldsymbol{q}_i, \boldsymbol{q}_j)$ are unchanged by Girsanov's transformation, and it is hard to find a linear SDE to work with.

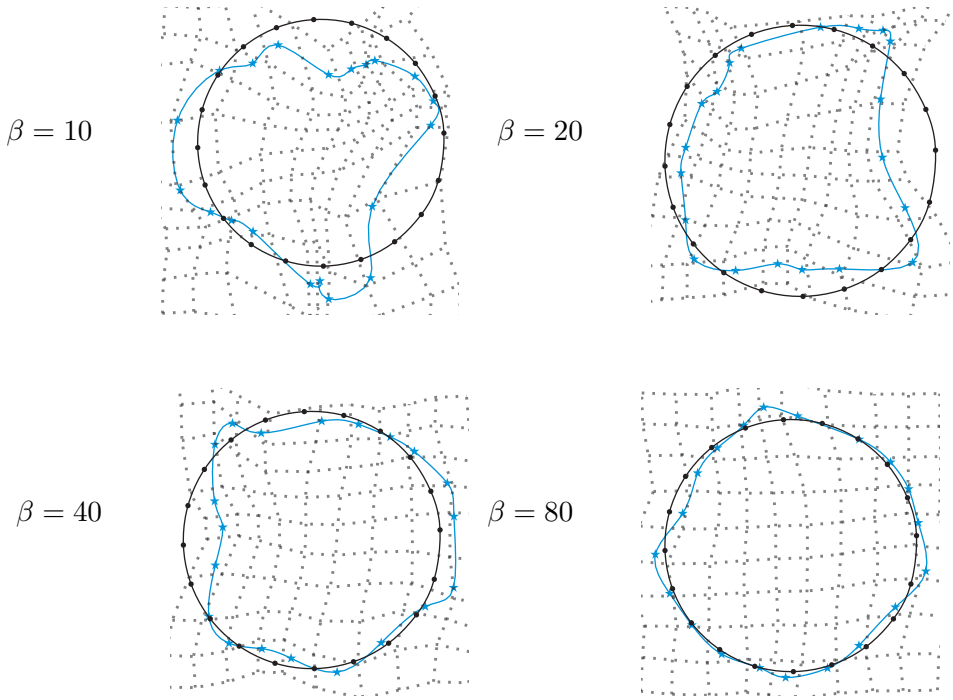

Figure 1. *Push-forward maps* $\boldsymbol{\Phi}$ *applied to a grid on* $B = [-1, 1]^2$ *and the unit circle (shown in blue) with* $G(\boldsymbol{q}_1, \boldsymbol{q}_2) = \exp(-\|\boldsymbol{q}_1 - \boldsymbol{q}_2\|^2/\ell^2)$ *for* $\ell = 0.5$, $\lambda = 0.5$, *and* $\beta = 10, 20, 40, 80$, *based on* $\boldsymbol{q}_i^r$ *as the marked points* ($\star$) *and* $\boldsymbol{p}(0) \sim \mathrm{N}(\boldsymbol{0}, C)$ *for* $C^{-1} = \beta \mathcal{G}(\boldsymbol{q}^r) \otimes I_2$. *As the inverse temperature* β *is increased, the circle is pushed forward to smoother shapes.*

Other approaches to sampling diffusion bridges include [2], which is not developed in the hypoelliptic case, or the Doob h-transform [25], which is computationally very demanding, as it involves computing the full pdf of the diffusion. Unfortunately, none of the known methods for diffusion bridges works with (2.1) to give computationally convenient algorithms.

Without an efficient method for sampling the diffusion bridge, it is hard to formulate an MCMC method with good acceptance rates. Consequently, the generalized Langevin prior distribution is difficult to use in Bayesian statistics, and we now turn to simpler prior distributions, which arise by approximating the Langevin equation. We introduce three priors, one based on a linearized Langevin equation and two based on the Baker–Campbell–Hausdorff formula for operator splittings.

All three of these methods are based on a regime of small dissipation λ and large inverse temperature β. In this case, sample paths of the Langevin equation are close to those of the Hamiltonian system on the time interval $[0, 1]$. This is a reasonable assumption in applications, as we want the time scale $1/\lambda \gg 1$, so that the landmarks $\boldsymbol{q}_i(t)$ at $t = 0$ and $t = 1$ are well coupled, but there is some drift in them. As we saw in Figure 1, small β leads to large perturbations of the initial shape. Therefore, we assume that $\sigma^2 = 2\lambda/\beta$ is small for computational convenience. In cases where these assumptions are not sufficient, it may be necessary to consider a higher-order method, but we do not do that here.

3.1. Linearized Langevin equation. In this section, based on small σ^2, we linearize the Langevin equation about the Hamiltonian solution to define a Gaussian prior distribution.

Let $\hat{\boldsymbol{z}}(t) = [\hat{\boldsymbol{p}}(t), \hat{\boldsymbol{q}}(t)]$ denote a solution of (1.3). Write the solution $\boldsymbol{z}(t) = [\boldsymbol{p}(t), \boldsymbol{q}(t)]$ of (2.1) as $\boldsymbol{z}(t) = \hat{\boldsymbol{z}}(t) + \boldsymbol{\delta}(t) + \boldsymbol{R}(t)$, where $\boldsymbol{\delta}(t)$ is a first-order correction given by linearizing (2.1) around $\hat{\boldsymbol{z}}(t)$. With initial conditions $\boldsymbol{\delta}(t^*) = \boldsymbol{z}(t^*) - \hat{\boldsymbol{z}}(t^*)$, it is defined by the linear system of SDEs,

$$d\boldsymbol{\delta} = \left[-\lambda \begin{pmatrix} \nabla_{\boldsymbol{p}} H(\hat{\boldsymbol{z}}(t)) \\ \boldsymbol{0} \end{pmatrix} + B^+(t)\boldsymbol{\delta}\right] dt + \begin{pmatrix} \sigma I_{dN} \\ 0 \end{pmatrix} d\boldsymbol{W}(t), \tag{3.1}$$

where $\boldsymbol{W}(t)$ is a $\mathbb{R}^{dN}$ Brownian motion and

$$B^+(t) = \begin{pmatrix} -\lambda \nabla_{\boldsymbol{pp}} H - \nabla_{\boldsymbol{qp}} H & -\lambda \nabla_{\boldsymbol{pq}} H - \nabla_{\boldsymbol{qq}} H \\ \nabla_{\boldsymbol{pp}} H & \nabla_{\boldsymbol{pq}} H \end{pmatrix},$$

all evaluated at $\hat{\boldsymbol{z}}(t)$. In the case $\lambda = \sigma = 0$, $\boldsymbol{\delta} = \boldsymbol{0}$ solves (3.1). With smoothness and growth conditions on H, it can be shown that the remainder $\boldsymbol{R}(t) = \mathcal{O}(\sigma^2 + \lambda^2)$ [8].

To preserve the symmetry of the system, we specify an initial distribution at $t^* = 1/2$ and ask that $\boldsymbol{\delta}(t^*) \sim \mu^*$. For $t < 1/2$, we use

$$d\boldsymbol{\delta} = \left[\lambda \begin{pmatrix} \nabla_{\boldsymbol{p}} H(\boldsymbol{p}^*(t), \boldsymbol{q}^*(t)) \\ \boldsymbol{0} \end{pmatrix} + B^-(t)\boldsymbol{\delta}\right] dt + \begin{pmatrix} \sigma I_{dN} \\ 0 \end{pmatrix} d\boldsymbol{W}(t), \tag{3.2}$$

for

$$B^-(t) = \begin{pmatrix} \lambda \nabla_{\boldsymbol{pp}} H - \nabla_{\boldsymbol{qp}} H & \lambda \nabla_{\boldsymbol{pq}} H - \nabla_{\boldsymbol{qq}} H \\ \nabla_{\boldsymbol{pp}} H & \nabla_{\boldsymbol{pq}} H \end{pmatrix}.$$

That is, the sign of the dissipation is switched, as we are specifying a final condition for this system. B^- differs by a sign in the conservative terms, as time is reversed.

Equation (3.1) is linear and its solution is a Gaussian process, and exact expressions are available for the mean and covariance in terms of deterministic integrals [16]. We prefer to use a time-stepping method to approximate (3.1). We specify the distribution at some intermediate time and need forward and backward integrators: the Euler–Maruyama method gives approximations $\boldsymbol{\delta}_n \approx \boldsymbol{\delta}(t_n)$ defined by

$$\boldsymbol{\delta}_{n+1} = \underbrace{\left(I + B_n^+ \Delta t\right)}_{=:M_n^+} \boldsymbol{\delta}_n + \boldsymbol{A}_n + \begin{pmatrix} \sigma \Delta \boldsymbol{W}_n \\ \boldsymbol{0} \end{pmatrix}, \quad \text{use for } t_{n+1} > 1/2,$$

$$\boldsymbol{\delta}_{n-1} = \underbrace{\left(I + B_n^- \Delta t\right)}_{=:M_n^-} \boldsymbol{\delta}_n + \boldsymbol{A}_n + \begin{pmatrix} \sigma \Delta \boldsymbol{W}_n \\ \boldsymbol{0} \end{pmatrix}, \quad \text{use for } t_{n-1} < 1/2,$$

where

$$\boldsymbol{A}_n = -\Delta t\, \lambda \begin{pmatrix} \nabla_{\boldsymbol{p}} H \\ \boldsymbol{0} \end{pmatrix},$$

$$B_n^+ = B(t_n), \qquad B_n^- = -B^-(t_n) = \begin{pmatrix} -\lambda \nabla_{\boldsymbol{pp}} H + \nabla_{\boldsymbol{qp}} H & -\lambda \nabla_{\boldsymbol{pq}} H + \nabla_{\boldsymbol{qq}} H \\ -\nabla_{\boldsymbol{pp}} H & -\nabla_{\boldsymbol{pq}} H \end{pmatrix}.$$

For a Gaussian initial distribution μ^*, the resulting distribution on paths and their Euler–Maruyama approximation are Gaussian. In Appendix A, we give equations for calculating the mean and covariance of the Euler–Maruyama approximations $[\boldsymbol{\delta}_0, \dots, \boldsymbol{\delta}_{N_{\Delta t}}]$.

The Gaussian distributions can be sampled to generate paths $[\boldsymbol{p}(t), \boldsymbol{q}(t)] \approx [\hat{\boldsymbol{p}}(t), \hat{\boldsymbol{q}}(t)] + \boldsymbol{\delta}(t)$. This then defines a map $\boldsymbol{\Phi}$ via (1.1) and (1.4). Note, however, that the consistency is broken, and $\boldsymbol{\Phi}(\boldsymbol{q}_i(0))$ may not equal $\boldsymbol{q}_i(1)$.

