

# A GEOMETRIC FRAMEWORK FOR STOCHASTIC SHAPE ANALYSIS

ALEXIS ARNAUDON, DARRYL D HOLM, AND STEFAN SOMMER

**ABSTRACT.** We introduce a stochastic model of diffeomorphisms, whose action on a variety of data types descends to stochastic models of shapes, images and landmarks. The stochasticity is introduced in the vector field which transports the data in the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework for shape analysis and image registration. The stochasticity thereby models errors or uncertainties of the flow in following the prescribed deformation velocity. The approach is illustrated in the example of finite dimensional landmark manifolds, whose stochastic evolution is studied both via the Fokker-Planck equation and by numerical simulations. We derive two approaches for inferring parameters of the stochastic model from landmark configurations observed at discrete time points. The first of the two approaches matches moments of the Fokker-Planck equation to sample moments of the data, while the second approach employs an Expectation-Maximisation based algorithm using a Monte Carlo bridge sampling scheme to optimise the data likelihood. We derive and numerically test the ability of the two approaches to infer the spatial correlation length of the underlying noise.

## CONTENTS

1. Introduction	1
2. Stochastic Large Deformation Matching	5
3. Exact stochastic Landmark Matching	8
4. Estimating the Spatial Correlation of the Noise	15
5. Numerical examples	20
6. Discussion and Outlook	25
References	29
Appendix A. Bridge Sampling	31
Appendix B. Moment equation for stochastic landmark	34

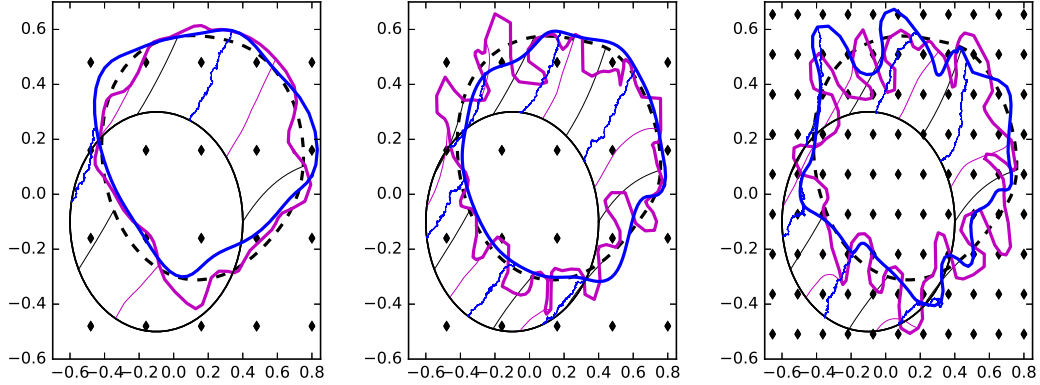
## 1. INTRODUCTION

In this work, we aim at modelling variability of shapes using a theory of stochastic perturbations consistent with the geometry of the diffeomorphism group corresponding to the Large Deformation Diffeomorphic Metric Mapping framework

(LDDMM, see [You10]). In applications, such variability arises and can be observed, for example, when human organs are influenced by disease processes, as analysed in computational anatomy [YAM09]. Spatially independent white noise contains insufficient information to describe these large-scale variabilities of shapes. In addition, the coupling of the spatial correlations of the noise must be adapted to a variety of transformation properties of the shape spaces. The theory developed here addresses this problem by introducing spatially correlated noise which respects the geometric structure of the data. This method provides a new way of characterising stochastic variability of shapes using spatially correlated noise in the context of the standard LDDMM framework.