3.2. Operator splitting. Let L denote the generator associated to the generalized Langevin equation (2.1). Then $L = L_0 + \sigma^2 L_1$ for

$$L_0 = \nabla_{\boldsymbol{p}} H \nabla_{\boldsymbol{q}} - \nabla_{\boldsymbol{q}} H \nabla_{\boldsymbol{p}}, \qquad \text{known as the Liouville operator, and}$$
$$L_1 = \frac{1}{\sigma^2}\left(-\lambda \nabla_{\boldsymbol{p}} H \nabla_{\boldsymbol{p}} + \frac{1}{2}\sigma^2 \nabla_{\boldsymbol{p}} \cdot \nabla_{\boldsymbol{p}}\right) = -\frac{\beta}{2}\nabla_{\boldsymbol{p}} H \nabla_{\boldsymbol{p}} + \frac{1}{2}\nabla_{\boldsymbol{p}} \cdot \nabla_{\boldsymbol{p}}.$$

The Fokker–Planck equation is $\rho_t = L^*\rho$, where L^* denotes the adjoint of L, and describes the evolution of the pdf from a given initial density $\rho(0, \cdot) = \rho_0$. Using semigroup theory, we write $\rho(t, \cdot) = e^{L^* t}\rho_0$. We can approximate e^{A+B} via $e^A e^B + \mathcal{O}([A, B])$ or via the Strang splitting as

$$e^{A+B} \approx e^{A/2} e^B e^{A/2} + \mathcal{O}\big([B, [B, A]] + [A, [A, B]]\big),$$

where $[\cdot, \cdot]$ denotes the operator commutator. This can be applied with $A = L_0^*$ and $B = \sigma^2 L_1^*$ to simplify (2.1). In the small-noise limit, $\sigma^2 L_1^* \to 0$, but L_0 is order one, and the error is $\mathcal{O}(\sigma^2)$. These approximation strategies do preserve the Gibbs invariant measure, as $e^{\sigma^2 L_1^*}\mu = e^{L_0}\mu = 0$ for $\mu = \exp(-\beta H)$. They are also much easier to compute with than the full e^{L^*}. We look at two uses of the Strang splitting.

First splitting. Approximate

$$e^{L^*} \approx e^{\sigma^2 L_1^*/2} e^{L_0^*} e^{\sigma^2 L_1^*/2}.$$

The semigroup on the right-hand side maps

$$[\boldsymbol{p}(0), \boldsymbol{q}(0)] \underbrace{\longmapsto}_{e^{\sigma^2 L_1^*/2}} [\boldsymbol{p}(1/2), \boldsymbol{q}(0)] \underbrace{\longmapsto}_{e^{L_0^*}} [\tilde{\boldsymbol{p}}(1/2), \boldsymbol{q}(1)] \underbrace{\longmapsto}_{e^{\sigma^2 L_1^*/2}} [\boldsymbol{p}(1), \boldsymbol{q}(1)].$$

The two steps with $e^{\sigma^2 L_1^*/2}$ are described by the time-half evolution governed by the Ornstein–Uhlenbeck SDE,

$$d\boldsymbol{p} = -\lambda \nabla_{\boldsymbol{p}} H(\boldsymbol{p}, \boldsymbol{q}_0)\, dt + \sigma\, d\boldsymbol{W}(t), \qquad \boldsymbol{p}(0) = \boldsymbol{p}_0, \tag{3.3}$$

for $[\boldsymbol{p}_0, \boldsymbol{q}_0] = [\boldsymbol{p}(0), \boldsymbol{q}(0)]$ or $[\tilde{\boldsymbol{p}}(1/2), \boldsymbol{q}(1)]$. This only involves a change in momentum. The middle step with $e^{L_0^*}$ is the time-one evolution with the Hamiltonian equations (1.3). If $[\boldsymbol{p}(0), \boldsymbol{q}(0)] \sim \exp(-\beta H)$, then so are $[\boldsymbol{p}(1/2), \boldsymbol{q}(0)]$, $[\tilde{\boldsymbol{p}}(1/2), \boldsymbol{q}(1)]$, and also $[\boldsymbol{p}(1), \boldsymbol{q}(1)]$. The effects of $e^{\sigma^2 L_0^*/2}$ at either end are superfluous, as they change the momentum only; any conditioning is applied on the position data. In this way, we see fit to disregard this term and define the prior as the push forward under the Hamiltonian flow of Gibbs distribution. The density of the prior on paths $\boldsymbol{z}(t) = [\boldsymbol{p}(t), \boldsymbol{q}(t)]$ for $t \in [0, 1]$ is

$$\exp(-\beta H(\boldsymbol{z}(0)))\delta_{\boldsymbol{z}(t) - \boldsymbol{S}(t;0,\boldsymbol{z}(0))},$$

where $\boldsymbol{S}(t; s, \boldsymbol{z}_0)$ is the solution of (1.3) at time t with initial data $[\boldsymbol{p}(s), \boldsymbol{q}(s)] = \boldsymbol{z}_0$.

Second splitting. Approximate

$$e^{L^*} \approx e^{L_0^*/2} e^{\sigma^2 L_1^*} e^{L_0^*/2}.$$

The semigroup on the right-hand side maps

$$[\boldsymbol{p}(0), \boldsymbol{q}(0)] \underbrace{\longmapsto}_{e^{L_0^*/2}} [\boldsymbol{p}(1/2), \boldsymbol{q}(1/2)] \underbrace{\longmapsto}_{e^{\sigma^2 L_1^*}} [\tilde{\boldsymbol{p}}(1/2), \boldsymbol{q}(1/2)] \underbrace{\longmapsto}_{e^{L_0^*/2}} [\boldsymbol{p}(1), \boldsymbol{q}(1)].$$

Again, if $[\boldsymbol{p}(0), \boldsymbol{q}(0)] \sim \exp(-\beta H)$, then so do each of the following sets of positions and momenta. It is important to preserve each of the three parts of the approximation, as the Hamiltonian flow at either end affects all components. The density is

$$\exp(-\beta H(\boldsymbol{p}(1/2), \boldsymbol{q}(1/2))\, \upsilon(1, \tilde{\boldsymbol{p}}(1/2)); [\boldsymbol{p}(1/2), \boldsymbol{q}(1/2)])\, \delta_{\boldsymbol{z}(t)-\boldsymbol{Z}(t)},$$

where $\upsilon(t, \boldsymbol{p}; [\boldsymbol{p}_0, \boldsymbol{q}_0])$ is the density at time t of the random variable $\boldsymbol{p}(t)$ defined by the SDE

$$d\boldsymbol{p} = -\lambda \nabla_{\boldsymbol{p}} H(\boldsymbol{p}, \boldsymbol{q}_0)\, dt + \sigma d\boldsymbol{W}(t), \qquad \boldsymbol{p}(0) = \boldsymbol{p}_0. \tag{3.4}$$

The function $\boldsymbol{Z}(t)$ describes the Hamiltonian flow and is defined by

$$\boldsymbol{Z}(t) = \begin{cases} S(t; 1/2, [\tilde{\boldsymbol{p}}(1/2), \boldsymbol{q}(1/2)]), & t > 1/2, \\ S(t; 1/2, [\boldsymbol{p}(1/2), \boldsymbol{q}(1/2)]), & t < 1/2. \end{cases} \tag{3.5}$$

It will be more convenient to have both halves flow forward and write

$$\boldsymbol{Z}(t) = \begin{cases} S(t - 1/2; 0, [\tilde{\boldsymbol{p}}(1/2), \boldsymbol{q}(1/2)]), & t > 1/2, \\ RS(1/2 - t; 0, R[\boldsymbol{p}(1/2), \boldsymbol{q}(1/2)]), & t < 1/2, \end{cases}$$

where $R[\boldsymbol{p}, \boldsymbol{q}] = [-\boldsymbol{p}, \boldsymbol{q}]$ expresses the time reversal.
The key variables for conditioning are the start and end positions, $\boldsymbol{q}(0)$ and $\boldsymbol{q}(1)$. These positions are deterministic maps of the time-half data, provided by a time-half push forward of the deterministic Hamiltonian dynamics. Thus, it is convenient to express the prior in terms of $\boldsymbol{p}(1/2), \boldsymbol{q}(1/2), \tilde{\boldsymbol{p}}(1/2)$ by the density proportional to

$$\exp(-\beta H(\boldsymbol{p}(1/2), \boldsymbol{q}(1/2))\, \upsilon(1/2, \tilde{\boldsymbol{p}}(1/2)); [\boldsymbol{p}(1/2), \boldsymbol{q}(1/2)]).$$

We now show how to simplify υ when β is large and λ is small. In (3.4), $\nabla_{\boldsymbol{p}} H(\boldsymbol{p}, \boldsymbol{q}) = (\mathcal{G}(\boldsymbol{q}) \otimes I_d)\boldsymbol{p}$ for $\mathcal{G}(\boldsymbol{q})$ defined in subsection 1.3. For a deterministic $\boldsymbol{p}_0$, the solution $\boldsymbol{p}(t)$ of (3.4) is an Ornstein–Uhlenbeck process, with a Gaussian distribution with mean $\mu_t = (e^{-\lambda \mathcal{G}(\boldsymbol{q}_0) t} \otimes I_d)\, \boldsymbol{p}_0$ and covariance $C_t \otimes I_d$, for

$$C_t := \sigma^2 \frac{1}{2\lambda} \mathcal{G}(\boldsymbol{q}_0)^{-1} \Big(I_N - e^{-2\lambda\, t\, \mathcal{G}(\boldsymbol{q}_0)} \Big) = \frac{1}{\beta} \mathcal{G}(\boldsymbol{q}_0)^{-1} \Big(I_N - e^{-2\lambda\, t\, \mathcal{G}(\boldsymbol{q}_0)} \Big).$$

By Taylor's theorem, $e^{-\lambda A} = I_N - \lambda A + \int_0^1 \lambda^2 A^2 e^{-\lambda A s}(1 - s)\, ds$ for any $N \times N$ matrix A. Hence,

$$\begin{aligned} C_t &= \sigma^2\, t\, I_N + \frac{1}{\beta}\, \mathcal{G}(\boldsymbol{q}_0)^{-1} \int_0^1 4\, \lambda^2\, t^2\, \mathcal{G}(\boldsymbol{q}_0)^2 e^{-2\, \lambda\, t\, \mathcal{G}(\boldsymbol{q}_0)} (1 - s)\, ds \\ &= \sigma^2\, t\, I_N + 4 \frac{1}{\beta}\, \lambda\, t\, K \quad \text{for } K := \int_0^1 \lambda\, t\, \mathcal{G}(\boldsymbol{q}_0)\, e^{-2\, \lambda\, t\, \mathcal{G}(\boldsymbol{q}_0)} (1 - s)\, ds. \end{aligned}$$

When G is a positive-definite function, K is uniformly bounded over any $\boldsymbol{q}_0 \in \mathbb{R}^{dN}$ and $t \in [0,1]$. Therefore,

$$C_t = \sigma^2\, t\, I_N + \mathcal{O}\big(\lambda\, t/\beta\big). \tag{3.6}$$

As explained in section 3, we are interested in large β and small λ, and hence we are justified in approximating $C_t \approx \sigma^2\, t\, I_N$ for $t \in [0,1]$. Then

$$e^{\sigma^2 L_1^* t}\delta_{(\boldsymbol{p}_0,\boldsymbol{q}_0)} \approx \mathrm{N}((e^{-\lambda\, t\, \mathcal{G}(\boldsymbol{q}_0)} \otimes I_d)\boldsymbol{p}_0, \sigma^2\, t\, I_{dN}) \times \delta_{\boldsymbol{q}_0}.$$