We will show that this specific type of noise can be used for all the data structures that the LDDMM framework applies to. The LDDMM theory was initiated by [Tro95, CRM96, DGM98, MTY02, BMTY05] building on the pattern theory of [Gre94]. LDDMM models the dynamics of shapes by the action of diffeomorphisms on shape spaces. It gives a unified approach to shape modelling and shape analysis, valid for a range of structures such as landmarks, curves, surfaces, images, densities or even tensor-valued images. For any such data structure, the optimal shape deformations are described via the Euler-Poincaré equation of the diffeomorphism group, usually referred to as the EPDiff equation [HMR98, HM05, YAM09]. In this work, we will show how to obtain a stochastic EPDiff equation valid for any data structure, and in particular for the finite dimensional representation of images based on landmarks. For this, we will follow the LDDMM derivation of [BGBHR11] based on geometric mechanics [MR99, Hol11] and the existence of momentum maps to represent images and shapes. The momentum maps are the key structures to retain after the introduction of noise into the EPDiff equation, as they will allow us to use most of the technology developed for shape analysis in the deterministic context and in computational anatomy.

This work is not the first to consider stochastic evolutions in LDDMM. Indeed, [TV12, Via13] and more recently [MS16] already investigated the possibility of stochastic perturbations of landmark dynamics. In these works, the noise is introduced into the momentum equation, as though it were an external random force acting on each landmark independently. In [MS16], an extra dissipative force was added to balance the energy input from the noise and to make the dynamics correspond to a certain type of heat bath used in statistical physics. Here, we will instead introduce an Eulerian noise directly into the reconstruction relation used to find the deformation flows from the velocity fields, which are solutions of the EPDiff equation [HM05, You10]. As we will see, this derivation of stochastic models is compatible with variational principles, preserves the momentum map structure and yields a stochastic EPDiff equation with a novel type of multiplicative noise, depending on the gradient of the solution, as well as its magnitude. This model is based on the previous works [Hol15, ACH16], where, respectively, stochastic perturbations of infinite and finite dimensional mechanical systems were considered. The Eulerian nature of the noise discussed here implies that the noise correlation depends on the image position and not, as for example in [TV12, MS16], on the landmarks themselves. This property will explain why the noise is compatible with any data structure while retaining the freedom in the choice of its spatial correlation.



(a) Low resolution and large noise correlation (100 landmarks,  $6 \times 6$  noise fields) (b) High resolution and large noise correlation (200 landmarks,  $6 \times 6$  noise fields) (c) High resolution and small noise correlation (200 landmarks,  $12 \times 12$  noise fields)

FIGURE 1. In this figure, we compare the deterministic evolution of landmarks arranged in an ellipse (black line) with a translated ellipse as final position (black dashed line), to two different stochastically perturbed evolutions. The radius for the landmark kernel is twice their average initial distances. In blue is the stochastic perturbation developed in this paper. The black dots represent the  $J$  Eulerian noise fields arranged in a grid configuration. In magenta is the evolution resulting from additive noise in the momentum equation, different for each landmark but with the same amplitude as the Eulerian noise. We run three initial value simulations to compare the limit of a large number of landmarks and small noise correlation. The Eulerian noise model (blue) is robust to the continuum limit and can reproduce the general behaviour of the additive noise model. Furthermore, the choice of the noise fields provides an additional freedom in parameterisation which will be studied and exploited in this work.

To illustrate this framework and give an early demonstration of the stochastic landmark dynamics, we display in Figure 1 three experiments which compare the proposed model with a stochastic forcing model, such as studied for example by [TV12]. The proposed model is the stochastic Hamiltonian system for the positions of the landmarks  $\mathbf{q}_i$  and their canonical momenta  $\mathbf{p}_i$

$$\begin{aligned} d\mathbf{q}_i &= \frac{\partial h}{\partial \mathbf{p}_i} dt + \sum_l \sigma_l(\mathbf{q}_i) \circ dW_t^l, \\ d\mathbf{p}_i &= -\frac{\partial h}{\partial \mathbf{q}_i} dt - \sum_l \frac{\partial}{\partial \mathbf{q}_i} (\mathbf{p}_i \cdot \sigma_l(\mathbf{q}_i)) \circ dW_t^l. \end{aligned} \tag{1.1}$$

Here, the  $\sigma_i$  are given functions of space which represent the spatial correlations of the noise. In Figure 1, the  $\sigma_i$  fields are Gaussian fields whose variance is equal to twice their separation distance and locations are indicated by black dots. We

compare this model with

$$d\mathbf{q}_i^\alpha = \frac{\partial h}{\partial \mathbf{p}_i^\alpha} dt \quad \text{and} \quad d\mathbf{p}_i^\alpha = -\frac{\partial h}{\partial \mathbf{q}_i^\alpha} dt + \sigma dW_t^i, \quad (1.2)$$

where  $\sigma$  is a constant. In this case, the noise corresponds to a stochastic force acting on the landmarks, whose corresponding Brownian motion is different for each landmark. We show on the first panel of Figure 1 that for a small number of landmarks and a large range of spatial correlations of the noise, both types of stochastic deformations in (1.1) and (1.2) visually coincide. This is shown for a simple experiment in translating a circle (from the black circle to the black dashed circle). By doubling the number of landmarks (middle panel), the dynamics of (1.2) results in small-scale noise correlation (magenta), whereas the proposed model (blue) remains equivalent to the first experiment. This figure illustrates shape evolution when the noise is Eulerian and independent of the data structure. Indeed, the limit of a large number of landmarks corresponds to a certain continuum limit, in this case corresponding to curve dynamics. Finally, in the right-most panel, we reduce the range of the spatial correlation of the noise by adding more noise fields. This arrangement allows us to qualitatively reproduce the dynamics of the equation (1.2) with the same number of landmarks.

Modelling large-scale shape variability with noise is of interest for applications in computational anatomy where sources of variability include natural ageing, the influence of diseases such as Alzheimer’s disease, and intra-subject population scale variations. In the LDDMM context, these effects are sometimes modelled using the random orbit model [MBC<sup>+</sup>97]. The random orbit approach models variability in the observed data by using an ensemble of initial velocities in matching a template to a set of observations via geodesic flows, see [VMYT04]. The randomness is confined to the initial velocity as opposed to the evolving stochastic processes that we will use here. A prior can be defined by assuming a distribution of the initial velocities, and Bayesian approaches can then be used for inference of the template shape as well as additional unknown parameters [AAT07, MMTY08, ZSF13]. The stochastic model developed here can also be applied to model random warps and generate distributions used in Bayesian shape modelling, and for coupling warps and functional variations such as for example in [RSM14, KSPR17]. Indeed, because the proposed probabilistic approach assigns a likelihood to random deformations, the model can be used for general likelihood-based inference tasks.

In the present model, the observed shape variability indicates the spatial correlation of the noise. As this correlation is generally unknown, estimating the parameters of the correlation structure becomes an important part of the framework. We will address this problem by considering two different methods. One is based on the estimation of the time evolution of the probability distribution of each landmark. For this method, we will derive a set of differential equations approximating the time evolution of the complete distribution via its first moments. We can then solve the inverse problem of estimating the noise correlation from a known initial and final distribution of landmarks by minimization of a certain cost function, solved using a genetic algorithm. The other method is based on an Expectation-Maximisation (EM) algorithm which can infer unknown parameters for a parametric statistical

model from observed data. In this context, since only initial and final landmarks positions are observed, the full stochastic trajectories are considered missing information. For this algorithm, we need to estimate the likelihood of stochastic paths connecting sets of observed landmarks. We achieve this by adapting the theory of diffusion bridges to the stochastic landmark equation.

**Plan of this work.** We begin by developing a general theory of stochastic perturbations for inexact matching in section 2. We then focus on exact landmark matching in section 3, which is the simplest example of this theory. In particular, we derive the Fokker-Planck equation in section 3.2 and diffusion bridge simulation in section 3.3. In section 4, we describe the two methods for estimating parameters of the noise from observations. The Fokker-Planck based method is in section 4.2 and the Expectation-Maximisation algorithm in section 4.3. We end the paper with numerical examples in section 5 in which we investigate the effect of the noise on landmark dynamics and compare the two methods for estimating the noise amplitude.