For the prior, we are interested in $\upsilon(1, \tilde{\boldsymbol{p}}(1/2); (\boldsymbol{p}(1/2), \boldsymbol{q}(1/2)))$ and, by this approximation,

$$\upsilon(1, \cdot; (\boldsymbol{p}(1/2), \boldsymbol{q}(1/2))) \approx \mathrm{N}((e^{-\lambda\, \mathcal{G}(\boldsymbol{q}(1/2))} \otimes I_d)\boldsymbol{p}(1/2), \sigma^2 I_{dN}) \times \delta_{\boldsymbol{q}(1/2)}.$$

Hence, the prior distribution on $(\boldsymbol{p}(1/2), \boldsymbol{q}(1/2), \tilde{\boldsymbol{p}}(1/2))$ has density proportional to

$$\exp\big(-\beta H(\boldsymbol{p}(1/2), \boldsymbol{q}(1/2))\big) \exp\left(-\frac{1}{2\sigma^2}\left\|\tilde{\boldsymbol{p}}(1/2) - (e^{-\lambda\, \mathcal{G}(\boldsymbol{q}(1/2))} \otimes I_d)\boldsymbol{p}(1/2)\right\|^2\right). \tag{3.7}$$

Distributions on the paths $[\boldsymbol{p}(t), \boldsymbol{q}(t)]$ are implied by solving (1.3) with initial data $[\boldsymbol{p}(1/2),\ \boldsymbol{q}(1/2)]$ for $t > 1/2$ and with final data $[\tilde{\boldsymbol{p}}(1/2), \boldsymbol{q}(1/2)]$ for $t < 1/2$.

4. Data and experiments. We now show how to work with the prior distributions using data. For a prior distribution on the diffeomorphisms $\boldsymbol{\Phi}$, we would like to compute or sample from the conditional distribution of $\boldsymbol{\Phi}$ given that $\boldsymbol{q}_i(0) = \boldsymbol{q}_i^t + \boldsymbol{\eta}_i^t$ and $\boldsymbol{q}_i(1) = \boldsymbol{q}_i^r + \boldsymbol{\eta}_i^r$, where $\boldsymbol{\eta}_i^t, \boldsymbol{\eta}_i^r \sim \mathrm{N}(\boldsymbol{0}, \delta^2 I_d)$ are i.i.d. for some parameter $\delta > 0$. We present three cases as follows:

1. The linearized Langevin prior is Gaussian, and conditioning by observations of the landmarks with i.i.d. Gaussian errors yields a Gaussian posterior distribution. We show how to compute the posterior distribution for the Euler–Maruyama discretized equations.
2. The first splitting prior consists of a Gibbs distribution on the initial data and Hamiltonian flow equations. As such, the distribution is specified by the distribution on the initial landmarks and generalized momenta. We condition this on landmarks also with i.i.d. Gaussian errors. The posterior is not Gaussian. We show how to compute the maximum a posteriori (MAP) point and approximate the posterior covariance matrix by the Laplace method. The MAP point is a set of initial landmark positions and generalized momenta.
3. The second splitting prior consists of a Gibbs distribution on the midpoint, a second momentum for each landmark (correlated to the first) at the midpoint, and Hamiltonian flow equations. This distribution is parameterized by one set of landmarks and two sets of generalized momenta. We show how to examine the posterior distribution (again conditioning on Gaussian observations) via the MAP point and Laplace method. We interpret the MAP point as an average set of landmarks by extending the prior to allow for multiple sets of landmarks.

The discussion includes computational examples. The calculations were performed in Python using the Numpy, Matplotlib, and Scipy libraries, and the code is available for download [18]. For information about the code and for a set of further examples, see the supplementary material (M107928_01.pdf [local/web 3.06MB]). In all cases, the landmarks in each image were centered to have zero mean and then aligned using an orthogonal Procrustes transformation in order to remove potentially confusing global transformations.

4.1. Noisy landmarks via the linearized Langevin equation. The key step in defining the linearized Langevin prior is distinguishing paths about which to linearize. We choose paths $[\boldsymbol{p}(t), \boldsymbol{q}(t)]$ by solving the Hamiltonian boundary-value problem (1.3) based on the landmark data $\boldsymbol{q}_i^t$ and $\boldsymbol{q}_i^r$. Then the linearized Langevin prior is a Gaussian distribution on the paths $[\boldsymbol{p}(t), \boldsymbol{q}(t)]$ generated by (3.1), the linearization of the Langevin equations about the distinguished paths. We denote the Euler–Maruyama approximation with time step $\Delta t = 1/N_{\Delta t}$ to $[\boldsymbol{p}^*, \boldsymbol{q}^*] + \boldsymbol{\delta}$ at t_n by $[\boldsymbol{P}_n, \boldsymbol{Q}_n]$ and the vector $[\boldsymbol{P}_0, \boldsymbol{Q}_0, \ldots, \boldsymbol{P}_{N_{\Delta t}}, \boldsymbol{Q}_{N_{\Delta t}}]$ by $\boldsymbol{X}$. The mean $\boldsymbol{M}_1$ and covariance $\mathcal{C}$ of $\boldsymbol{X}$ can be found using the equations in Appendix A.

Let $\widehat{\boldsymbol{Q}}^r = \boldsymbol{Q}_0 + \boldsymbol{\eta}^r$ and $\widehat{\boldsymbol{Q}}^t = \boldsymbol{Q}_{N_{\Delta t}} + \boldsymbol{\eta}^t$ for $\boldsymbol{\eta}^r, \boldsymbol{\eta}^t \sim \mathrm{N}(\boldsymbol{0}, \delta^2 I_{dN})$ i.i.d. (the distributions are independent of each other and also of the Brownian motions). Let $\boldsymbol{Y} = [\widehat{\boldsymbol{Q}}^r, \widehat{\boldsymbol{Q}}^t]$ and $\boldsymbol{Z} = [\boldsymbol{X}, \boldsymbol{Y}]$. $\boldsymbol{Z}$ is then Gaussian with mean $[\boldsymbol{M}_1, \boldsymbol{M}_2] = \big[\boldsymbol{M}_1, \big[\mathbb{E}[\boldsymbol{Q}_0], \mathbb{E}\big[\boldsymbol{Q}_{N_{\Delta t}}\big]\big]\big]$ and covariance

$$\begin{bmatrix} C_{11} & C_{21}^{\mathsf{T}} \\ C_{21} & C_{22} \end{bmatrix}, \quad \text{where } C_{11} = \mathcal{C},\ C_{22} = \begin{bmatrix} \mathrm{Cov}(\boldsymbol{Q}_0, \boldsymbol{Q}_0) + \delta^2 I_{dN} & C_{0N_{\Delta t}} \\ C_{0N_{\Delta t}}^{\mathsf{T}} & \mathrm{Cov}(\boldsymbol{Q}_{N_{\Delta t}}, \boldsymbol{Q}_{N_{\Delta t}}) + \delta^2 I_{dN} \end{bmatrix},$$
$$\text{and } C_{21} = \begin{bmatrix} \mathrm{Cov}(\boldsymbol{Q}_0, \boldsymbol{Q}_0) & \ldots & \mathrm{Cov}(\boldsymbol{Q}_0, \boldsymbol{Q}_{N_{\Delta t}}) \\ \mathrm{Cov}(\boldsymbol{Q}_{N_{\Delta t}}, \boldsymbol{Q}_0) & \ldots & \mathrm{Cov}(\boldsymbol{Q}_{N_{\Delta t}}, \boldsymbol{Q}_{N_{\Delta t}}) \end{bmatrix}.$$

The distribution of $\boldsymbol{X}$ given observations $\widehat{\boldsymbol{Q}}^t = \boldsymbol{q}^t$ and $\widehat{\boldsymbol{Q}}^r = \boldsymbol{q}^r$ is $\mathrm{N}(\boldsymbol{M}_{1|2}, C_{1|2})$ with

$$\boldsymbol{M}_{1|2} = \boldsymbol{M}_1 + C_{12} C_{22}^{-1} (\boldsymbol{y} - \boldsymbol{M}_2), \qquad \boldsymbol{y} = [\boldsymbol{q}^r, \boldsymbol{q}^t],$$
$$C_{1|2} = C_{11} - C_{12} C_{22}^{-1} C_{21}.$$

For the number of landmarks that we consider (less than 100), this is readily computed using standard linear-algebra routines. The two inverse matrices involved are of size $dN \times dN$. The full covariance matrix is memory demanding though, as it has size $(N_{\Delta t}+1)2dN \times (N_{\Delta t}+1)2dN$.

Figure 2 shows the solution of (1.3) and the associated registration for a set of known landmarks. We linearize about the solution $[\boldsymbol{p}(t), \boldsymbol{q}(t)]$ to define a linearized Langevin prior, and Figure 3 shows the standard deviations of the computed posterior distribution at the landmark positions. Figure 4 shows the standard deviation of the posterior throughout the image space, in both the original and warped coordinate systems. The difference in standard deviations shown in Figures 3 and 4 is significant, as one comes from the posterior distribution matrix at the landmarks and the other from a Monte Carlo estimator of the distribution of $\boldsymbol{\Phi}(\boldsymbol{Q})$ for $\boldsymbol{Q}$ away from landmark points. In this linearized situation, $\boldsymbol{\Phi}$ may not agree with the linearized Langevin equation. We see this weakness again for large deformations in Figure 5, where we compare the random diffeomorphisms and the paths $\boldsymbol{q}_i(t)$ defined by samples of the posterior distribution. Though $\boldsymbol{\Phi}(\boldsymbol{q}_i^r)$ and $\boldsymbol{q}_i(1)$ agree when the data is regular, for larger

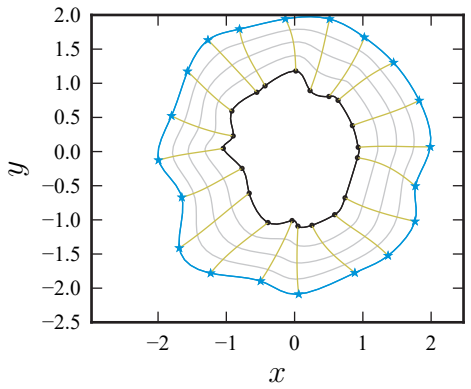

Figure 2. *The blue and black stars mark* 20 *noisy observations of regularly spaced points on two concentric circles. Using the Hamiltonian boundary-value problem* (1.3)*, we compute a diffeomorphism and show paths* $\boldsymbol{q}_i(t)$*. Three intermediate shapes are shown in grey. The yellow lines show the paths taken by the landmarks through the interpolating shapes.*

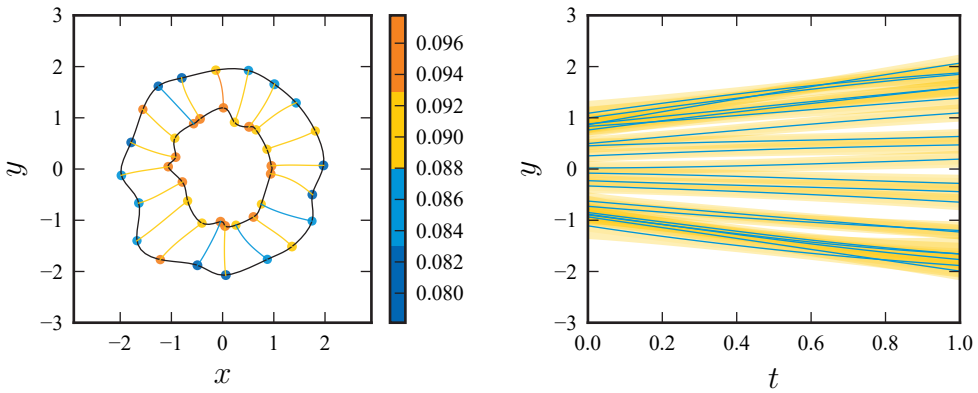

Figure 3. *A registration between two noisy observations of a circle at different scales (radii of* 1 *and* 2 *corresponding to times* $t = 0, 1$*, respectively) using the linearized Langevin prior (with* $\lambda = 0.1$ *and* $\beta = 25$*), with landmarks observed with i.i.d.* $\mathrm{N}(\mathbf{0}, \delta^2 I_d)$ *errors for* $\delta^2 = 0.01$*. The discs on the left-hand plot and the yellow shadows on the right-hand plot indicate one standard deviation of the computed posterior covariance matrix.*

deformations, there is significant disagreement. This is because $\boldsymbol{\Phi}$ is defined by (1.1) and (1.4), which is no longer identical to the linear equation (3.1) used to define $\boldsymbol{q}_i(t)$.

4.2. Noisy landmarks by operator splitting. The first splitting prior demands much less memory than the linearized Langevin prior, as the randomness concerns only the initial position and momentum. It also has the advantage of preserving the Gibbs distribution and maintaining consistency with the definition of $\boldsymbol{\Phi}$. We show how to use this prior in the same scenario as subsection 4.1. This time we are unable to sample the posterior distribution. Instead, we formulate a MAP estimator and apply a Laplace approximation to estimate the posterior covariance.

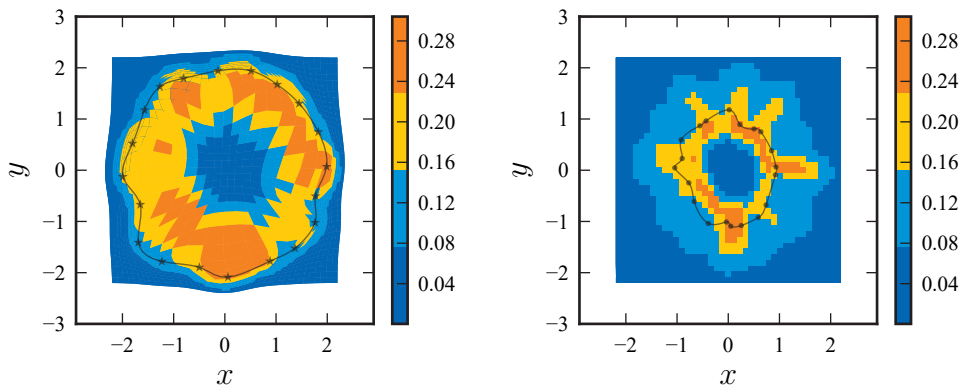

Figure 4. *The colors show the standard deviation of $\Phi(Q)$ at $\Phi(Q)$ (left-hand side) and at Q (right-hand side) for a set of uniformly spaced Q on a rectangular grid, when Φ is defined by the posterior distribution for the linearized Langevin prior.*

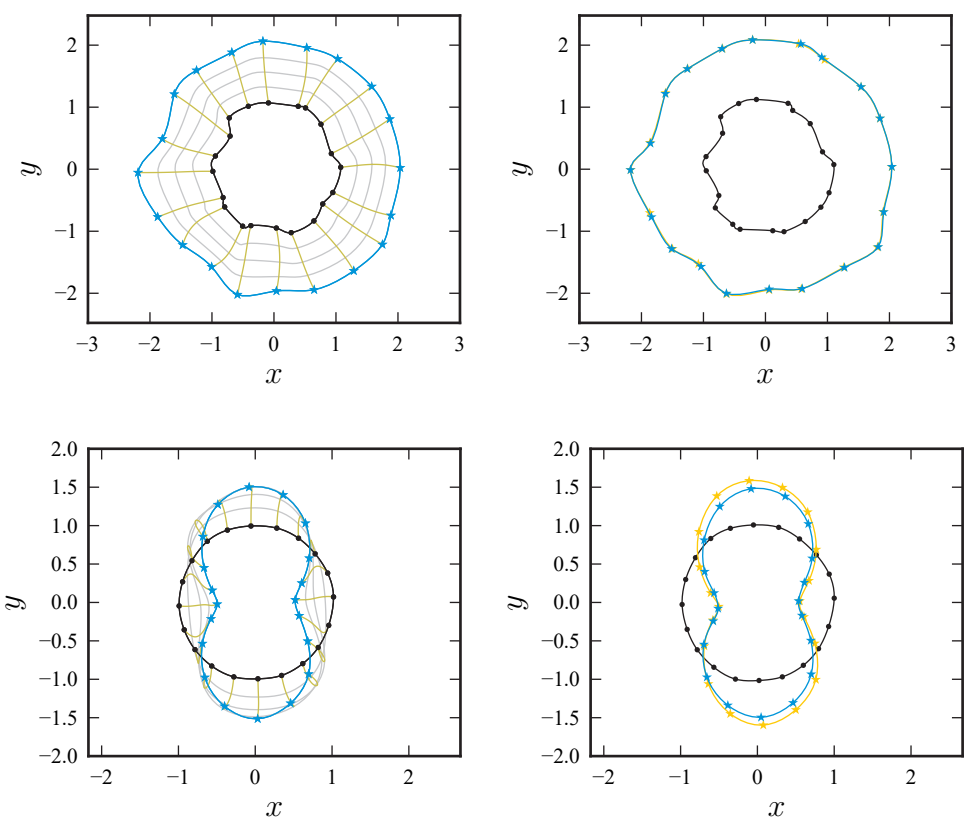

Figure 5. *In the right-hand column, blue stars mark $\Phi(q_i(0))$ and the yellow stars mark $q_i(1)$, where $q_i(t)$ and Φ are given via samples of the linearized Langevin posterior distribution, with paths shown in the left-hand column. The inner black loop marks $q_i(0)$. Due to the linearization, $\Phi(q_i(0)) \neq q_i(1)$, although it is closer on the top row where the deformation field is much smoother.*

As the observation error is independent of the prior, the posterior density is given by the pdf of the prior on $[\boldsymbol{p}_0, \boldsymbol{q}_0]$ times the data likelihood for the observations $\boldsymbol{q}_0$ and $S_q(1, 0; [\boldsymbol{p}_0, \boldsymbol{q}_0])$ of $\boldsymbol{q}^r$ and $\boldsymbol{q}^t$. The density of the first splitting prior is $\exp(-\beta H(\boldsymbol{p}_0, \boldsymbol{q}_0))$. For Gaussian observations, the data likelihood is proportional to

$$\exp\left(-\frac{1}{2\delta^2}\left(\left\|\boldsymbol{q}^r - \boldsymbol{q}_0\right\|^2 + \left\|\boldsymbol{q}^t - S_q(1; 0, [\boldsymbol{p}_0, \boldsymbol{q}_0])\right\|^2\right)\right),$$

where S_q denotes the position components in S (the Hamiltonian flow map). The posterior density is proportional to

$$\exp(-\beta H(\boldsymbol{p}_0, \boldsymbol{q}_0)) \exp\left(-\frac{1}{2\delta^2}\left(\left\|\boldsymbol{q}^r - \boldsymbol{q}_0\right\|^2 + \left\|\boldsymbol{q}^t - S_q(1; 0, [\boldsymbol{p}_0, \boldsymbol{q}_0])\right\|^2\right)\right).$$

To find the MAP point, we minimize

$$F(\boldsymbol{p}_0, \boldsymbol{q}_0) \coloneqq \beta H(\boldsymbol{p}_0, \boldsymbol{q}_0) + \frac{1}{2\delta^2}\left(\left\|\boldsymbol{q}^r - \boldsymbol{q}_0\right\|^2 + \left\|\boldsymbol{q}^t - S_q(1; 0, [\boldsymbol{p}_0, \boldsymbol{q}_0])\right\|^2\right). \tag{4.1}$$

This comprises the regularizer that comes from the Gibbs distribution and two landmark-matching terms, and can also be derived as a Tychonov regularization of the standard landmark registration problem. There is one parameter β from the Gibbs distribution, and the dissipation λ is not present. We minimize F to find an approximation to the MAP point, using standard techniques from unconstrained optimization and finite-difference methods for (1.3).

The Laplace method gives an approximation to the posterior covariance matrix by a second-order approximation to F at the MAP point $\boldsymbol{z}_0$. Thus, we evaluate the Hessian $\nabla^2 F$ of F at the MAP point. Second derivatives of F are approximated by using a Gauss–Newton approximation for the last term, so we use

$$\nabla^2 F \approx \beta \nabla^2 H + \frac{1}{\delta^2}\left[\begin{pmatrix} 0 & 0 \\ 0 & I_{dN} \end{pmatrix} + J^\mathsf{T} J\right],$$

where J is the Jacobian matrix of $S_q(1; 0, \boldsymbol{z}_0)$. The Gauss–Newton approximation guarantees that the second term is positive definite (though the Hessian of H and the overall expression may not be). To make sure the covariance is a well-defined symmetric positive-definite matrix, we form a spectral decomposition of $\nabla^2 F$, throw away any negative eigenvalues, and form the inverse matrix from the remaining eigenvalues to define a covariance matrix $C \approx \nabla^2 F^{-1}$. See Figure 6 for an example.

4.3. Second splitting prior and landmark-set averages. Averages are an important way of summarizing a data set. Under our Bayesian formulation, it is relatively simple to define a consistent average for sets of landmarks defined on multiple images, as we demonstrate in this section. The approach is similar in spirit to the arithmetic mean, which arises in calculations of the MAP point for Gaussian samples.