## 2. STOCHASTIC LARGE DEFORMATION MATCHING

In this section, we will first review the geometrical framework of LDDMM, following [BGBHR11], and then introduce noise following [Hol15] to preserve the geometrical structure of LDDMM. The key ingredient for both topics is the momentum map, which we will use as the main tool for reducing the infinite dimensional equation on the diffeomorphism group to equations on shape spaces.

**2.1. The Deterministic LDDMM Model.** Here, we will briefly review the theory of reduction by symmetry, as applied to the LDDMM context, following the presentation of [BGBHR11]. We detail the proof of the formulas below in the next section when we include noise. Define an energy functional  $E$  by

$$E(u_t) = \int_0^1 l(u_t) dt + \frac{1}{2\lambda^2} \|g_1 \cdot I_0 - I_1\|^2, \quad (2.1)$$

where  $I_0, I_1 \in V$  are shapes represented in a vector space  $V$  on which the diffeomorphism group  $\text{Diff}(\mathbb{R}^d)$  acts,  $u_t$  is a time-dependent vector field, and  $\lambda$  is a weight, or tolerance, which allows the matching to be inexact. The flow  $g_t \in \text{Diff}(\mathbb{R}^d)$  corresponding to  $u_t$  is found by solving the reconstruction relation

$$\partial_t g_t = u_t g_t, \quad (2.2)$$

and  $I_0$  is matched against  $I_1$  through the action  $g_1 \cdot I_0$  of  $g_1$  on  $s_0$ . The vector field  $u_t$  can be considered an element of the Lie algebra  $\mathfrak{X}(\mathbb{R}^d)$ . In the case of  $I_0, I_1$  being images  $I : \mathbb{R}^d \rightarrow \mathbb{R}$ , the action is by push-forward,  $g \cdot I = I \circ g^{-1}$ , and when  $I$  represents  $N$  landmarks with positions  $\mathbf{q}_i \in \mathbb{R}^d$ , the action is by evaluation  $g \cdot \mathbf{q} = (g(\mathbf{q}_1), \dots, g(\mathbf{q}_N))$  (see [BGBHR11] for more details). The group elements can act on various additional shape structures such as tensor fields.

**Remark 2.1** (Nonlinear shape structures). *This framework can be extended to structures that are not represented by a vector space  $V$ , such as curves or surfaces. We leave this extension for future work.*

Using the calculus of variations for the functional (2.1) results in the equation of motion for  $u_t$  of the form

$$\frac{d}{dt} \frac{\delta l}{\delta u} + \text{ad}_{u_t}^* \frac{\delta l}{\delta u} = 0, \quad (2.3)$$

which is called the Euler-Poincaré equation. The operation  $\text{ad}^*$  is the coadjoint action of the Lie algebra of vector fields associated to the diffeomorphism group. The operation  $\text{ad}^*$  acts on the variations  $\delta l/\delta u$ , which are 1-form densities, in the dual of the Lie algebra of vector fields, under the  $L^2$  pairing. When  $l(u)$  is a norm, this equation is the geodesic equation for that norm, in the case that  $\lambda = \infty$ ; that is, with exact matching. We will focus on this case later in section 3 when discussing landmark dynamics. Here, the inexact matching term constrains the form of the momentum  $m = \frac{\partial l}{\partial u}$  to depend on the geodesic path. Following the notation of [BGBHR11], the momentum map is defined as

$$m(t) = -\frac{1}{\lambda^2} J_t^0 \diamond (g_{t,1}(J_1^0 - J_1^1)^\flat), \quad (2.4)$$

where  $g_{t,s}$  is the solution of (2.2) at time  $t$  with initial conditions at time  $s$ , while  $J_t^0 = g_{t,0}I_0$  and  $J_t^1 = g_{t,1}I_1$ . The value  $J_1^0$  corresponds to the initial image, pushed forward to time  $t = 1$ , and  $J_1^1 = I_1$  is the target image.