We use the second splitting prior and start with two sets of landmark points $\boldsymbol{q}^a$ and $\boldsymbol{q}^b$. We wish to find a third set of landmark points $\boldsymbol{q}^*$ that match both sets a, b according to some measure. We introduce momenta $\boldsymbol{p}^{*a}$ and $\boldsymbol{p}^{*b}$. Classical landmark matching gives a momentum

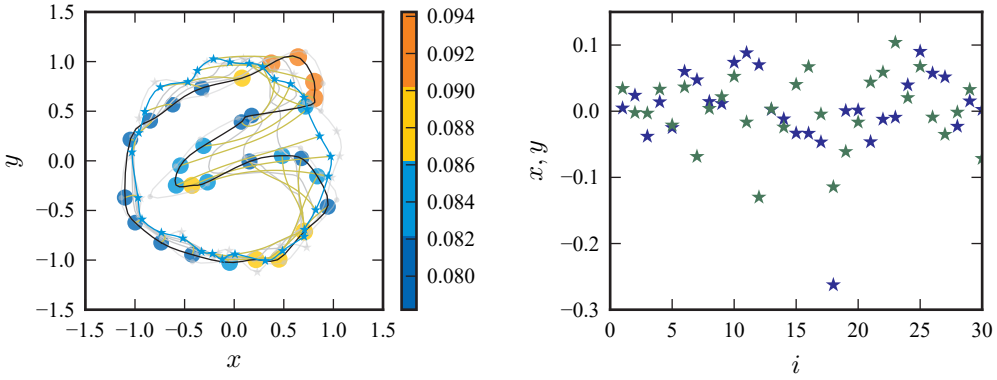

Figure 6. *Noisy landmark registration (with $\delta = \sqrt{0.005} \approx 0.07$) using the first splitting prior (with $\beta = 25$). In the left-hand plot, the grey lines mark the original landmark data. The landmarks given by the MAP registration algorithm are marked in color, with discs indicating the standard deviation on the target landmarks by the Laplace approximation for the computed posterior covariance. The right-hand plots show the difference between MAP landmarks and data landmarks.*

$\boldsymbol{p}^{*a}$ that flows $\boldsymbol{q}^*$ to $\boldsymbol{q}^a$, and similarly for b. This can be done for any $\boldsymbol{q}^*$. The second splitting prior expresses our preference for less deformation and coupling of the two momenta, and makes $\boldsymbol{q}^*$ well defined.

The second splitting prior gives a distribution on $(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}, \tilde{\boldsymbol{p}}_{1/2})$ proportional to (3.7). Substituting $\beta = 2\lambda/\sigma^2$, it is

$$\exp\bigl(-\beta H(\boldsymbol{p}(1/2), \boldsymbol{q}(1/2))\bigr) \exp\left(-\frac{\beta}{4\lambda}\left\|\tilde{\boldsymbol{p}}(1/2) - (e^{-\lambda \mathcal{G}(\boldsymbol{q}(1/2))} \otimes I_d)\boldsymbol{p}(1/2)\right\|^2\right). \tag{4.2}$$

When coupled with the likelihood function for data $\boldsymbol{q}^r$ and $\boldsymbol{q}^t$ given by

$$\exp\left(-\frac{1}{2\delta^2}\left(\left\|\boldsymbol{q}^r - S_q(1/2; 0, [-\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}])\right\|^2 + \left\|\boldsymbol{q}^t - S_q(1/2; 0, [\tilde{\boldsymbol{p}}_{1/2}, \boldsymbol{q}_{1/2}])\right\|^2\right)\right),$$

we can write down the posterior pdf. Then, to find the MAP point, we minimize the objective function,

$$\begin{aligned} F(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}, \tilde{\boldsymbol{p}}_{1/2}) &:= \beta H(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}) + \frac{\beta}{4\lambda}\left\|\tilde{\boldsymbol{p}}_{1/2} - e^{-\lambda \mathcal{G}(\boldsymbol{q}_{1/2})}\boldsymbol{p}_{1/2}\right\|^2 \\ &+ \frac{1}{2\delta^2}\left(\left\|\boldsymbol{q}^r - S_q(1/2; 0, [-\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}])\right\|^2 + \left\|\boldsymbol{q}^t - S_q(1/2; 0, [\tilde{\boldsymbol{p}}_{1/2}, \boldsymbol{q}_{1/2}])\right\|^2\right). \end{aligned} \tag{4.3}$$

This comprises the regularizer due to the Gibbs distribution, a penalty for changing the momentum at $t = 1/2$, and two landmark-matching terms. The minimizer of F gives the MAP point. We are interested in using $\boldsymbol{q}^* = \boldsymbol{q}_{1/2}$ as the average landmark set.

Before discussing numerical experiments, we describe the limiting properties of the MAP point as λ, β are varied. In the following, we assume that B^N is a convex subset of $\mathbb{R}^{dN}$ and that $\boldsymbol{q}^r, \boldsymbol{q}^t \in B^N$.

Lemma 1. *With* $\boldsymbol{p}_{1/2} = \tilde{\boldsymbol{p}}_{1/2} = \mathbf{0}$, *the minimizer of*

$$f(\boldsymbol{q}) := \left\|\boldsymbol{q}^r - S_q(1/2; 0, [-\boldsymbol{p}_{1/2}, \boldsymbol{q}])\right\|^2 + \left\|\boldsymbol{q}^t - S_q(0; 1/2, [\tilde{\boldsymbol{p}}_{1/2}, \boldsymbol{q}])\right\|^2$$

over $\boldsymbol{q}_{1/2} \in B^N$ *is* $\boldsymbol{q}_{1/2} = (\boldsymbol{q}^r + \boldsymbol{q}^t)/2$. *Hence,*

$$\min_{(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}, \tilde{\boldsymbol{p}}_{1/2}) \in \mathbb{R}^{dN} \times B^N \times \mathbb{R}^{dN}} F(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}, \tilde{\boldsymbol{p}}_{1/2}) \leq \frac{1}{4\delta^2} \left\|\boldsymbol{q}^r - \boldsymbol{q}^t\right\|^2.$$

Proof. When $\boldsymbol{p} = \boldsymbol{p}_{1/2} = \tilde{\boldsymbol{p}}_{1/2} = \mathbf{0}$, $H = 0$ and $S_q(s; t, [\boldsymbol{p}, \boldsymbol{q}]) = \boldsymbol{q}$ for all s, t. Hence, $f(\boldsymbol{q}) = \|\boldsymbol{q}^r - \boldsymbol{q}\|^2 + \|\boldsymbol{q}^t - \boldsymbol{q}\|^2$, which is minimized by $\boldsymbol{q}_{1/2} = (\boldsymbol{q}^r + \boldsymbol{q}^t)/2$. ∎

Corollary 2. *Assume that* $G(\boldsymbol{q}_i)$ *is uniformly bounded over* $\boldsymbol{q}_i \in B \subset \mathbb{R}^d$. *Then, as* $\lambda \to 0$, $\boldsymbol{q}_{1/2}$ *converges to* $S_q(1/2; 0, [\boldsymbol{p}_0, \boldsymbol{q}_0])$, *where* $[\boldsymbol{p}_0, \boldsymbol{q}_0]$ *is the MAP point for* (4.1).

Proof. As $\min F$ is bounded independently of λ, we know that

$$\frac{\beta}{\lambda} \left\|\tilde{\boldsymbol{p}}_{1/2} - e^{-\lambda \mathcal{G}(\boldsymbol{q}_{1/2})} \boldsymbol{p}_{1/2}\right\|^2$$

is bounded as $\lambda \to 0$. Hence, $\tilde{\boldsymbol{p}}_{1/2} - e^{-\lambda \mathcal{G}(\boldsymbol{q}_{1/2})} \boldsymbol{p}_{1/2} \to \mathbf{0}$. When all entries of $\mathcal{G}$ are bounded, $e^{-\lambda \mathcal{G}(\boldsymbol{q}_{1/2})} \to I_N$ as $\lambda \to 0$. Therefore, $\boldsymbol{p}_{1/2} - \tilde{\boldsymbol{p}}_{1/2} \to \mathbf{0}$, and $\boldsymbol{Z}(t)$ as defined in (3.5) is the solution of the Hamiltonian equation (1.3) on $[0, 1]$ in the limit $\lambda \to 0$. Let $[\boldsymbol{p}_0, \boldsymbol{q}_0] = RS(1/2; 0, R[\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}])$ for $R[\boldsymbol{p}, \boldsymbol{q}] = [-\boldsymbol{p}, \boldsymbol{q}]$. Then

$$\begin{aligned}
\min F \to{}& \min \beta H(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}) + 0 \\
&+ \frac{1}{2\delta^2} \left[\left\|\boldsymbol{q}^r - S_q(1/2; 0, [-\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}])\right\|^2 + \left\|\boldsymbol{q}^t - S_q(1/2; 0, [\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}])\right\|\right] \\
={}& \min \beta H(\boldsymbol{p}_0, \boldsymbol{q}_0) + \frac{1}{2\delta^2} \left[\left\|\boldsymbol{q}^r - \boldsymbol{q}_0\right\|^2 + \left\|\boldsymbol{q}^t - S_q(1; 0, [\boldsymbol{p}_0, \boldsymbol{q}_0])\right\|\right].
\end{aligned}$$

Here we use the fact that H is constant along solutions of (1.3). The last expression is the same as (4.1), as required. ∎

Corollary 3. *If* $\mathcal{G}(\boldsymbol{q})$ *is uniformly positive definite over* $\boldsymbol{q} \in B^N \subset \mathbb{R}^{dN}$, *then in the limit* $\beta \to \infty$, $\boldsymbol{q}_{1/2}$ *converges to the arithmetic average* $(\boldsymbol{q}^r + \boldsymbol{q}^t)/2$.

Proof. Rescale the objective function

$$\begin{aligned}
\frac{1}{\beta} F(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}, \tilde{\boldsymbol{p}}_{1/2}) :={}& H(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}) + \frac{1}{4\lambda} \left\|\tilde{\boldsymbol{p}}_{1/2} - e^{-\lambda \mathcal{G}(\boldsymbol{q}_{1/2})} \boldsymbol{p}_{1/2}\right\|^2 \\
&+ \frac{1}{2\beta\delta^2} \left(\left\|\boldsymbol{q}^r - S_q(1/2; 0, [-\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}])\right\|^2 + \left\|\boldsymbol{q}^t - S_q(1/2; 0, [\tilde{\boldsymbol{p}}_{1/2}, \boldsymbol{q}_{1/2}])\right\|^2\right).
\end{aligned}$$

This converges to zero as $\beta \to \infty$. Hence, $H(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}) \to 0$, so that $\boldsymbol{p}_{1/2} \to \mathbf{0}$ if $\mathcal{G}$ is uniformly positive definite. The second term implies that $\tilde{\boldsymbol{p}}_{1/2} \to \mathbf{0}$. Then $\min F \to \frac{1}{2\delta^2}(\|\boldsymbol{q}^r - \boldsymbol{q}_{1/2}\|^2 + \|\boldsymbol{q}^t - \boldsymbol{q}_{1/2}\|^2)$. Lemma 1 gives $\boldsymbol{q}_{1/2}$ as the arithmetic average. ∎

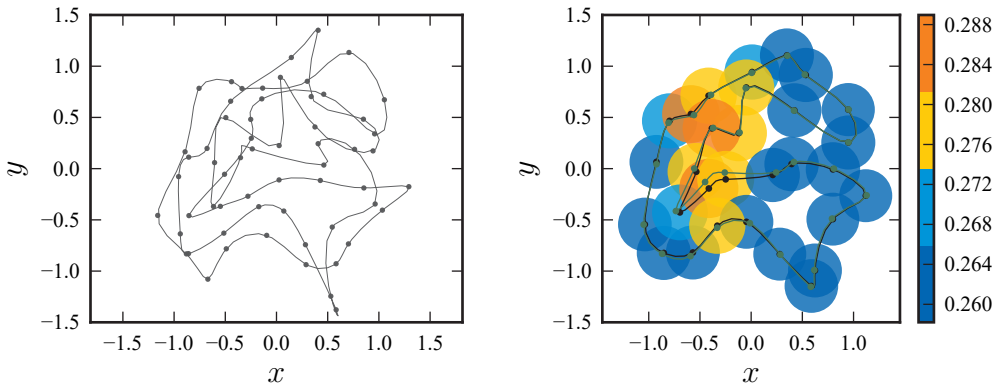

Figure 7. *The left-hand plot shows two sets of landmarks. The right-hand plot shows two versions of the average landmarks. The black shape is calculated using the second splitting prior (with parameters* $\lambda = 0.1$, $\beta = 25$*) and assuming landmarks are known to* $\mathrm{N}(\mathbf{0}, \delta^2 I)$ *errors (*$\delta^2 = 0.005$*, so* $\delta \approx 0.07$*). The discs indicate one standard deviation of the posterior distribution (via the Laplace/Gauss–Newton approximation). The dark green shape is an arithmetic average.*

The reverse limits are degenerate: as $\lambda \to \infty$, $\tilde{\boldsymbol{p}}_{1/2}$ and $\boldsymbol{p}_{1/2}$ are not coupled and may be chosen independently. In particular, the second of the data terms always can be made zero. The remaining terms are minimized by taking $\boldsymbol{q}_{1/2} = \boldsymbol{q}^r$ and $\boldsymbol{p}_{1/2} = 0$. For the limit as the noise grows and overwhelms the system, $\beta \to 0$, there is no Hamiltonian or momenta coupling, and only data terms remain. Then $\boldsymbol{q}_{1/2}$ can be placed anywhere, as the momenta can be chosen arbitrarily without cost. This case has a very shallow energy landscape and $\boldsymbol{q}_{1/2}$ is not well determined. Both of these are outside the regime used in the derivation of the approximation (3.6).

When the terms are balanced, the optimization must achieve some accuracy in flowing to the landmark points, coupling the momenta, and moderation of the energy in H. We see in Figure 10 examples where the arithmetic average and MAP average are very different.

4.3.1. Computations with two landmark sets. The MAP point can be found using unconstrained numerical optimization. The objective function is more complicated this time, due to the matrix exponential $e^{-\lambda \mathcal{G}(\boldsymbol{q}_{1/2})}$ and the required derivative of the matrix exponential (for gradient-based optimization methods). These functions are available in Python's SciPy library, among others. The Laplace method can be applied, again using Gauss–Newton approximations and removal of negative eigenvalues, to determine an approximation to the covariance matrix of the posterior distribution.

To define an average of two sets of landmarks $\boldsymbol{q}^{a,b}$, we choose $\boldsymbol{q}^r = \boldsymbol{q}^a$ and $\boldsymbol{q}^t = \boldsymbol{q}^b$ and find the MAP point $(\boldsymbol{p}_{1/2}, \boldsymbol{q}_{1/2}, \tilde{\boldsymbol{p}}_{1/2})$. The landmarks $\boldsymbol{q}^* = \boldsymbol{q}_{1/2}$ are used as the average of $\boldsymbol{q}^r$ and $\boldsymbol{q}^t$. An example of the resulting average is compared to the arithmetic average in Figure 7.

4.3.2. Generalization to multiple landmark sets. We generalize the second splitting prior to allow for more landmark sets and thereby define an average of multiple landmark sets. Let $\boldsymbol{q}^* \in B \subset \mathbb{R}^{dN}$ be the desired average, and let $\boldsymbol{p}^* \in \mathbb{R}^{dN}$ be an associated momentum. For the prior distribution, we assume $[\boldsymbol{p}^*, \boldsymbol{q}^*]$ follow the Gibbs distribution. Let $\boldsymbol{q}^j \in B \subset \mathbb{R}^{dN}$

for $j = 1, \ldots, J$ denote the given data set of landmarks, and associate to each a momentum $\boldsymbol{p}^j$. We couple each $\boldsymbol{p}^j$ to $[\boldsymbol{p}^*, \boldsymbol{q}^*]$ via the time-one evolution of (3.4). With Gaussian errors in the approximation of the data $\boldsymbol{q}^j$ by the time-half evolution of the Hamiltonian system from $[\boldsymbol{p}^j, \boldsymbol{q}^*]$, this leads to the following objective function for the MAP point:

$$F(\boldsymbol{p}^*, \boldsymbol{q}^*, \boldsymbol{p}^j) \coloneqq \beta H(\boldsymbol{p}^*, \boldsymbol{q}^*) + \frac{\beta}{4\lambda} \sum_{j=1}^{J} \left\| \boldsymbol{p}^j - e^{-\lambda \mathcal{G}(\boldsymbol{q}^*)} \boldsymbol{p}^* \right\|^2 + \frac{1}{2\delta^2} \sum_{j=1}^{J} \left\| \boldsymbol{q}^j - S_q(1/2; 0, [\boldsymbol{p}^j, \boldsymbol{q}^*]) \right\|^2. \tag{4.4}$$

There are $J + 1$ momenta, and this objective does not reduce to (4.3), which depends on two momenta for $J = 2$ landmark sets (see Figure 8). The limit as $\lambda \to 0$ is different, and $\boldsymbol{q}^*$ cannot converge to the midpoint on the paths, as there is no such thing as a single flow between the landmark points for $J > 2$. The extra momentum $\boldsymbol{p}^*$ is introduced as a substitute and provides a means of coupling the deformation for each landmark set to a single momentum. In contrast, as we now show, the limiting behavior as $\beta \to \infty$ resembles the two-landmark average found by studying (4.3).

Theorem 4. *Let $[\boldsymbol{p}^*, \boldsymbol{q}^*, \boldsymbol{p}^j]$ denote the minimizer of (4.4). Suppose that*

1. *$G(\boldsymbol{q}_i)$ is uniformly bounded over $\boldsymbol{q}_i \in B \subset \mathbb{R}^d$ and $\lambda \to 0$, or*
2. *$\mathcal{G}(\boldsymbol{q})$ is uniformly positive definite over $\boldsymbol{q} \in B^N \subset \mathbb{R}^{dN}$ and $\beta \to \infty$.*

In the limit, $\boldsymbol{q}^*$ converges to the arithmetic average $(\boldsymbol{q}^1 + \cdots + \boldsymbol{q}^J)/J$.

Proof. The argument for $\beta \to \infty$ is the same as Corollary 3. We concentrate on the case $\lambda \to 0$. By arguing similarly to Corollary 2, $\min F$ and $(\beta/\lambda)\|\boldsymbol{p}^j - e^{-\lambda \mathcal{G}(\boldsymbol{q}^*)} \boldsymbol{p}^*\|^2$ are bounded as $\lambda \to 0$. Hence, $\boldsymbol{p}^j - e^{-\lambda \mathcal{G}(\boldsymbol{q}^*)} \boldsymbol{p}^* \to \boldsymbol{0}$, and because entries of $\mathcal{G}$ are bounded, we know that $\boldsymbol{p}^j - \boldsymbol{p}^* \to \boldsymbol{0}$. We can minimize the two remaining terms separately: $\beta H(\boldsymbol{p}^*, \boldsymbol{q}^*)$ is minimized by $\boldsymbol{p}^* = \boldsymbol{0}$ and the data term is minimized when $S_q(1/2; 0, [\boldsymbol{p}^j, \boldsymbol{q}^*])$ equals the arithmetic average. This is achieved when $\boldsymbol{p}^j = \boldsymbol{p}^* = \boldsymbol{0}$ and $\boldsymbol{q}^*$ is the arithmetic average. ∎

The methodology for this objective is similar to (4.3): the minimum is found by unconstrained numerical optimization, and $\boldsymbol{q}^*$ is used as an average. The Hessian can be evaluated at the MAP point to define an approximate posterior covariance matrix.

An example of the resulting average for 16 samples is compared to the arithmetic average in Figure 9. The standard deviation is reduced in comparison to Figure 7, from the range $[0.26, 0.29]$ down to $[0.15, 0.18]$, which is roughly a factor 1.6 decrease from a factor 8 increase in the number of samples, and less than expected from the central limit theorem. Figure 10 shows computations of 64 and 256 samples from the same distribution of landmark sets. The distinction between arithmetic and MAP averages is even stronger. The standard deviations are moderately reduced, compared to the factor of 2 expected from a factor 4 increase in the number of samples.

The final example in Figure 11 shows how the MAP average moves closer to the arithmetic average when the value of β is increased from $\beta = 50$ to $\beta = 100$, as discussed in Theorem 4.

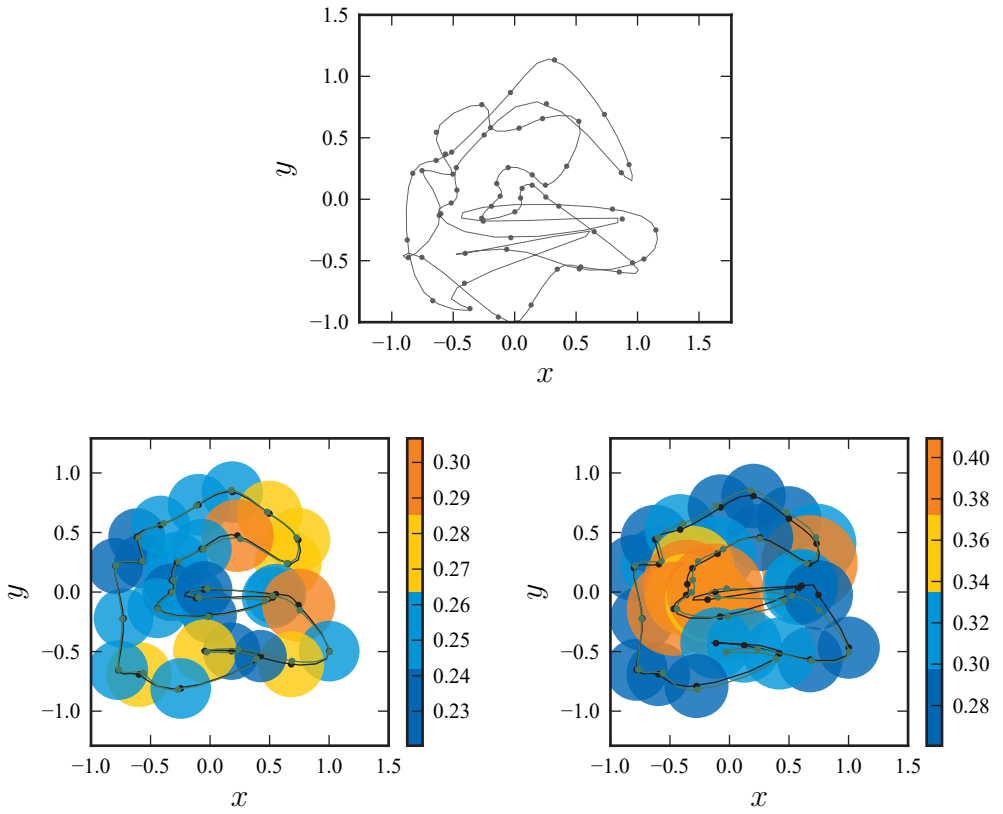

Figure 8. *For the two sets of landmarks, the plots show averages (black lines) according to (4.3) (left) and (4.4) (right) with colored discs showing one standard deviation. Both are close to the arithmetic average, shown in green, with the multiset objective function not as close and having large standard deviations.*