The operations  $\diamond$  and  $\flat$  in the momentum map formula (2.4) are defined, as follows. The Lagrangian  $l$  in (2.1) may be taken as kinetic energy, which defines a scalar product and norm  $l(u) = \langle u, Lu \rangle_{L^2} = \|u\|_{L^2}^2$  on the space of vector fields  $\mathfrak{X}(\mathbb{R}^d)$ . The quantity  $Lu = \delta l/\delta u$  may then be regarded as the momentum conjugate to the velocity  $u$ . Similarly, for the image data space  $V$ , we define the dual space  $V^*$  with the  $L^2$  pairing  $\langle f, I \rangle = \int_{\Omega} f(x)I(x)dx$ , where  $f \in V^*$  and  $\Omega$  is the image domain  $\Omega \in \mathbb{R}^d$ . This identification defines the  $\flat$  operator as  $\flat : V \rightarrow V^*$ . When an element  $g_t$  of the diffeomorphism group acts on  $V$  by push-forward,  $I_t = g_t.I_0 = (g_t)_*I_0$ , the corresponding infinitesimal action of the velocity  $u$  in the Lie algebra of vector fields  $u \in \mathfrak{X}(\mathbb{R}^d)$  is given by  $u.I := [g_t^* \frac{d}{dt} (g_t)_* I_0]_{t=0}$ . In terms of this infinitesimal action, we can then define the operation  $\diamond : V \times V^* \rightarrow \mathfrak{g}^*$  as

$$\langle I \diamond f, u \rangle_{\mathfrak{g} \times \mathfrak{g}^*} := \langle f, u.I \rangle_{V \times V^*}. \quad (2.5)$$

A detailed derivation of this formula for the momentum map can be found in [BGBHR11].

**Remark 2.2** (Solving this equation). *We will just add here the important remark that the relation (2.4) introduces nonlocality into the problem, as the momentum implicitly depends on the value of the group at later times. This is exactly what is needed in order to solve the boundary value problem coming from the matching of images  $I_1$  and  $I_0$ . The optimal vector field can be found with a shooting method or a gradient descent algorithm on the energy functional (2.1), see [BMTY05]. For more information about the relation of the momentum map approach of [BGBHR11] to the LDDMM approach of [BMTY05], see [BH15].*

**2.2. Stochastic Reduction Theory.** The aim here is to introduce noise in the Euler-Poincaré equation (2.3) while preserving the momentum map (2.4) such that the noise descends to the shape spaces. Following [Hol15], we introduce noise in the

reconstruction relation (2.2) and proceed with the theory of reduction by symmetry. We will focus on a noise described by a set of  $J$  real-valued independent Wiener processes  $W_t^i$  together with  $J$  associated vector fields  $\sigma_i \in \mathfrak{X}(\mathbb{R}^d)$  on the data domain. We will later discuss particular forms of these fields and methods for estimating unknown parameters of the fields in the context of landmark matching.

**Remark 2.3** (Dimension of the noise). *We proceed here with a finite number of  $J$  associated vector fields and finite dimensional noise, while leaving possible extension to infinite dimensional noise such as done by [Via13] for later works.*

We replace the reconstruction relation (2.2) by the following stochastic process

$$dg_t = u g_t dt + \sum_i \sigma_i g_t \circ dW_t^i, \quad (2.6)$$

where  $\circ$  denotes Stratonovich integration. That is, the Lie group trajectory  $g_t$  is now a stochastic process. With this noise construction, the previous derivations of (2.3) and (2.4) in [BGBHR11] still apply and we obtain the following result for the stochastic vector field,  $u_t$ .