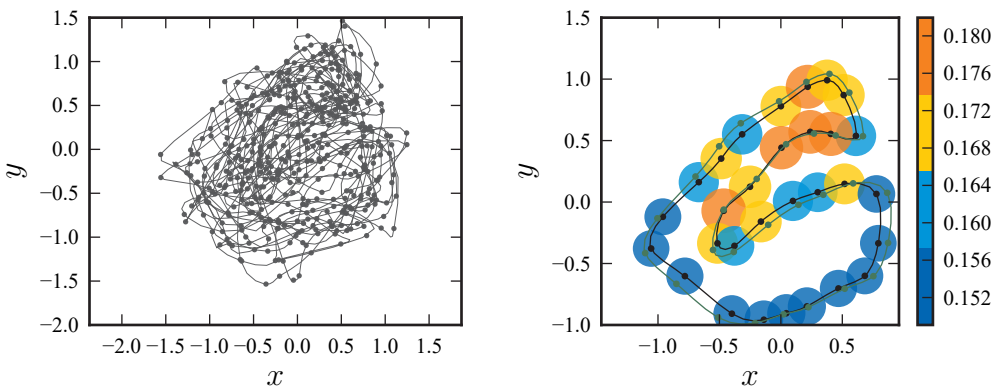

Figure 9. *This is similar to Figure 7, except* 16 *landmark sets are taken, and the MAP average is computed using the objective function (4.4).*

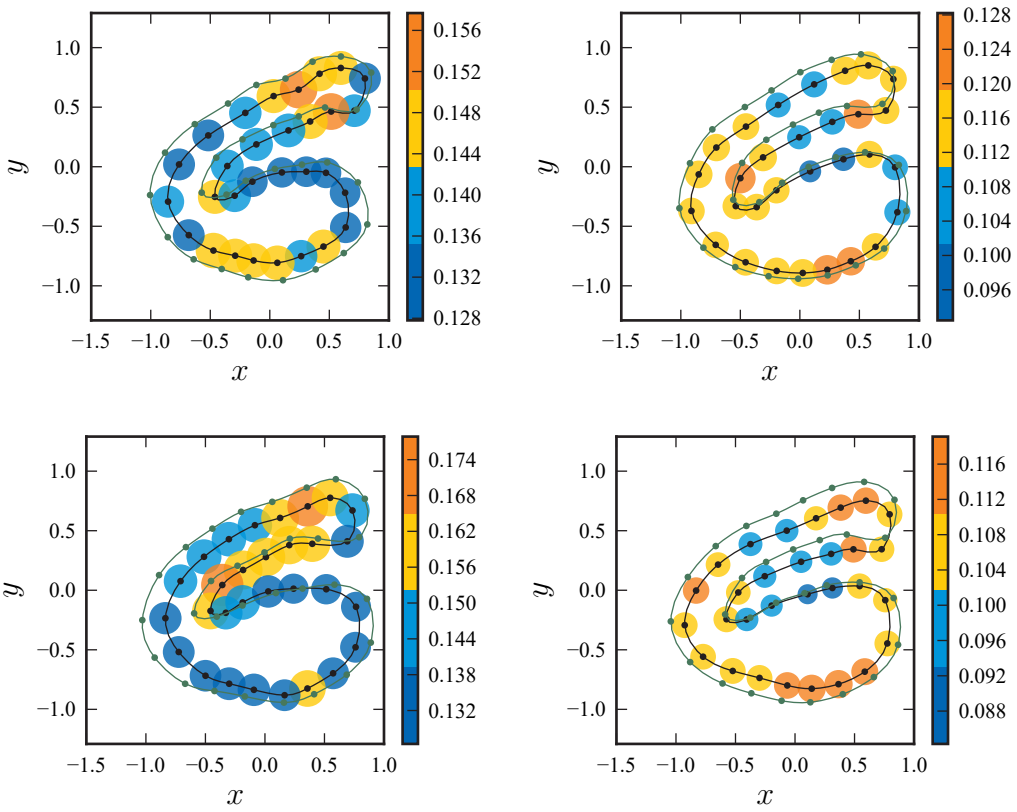

Figure 10. *Here we show two computations for the average of* 64 *(left) and* 256 *(right) independent samples, using the second splitting prior with* $\lambda = 0.1$ *and* $\beta = 25$ *(black line) and the arithmetic average (green line). The rows are calculations of the same averages for independent samples. The colors indicate one standard deviation of the computed posterior distribution.*

5. Conclusion. This paper introduces a type of Langevin equation for performing image registration by landmarks in the presence of uncertainty. The Langevin equation is used to define a prior distribution on the set of diffeomorphisms. It is computationally difficult to sample the diffusion bridge for the Langevin equation. To allow for computation, we introduced three approximate prior distributions: the first by linearizing the Langevin equation about the solution of a Hamiltonian problem, and the second and third by splitting the generator and using a Baker–Campbell–Hausdorff-type approximation to the solution of the Fokker–Planck equation. We give computational examples using the MAP point and Laplace method to find approximate variances for the posterior distribution.

The second splitting prior lends itself to formulating an average of two landmark sets. We defined the average of two landmark sets via the prior and studied the limits as the inverse temperature $\beta \to \infty$ (corresponding to the arithmetic average) and dissipation $\lambda \to 0$ (corresponding to the midpoint of the registration identified by the MAP point for the first splitting prior). This was extended to define an average for multiple landmark sets, with examples provided for both two and multiple landmark sets.

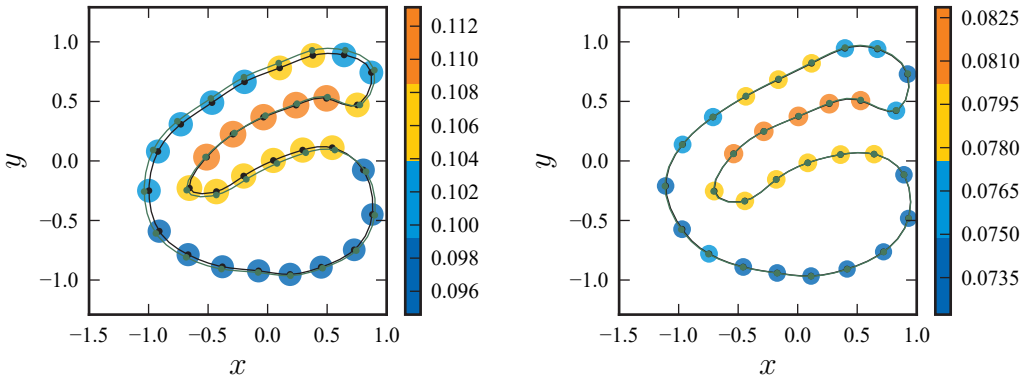

Figure 11. *The plots show the averages (black) provided by the MAP point for $\lambda = 0.1$ with $\beta = 50$ (left) and $\beta = 100$ (right) in comparison to the arithmetic average (green) for 64 landmark sets. As shown in Theorem 4, the averages become closer as β is increased.*

The work was limited by the current technology for sampling hypoelliptic diffusion bridges, and it will be interesting to see how this area develops.

Another avenue of future work is incorporating invariants into the prior distribution, such as conservation of average landmark position. The Langevin equation can be adjusted so that the dynamics live on a subspace of $\mathbb{R}^{2dN}$ where the Gibbs distribution may be a probability measure and landmark average is invariant. The following variation of (2.1) has invariant measure $\exp(-\beta H)$ and satisfies $\frac{d}{dt}\sum \boldsymbol{p}_i = \mathbf{0}$ for isotropic G:

$$
\begin{aligned}
d\boldsymbol{p}_i &= \left[-\lambda \sum_{j\neq i} w(q_{ij})^2 \hat{\boldsymbol{q}}_{ij}\hat{\boldsymbol{q}}_{ij} \cdot \nabla_{\boldsymbol{p}_i} H - \nabla_{\boldsymbol{q}_i} H \right] dt + \sigma \sum_{j\neq i} w(q_{ij}) \hat{\boldsymbol{q}}_{ij} dW_{ij}(t), \\
\frac{d\boldsymbol{q}_i}{dt} &= \nabla_{\boldsymbol{p}_i} H.
\end{aligned} \tag{5.1}
$$

Here $\hat{\boldsymbol{q}}_{ij}$ is the interparticle unit vector and $q_{ij} = \|\boldsymbol{q}_i - \boldsymbol{q}_j\|$. This time, $W_{ij}(t)$ are i.i.d. scalar Brownian motions for $i < j$ and $W_{ij} = W_{ji}$. Here $w\colon \mathbb{R} \to \mathbb{R}^+$ is a coefficient function, which could be identically equal to one for simplicity. For given $\bar{\boldsymbol{p}} \in \mathbb{R}^d$, we see that $\exp(-\beta H)$ is an invariant measure on the subspace of $\mathbb{R}^{2dN}$ with $\frac{1}{N}\sum_{i=1}^N \boldsymbol{p}_i = \bar{\boldsymbol{p}}$ (the center of mass is invariant for $\bar{\boldsymbol{p}} = \mathbf{0}$). This can be shown to be invariant by using the Fokker–Planck equation as above, with λ and σ replaced by position-dependent coefficients that still cancel out under the fluctuation–dissipation relation. See [7, 26, 27].

Appendix A. Linearized equations. We write down equations to compute the mean and covariance, using backward and forward Euler approximations. Suppose that $\boldsymbol{\delta}_{n_1} \sim \mathrm{N}(\mathbf{0}, C_1)$ for some given C_1. We wish to calculate the joint distribution of $\boldsymbol{\delta}_n$ for $n = 0, \ldots, N_{\Delta t}$. This is easy to do, as the joint distribution is Gaussian and we derive update rules for the mean and covariance: from

$$
\boldsymbol{\delta}_{n+1} = M_n^+ \boldsymbol{\delta}_n + \boldsymbol{A}_n + \begin{pmatrix} \sigma \Delta \boldsymbol{W}_n \\ 0 \end{pmatrix},
$$

we get an update rule for the mean

$$\boldsymbol{\mu}_{n+1} = \mathbb{E}\big[\boldsymbol{\delta}_{n+1}\big] = M_n^+ \boldsymbol{\mu}_n + \boldsymbol{A}_n.$$

Similarly, when time-stepping backward,

$$\boldsymbol{\mu}_{n-1} = \mathbb{E}\big[\boldsymbol{\delta}_{n-1}\big] = M_n^- \boldsymbol{\mu}_n + \boldsymbol{A}_n.$$

For the covariance update along the diagonal moving forward,

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{\delta}_{n+1}\boldsymbol{\delta}_{n+1}^{\mathsf{T}}\big] &= \mathbb{E}\left[\Big(M_n^+\boldsymbol{\delta}_n + \boldsymbol{A}_n\Big)\Big(M_n^+\boldsymbol{\delta}_n + \boldsymbol{A}_n\Big)^{\mathsf{T}}\right] + \begin{pmatrix}\sigma h I_{dN} & 0\\ 0 & 0\end{pmatrix}\\ &= M_n^+\mathbb{E}\big[\boldsymbol{\delta}_n\boldsymbol{\delta}_n^{\mathsf{T}}\big]M_n^{+\mathsf{T}} + \boldsymbol{A}_n\boldsymbol{\mu}_{n+1}^{\mathsf{T}} + \boldsymbol{\mu}_{n+1}\boldsymbol{A}_n^{\mathsf{T}} - \boldsymbol{A}_n\boldsymbol{A}_n^{\mathsf{T}} + \begin{pmatrix}\sigma h I_{dN} & 0\\ 0 & 0\end{pmatrix}.\end{aligned}$$

Similarly, moving backward,

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{\delta}_{n-1}\boldsymbol{\delta}_{n-1}^{\mathsf{T}}\big] &= \mathbb{E}\left[\Big(M_n^-\boldsymbol{\delta}_n + \boldsymbol{A}_n\Big)\Big(M_n^-\boldsymbol{\delta}_n + \boldsymbol{A}_n\Big)^{\mathsf{T}}\right] + \begin{pmatrix}\sigma h I_{dN} & 0\\ 0 & 0\end{pmatrix}\\ &= M_n^-\mathbb{E}\big[\boldsymbol{\delta}_n\boldsymbol{\delta}_n^{\mathsf{T}}\big]M_n^{-\mathsf{T}} + \boldsymbol{A}_n\boldsymbol{\mu}_{n-1}^{\mathsf{T}} + \boldsymbol{\mu}_{n-1}\boldsymbol{A}_n^{\mathsf{T}} - \boldsymbol{A}_n\boldsymbol{A}_n^{\mathsf{T}} + \begin{pmatrix}\sigma h I_{dN} & 0\\ 0 & 0\end{pmatrix}.\end{aligned}$$

The remaining parts of the matrix $\mathbb{E}\big[\boldsymbol{\delta}_j\boldsymbol{\delta}_k^{\mathsf{T}}\big]$ can be computed by sideways moves along either a row or column using the following rules: if $k \geq j$, then

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{\delta}_j\boldsymbol{\delta}_{k+1}^{\mathsf{T}}\big] &= \mathbb{E}\big[\boldsymbol{\delta}_j\boldsymbol{\delta}_k^{\mathsf{T}}\big]M_k^{+\mathsf{T}} + \boldsymbol{\mu}_j\boldsymbol{A}_k^{\mathsf{T}},\\ \mathbb{E}\big[\boldsymbol{\delta}_{k+1}\boldsymbol{\delta}_j^{\mathsf{T}}\big] &= M_k^+\mathbb{E}\big[\boldsymbol{\delta}_k\boldsymbol{\delta}_j^{\mathsf{T}}\big] + \boldsymbol{A}_k\boldsymbol{\mu}_j^{\mathsf{T}},\end{aligned}$$

and if $k \leq j$, then

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{\delta}_j\boldsymbol{\delta}_{k-1}^{\mathsf{T}}\big] &= \mathbb{E}\big[\boldsymbol{\delta}_j\boldsymbol{\delta}_k^{\mathsf{T}}\big]M_k^{-\mathsf{T}} + \boldsymbol{\mu}_j\boldsymbol{A}_k^{\mathsf{T}},\\ \mathbb{E}\big[\boldsymbol{\delta}_{k-1}\boldsymbol{\delta}_j^{\mathsf{T}}\big] &= M_k^-\mathbb{E}\big[\boldsymbol{\delta}_k\boldsymbol{\delta}_j^{\mathsf{T}}\big] + \boldsymbol{A}_k\boldsymbol{\mu}_j^{\mathsf{T}}.\end{aligned}$$

Finally, $\mathrm{Cov}(\boldsymbol{\delta}_j, \boldsymbol{\delta}_n) = \mathbb{E}\big[\boldsymbol{\delta}_j\boldsymbol{\delta}_n^{\mathsf{T}}\big] - \boldsymbol{\mu}_j\boldsymbol{\mu}_n^{\mathsf{T}}$.

REFERENCES

[1] A. Arnaudon, D. D. Holm, A. Pai, and S. Sommer, *A Stochastic Large Deformation Model for Computational Anatomy*, preprint, https://arxiv.org/abs/1612.05323v1, 2016.

[2] M. Bladt, S. Finch, and M. Sørensen, *Simulation of multivariate diffusion bridges*, J. R. Stat. Soc. Ser. B Stat. Methodol., 78 (2016), pp. 343–369, https://doi.org/10.1111/rssb.12118.

[3] F. L. Bookstein, *Principal warps: Thin-plate splines and the decomposition of deformations*, IEEE Trans. Pattern Anal. Mach. Intell., 11 (1989), pp. 567–585, https://doi.org/10.1109/34.24792.

[4] C. J. Cotter, S. L. Cotter, and F.-X. Vialard, *Bayesian data assimilation in shape registration*, Inverse Problems, 29 (2013), 045011, https://doi.org/10.1088/0266-5611/29/4/045011.

[5] B. Delyon and Y. Hu, *Simulation of conditioned diffusion and application to parameter estimation*, Stochast. Process. Appl., 116 (2006), pp. 1660–1675, https://doi.org/10.1016/j.spa.2006.04.004.

[6] J. Duchon, *Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces*, Rev. Française Automat. Informat. Recherche Opérationelle Sér., 10 (1976), pp. 5–12.
[7] P. Español and P. Warren, *Statistical mechanics of dissipative particle dynamics*, Europhys. Lett., 30 (1995), pp. 191–196, https://doi.org/10.1209/0295-5075/30/4/001.
[8] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer, Berlin, 2012, https://doi.org/10.1007/978-3-642-25847-3.
[9] A. Golightly and D. J. Wilkinson, *Bayesian inference for nonlinear multivariate diffusion models observed with error*, Comput. Stat. Data Anal., 52 (2008), pp. 1674–1693, https://doi.org/10.1016/j.csda.2007.05.019.
[10] M. Hairer, A. M. Stuart, and J. Voss, *Analysis of SPDEs arising in path sampling part* II: *The nonlinear case*, Ann. Appl. Probab., 17 (2007), pp. 1657–1706, https://doi.org/10.1214/07-aap441.
[11] M. Hairer, A. M. Stuart, and J. Voss, *Sampling conditioned hypoelliptic diffusions*, Ann. Appl. Probab., 21 (2011), pp. 669–698, https://doi.org/10.1214/10-AAP708.
[12] D. D. Holm, *Variational principles for stochastic fluid dynamics*, Proc. A, 471 (2015), 20140963, https://doi.org/10.1098/rspa.2014.0963.
[13] D. D. Holm and J. E. Marsden, *Momentum maps and measure-valued solutions (peakons, filaments, and sheets) for the EPDiff equation*, in The Breadth of Symplectic and Poisson Geometry, Progr. Math. 232, J. E. Marsden and T. S. Ratiu, eds., Birkhäuser Boston, Boston, MA, 2005, pp. 203–235, https://doi.org/10.1007/0-8176-4419-9_8.
[14] D. D. Holm, J. T. Ratnanather, A. Trouvé, and L. Younes, *Soliton dynamics in computational anatomy*, Neuroimage, 23 Suppl. 1 (2004), pp. S170–S178, https://doi.org/10.1016/j.neuroimage.2004.07.017.
[15] D. D. Holm and T. M. Tyranowski, *Variational principles for stochastic soliton dynamics*, Proc. A, 472 (2016), 20150827, https://doi.org/10.1098/rspa.2015.0827.
[16] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer, Berlin, 1988, https://doi.org/10.1007/978-1-4684-0302-2.
[17] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. 23, Springer, Berlin, 1992, https://doi.org/10.1007/978-3-662-12616-5.
[18] S. Marsland and T. Shardlow, *Python Codes for Langevin Equations for Landmark Image Registration with Uncertainty*, GitHub Repository, 2016, https://doi.org/10.5281/zenodo.220875, https://github.com/tonyshardlow/reg_sde.
[19] S. Marsland and C. J. Twining, *Clamped-plate splines and the optimal flow of bounded diffeomorphisms*, in Statistics of Large Datasets, Proceedings of Leeds Annual Statistical Research Workshop, 2002, pp. 91–95.
[20] S. Marsland and C. J. Twining, *Constructing diffeomorphic representations for the groupwise analysis of nonrigid registrations of medical images*, IEEE Trans. Med. Imaging, 23 (2004), pp. 1006–1020, https://doi.org/10.1109/TMI.2004.831228.
[21] R. I. McLachlan and S. Marsland, *N-particle dynamics of the Euler equations for planar diffeomorphisms*, Dyn. Syst., 22 (2007), pp. 269–290, https://doi.org/10.1080/14689360701191931.
[22] A. Mills and T. Shardlow, *Analysis of the geodesic interpolating spline*, Eur. J. Appl. Math., 19 (2008), pp. 519–539, https://doi.org/10.1017/S0956792508007493.
[23] A. Mills, T. Shardlow, and S. Marsland, *Computing the geodesic interpolating spline*, in Proceedings of Biomedical Image Registration: Third International Workshop (WBIR 2006, Utrecht, The Netherlands), Lecture Notes in Comput. Sci. 4057, J. P. W. Pluim, B. Likar, and F. A. Gerritsen, eds., Springer, Berlin, 2006, pp. 169–177, https://doi.org/10.1007/11784012_21.
[24] J. Modersitzki, *Numerical Methods for Image Registration*, Oxford University Press, Oxford, UK, 2003, https://doi.org/10.1093/acprof:oso/9780198528418.001.0001.
[25] O. Papaspiliopoulos and G. Roberts, *Importance sampling techniques for estimation of diffusion models*, in Statistical Methods for Stochastic Differential Equations, Monogr. Statist. Appl. Probab., M. Sørensen, ed., Chapman and Hall/CRC, Boca Raton, FL, 2012, pp. 311–340, https://doi.org/10.1201/b12126-5.
[26] T. Shardlow, *Splitting for dissipative particle dynamics*, SIAM J. Sci. Comput., 24 (2003), pp. 1267–1282, https://doi.org/10.1137/S1064827501392879.

[27] T. Shardlow and Y. Yan, *Geometric ergodicity for dissipative particle dynamics*, Stoch. Dyn., 6 (2006), pp. 123–154, https://doi.org/10.1142/S0219493706001670.

[28] C. Soize, *The Fokker–Planck Equation for Stochastic Dynamical Systems and Its Explicit Steady State Solutions*, Ser. Adv. Math. Appl. Sci. 17, World Scientific, River Edge, NJ, 1994, https://doi.org/10.1142/9789814354110_0006.

[29] A. Trouvé and F.-X. Vialard, *Shape splines and stochastic shape evolutions: A second order point of view*, Quart. Appl. Math., 70 (2012), pp. 219–251, https://doi.org/10.1090/s0033-569x-2012-01250-4.

[30] L. Younes, *Shapes and Diffeomorphisms*, Springer, Berlin, 2010, https://doi.org/10.1007/978-3-642-12055-8.