**Proposition 2.4.** *Under stochastic perturbations of the form (2.6), the momentum map (2.4) persists, and the Euler-Poincaré equation takes the form*

$$d \frac{\delta l}{\delta u} + \text{ad}_{u_t}^* \frac{\delta l}{\delta u} dt + \sum_i \text{ad}_{\sigma_i}^* \frac{\delta l}{\delta u} \circ dW_t^i = 0. \quad (2.7)$$

*Proof.* We first show that the momentum map formula (2.4) persists in the presence of noise. The key step in its computation is to prove the formula in the lemma 2.5 of [BGBHR11] which is given by  $\partial_t(g^{-1}\delta g) = \text{Ad}_g \delta u$ , where Ad is the adjoint action on the diffeomorphism group on its Lie algebra. We first compute the variations of (2.6)

$$\delta dg_t = \delta u g_t dt + u \delta g_t dt + \sum_i \sigma_i \delta g_t \circ dW_t^i,$$

and then prove this formula by a direct computation

$$\begin{aligned} d(g^{-1}\delta g) &= -g^{-1}dg g^{-1}\delta g + g^{-1}d\delta g \\ &= -g^{-1}(u dt + \sum_i \sigma_i \circ dW_t^i) \delta g + g^{-1}(\delta u g_t dt + u \delta g_t dt + \sum_i \sigma_i \delta g_t \circ dW_t^i) \\ &= g^{-1} \delta u g_t dt \\ &:= \text{Ad}_g \delta u dt. \end{aligned}$$

This key formula is the same as in [BMTY05] and [BGBHR11] for the deterministic case. In particular, it does not explicitly depend on the Wiener processes  $W_t^i$ . This ensures that the momentum map formula (2.4) remains the same as in the deterministic case. The last step of the proof is to derive the stochastic Euler-Poincaré equation (2.7). This is done by computing the stochastic evolution of the momentum, given by

$$\frac{\delta l}{\delta u} = \text{Ad}_{g^{-1}}^*(I_0 \diamond (g_1^{-1}\pi)), \quad \text{where} \quad \pi = \frac{1}{\lambda^2}(g_i I_0 - I_1)^{\flat}.$$

The only time dependence is in the coadjoint action, and, by the standard formula

$$d\text{Ad}_{g^{-1}}^* \eta = -\text{ad}_{dgg^{-1}}^* \text{Ad}_g^* \eta,$$

we obtain the result

$$\begin{aligned} d\frac{\delta l}{\delta u} &= -d\text{Ad}_{g^{-1}}^*(I_0 \diamond (g_1^{-1}\pi)) \\ &= \text{ad}_{dgg^{-1}}^* \text{Ad}_{g^{-1}}^*(I_0 \diamond (g_1^{-1}\pi)) \\ &= \text{ad}_{dgg^{-1}}^* \frac{\delta l}{\delta u}, \end{aligned}$$

where we have used the stochastic reconstruction relation (2.6) in the form

$$dgg^{-1} = udt + \sum_i \sigma_i \circ dW_i^t.$$

□

In summary, this stochastic perturbation of the LDDMM framework preserves the form of momentum map (2.4), although it does affect the reconstruction relation (2.6) and the Euler-Poincaré equation (2.7). As shown in [BGBHR11], various data structures fit into this framework including landmarks, images, shapes, and tensor fields. In practice, for inexact matching, a gradient descent algorithm can be used to minimise the energy functional (2.1). The noise will only appear in the evaluation of the matching cost via the reconstruction relation. The algorithm of [BMTY05] then directly applies, provided the stochastic reconstruction relation can be integrated with enough accuracy. We will not treat the full inexact matching problem here. Instead, we will study the simpler case of exact matching, where the energy functional consists only of the kinetic term.

### 3. EXACT STOCHASTIC LANDMARK MATCHING

In this section, we apply the previous ideas of stochastic deformation of LDDMM to exact matching with landmark dynamics. This is the simplest data structure in the LDDMM framework, and it will serve to give interesting insights into the effect of the noise in this context. Since exact matching means that the energy functional contains only a kinetic energy, the optimal vector field is found from a boundary value problem with the Euler-Poincaré equation (2.3). For exact matching, the momentum map for landmarks takes the simple familiar form for the reduction of the EPDiff equation (see [CH93, HM05])

$$\mathbf{m}(x, t) = \sum_{i=0}^N \mathbf{p}_i(t) \delta(\mathbf{x} - \mathbf{q}_i(t)), \quad (3.1)$$

for  $N$  landmarks with momenta  $\mathbf{p}_i$  and positions  $\mathbf{q}_i$ , with  $i = 1, 2, \dots, N$ . A direct substitution of  $u = K * m$  into the stochastic Euler-Poincaré equation (2.7) gives the stochastic landmark equations in (3.6). Here,  $K$  is a given kernel corresponding to the Green's function of the differential operator  $L$  used to construct the Lagrangian. Below, we take a different approach and proceed from a variational principle in which the stochastic landmark dynamics is constrained. We refer the interested reader to



for example [JS14] for a detailed exposition of this derivation in the deterministic context.

**3.1. Stochastic Landmarks dynamics.** Recall that for  $N$  landmarks in  $\mathbb{R}^d$ , the diffeomorphism group elements  $g$  act on the landmarks by evaluation of their position  $g \cdot \mathbf{q} = (g(q_1), \dots, g(q_N))$ , and the associated momentum map is (3.1). The original action functional (2.1) can be equivalently written as a constrained variational principle where the  $\mathbf{p}_i$  play the role of Lagrange multipliers enforcing the stochastic reconstruction relation (2.6). This procedure is based on the Clebsch action principle, which for landmark dynamics has been studied for one dimensional motion of landmarks on the real line in [HT16b]

$$S(\mathbf{u}, \mathbf{q}, \mathbf{p}) = \iint l(\mathbf{u}) d\mathbf{x} dt + \sum_i \int \mathbf{p}_i \cdot \left( \circ d\mathbf{q}_i - \mathbf{u}(\mathbf{q}_i) dt + \sum_l \sigma_l(\mathbf{q}_i) \circ dW_t^l \right). \quad (3.2)$$

Notice that only the Lagrangian depends on the spatial (Eulerian) variable  $\mathbf{x}$  on the image domain. We now use the momentum map (3.1) which provides us with the relation

$$2l(\mathbf{u}) = \int \mathbf{m}(\mathbf{q}, \mathbf{p})(\mathbf{x}) \cdot \mathbf{u}(\mathbf{x}) d\mathbf{x} = \sum_i \mathbf{p}_i \cdot \mathbf{u}(\mathbf{q}_i),$$

which reduces this action functional to the finite dimensional space of landmarks. We arrive at the action integral

$$S(\mathbf{q}, \mathbf{p}) = \int h(\mathbf{q}, \mathbf{p}) dt + \sum_i \int \mathbf{p}_i \cdot \left( \circ d\mathbf{q}_i + \sum_l \sigma_l(\mathbf{q}_i) \circ dW_t^l \right), \quad (3.3)$$

where the Hamiltonian only depends on the landmark variables, as

$$h(\mathbf{q}, \mathbf{p}) = \frac{1}{2} \sum_{i,j=1}^N (\mathbf{p}_i \cdot \mathbf{p}_j) K(\mathbf{q}_i - \mathbf{q}_j). \quad (3.4)$$

The action integral in (3.3) involves the *phase space Lagrangian* (3.4) and the *stochastic potential*, given by

$$\phi_l(\mathbf{q}, \mathbf{p}) := \sum_i \mathbf{p}_i \cdot \sigma_l(\mathbf{q}_i). \quad (3.5)$$

Taking free variations of (3.3) gives the stochastic Hamilton equations in the form

$$\begin{aligned} d\mathbf{q}_i &= \frac{\partial h}{\partial \mathbf{p}_i} dt + \sum_l \frac{\partial \phi_l}{\partial \mathbf{p}_i} \circ dW_t^l, \\ d\mathbf{p}_i &= -\frac{\partial h}{\partial \mathbf{q}_i} dt - \sum_l \frac{\partial \phi_l}{\partial \mathbf{q}_i} \circ dW_t^l. \end{aligned} \quad (3.6)$$

Explicitly, we have

$$\begin{aligned} d\mathbf{q}_i &= \sum_j \mathbf{p}_j K(\mathbf{q}_i - \mathbf{q}_j) dt + \sum_l \sigma_l(\mathbf{q}_i) \circ dW_t^l, \\ dp_i &= - \sum_j \mathbf{p}_i \cdot \mathbf{p}_j \frac{\partial}{\partial \mathbf{q}_i} K(\mathbf{q}_i - \mathbf{q}_j) dt - \sum_l \frac{\partial}{\partial \mathbf{q}_i} (\mathbf{p}_i \cdot \sigma_l(\mathbf{q}_i)) \circ dW_t^l. \end{aligned} \quad (3.7)$$

In coordinates, the stochastic equations (3.6) become

$$\begin{aligned} dq_i^\alpha &= \frac{\partial h}{\partial p_i^\alpha} dt + \sum_l \sigma_l^\alpha(\mathbf{q}_i) \circ dW_t^l, \\ dp_i^\alpha &= - \frac{\partial h}{\partial q_i^\alpha} dt - \sum_{l,\beta} \frac{\partial \sigma_l^\beta(\mathbf{q}_i)}{\partial q_i^\alpha} p_i^\beta \circ dW_t^l, \end{aligned} \quad (3.8)$$

where  $\alpha, \beta$  run through the domain directions,  $\alpha, \beta = 1, \dots, d$ .

The particular form of the stochastic potential in (3.5) arises from the Legendre transformation of (3.2). The solutions of (3.8) represent the singular solutions of the stochastic EPDiff equation, corresponding to a stochastic path in the diffeomorphism group. In previous works such as [TV12, Via13, MS16], noise has been introduced additively and only in the momentum equation, corresponding to a stochastic force. Also, the noise has typically been taken to be different for each landmark, and one can interpret it having been attached to each landmark. In the present case, the noise is not additive and the Wiener processes are not related to the landmarks, but to the domain of the image. Nearby landmarks will thus be affected by a similar noise, controlled by the spatial correlations of the noise. We refer to Figure 1 in the Introduction for a numerical experiment demonstrating this effect.

**Remark 3.1** (Geometric noise). *The geometric origin of our noise deserves more explanation. In the position equation (3.6), the noise arises as the infinitesimal transformation by the action of the stochastic vector field in (2.6), namely  $dgg^{-1} = udt + \sum_i \sigma_i \circ dW_i^t$ , on the manifold of positions of the landmarks, which is generated by the  $J$  stochastic potentials,  $\Phi_l(\mathbf{q}_i, \mathbf{p}_i) := \mathbf{p}_i \cdot \sigma_l(\mathbf{q}_i)$ . Since this stochastic Hamiltonian is linear in the canonical momenta, the noise perturbing the evolution of the landmark positions is independent of the landmark momenta. On the other hand, the noise in the momentum equations arises as the cotangent lift of the action of the stochastic vector field  $dgg^{-1}$  on the positions of the landmarks. This cotangent lift determines the action on the momentum fibres attached to the perturbed position of each of the landmarks in phase space. The cotangent lift transformation is given explicitly by the product of the momentum and the gradient of the spatial fields  $\sigma_l$  with respect to the position  $\mathbf{q}_i$  of the  $i$ -th landmark. This dependence increases the effect of the noise in regions where the  $\sigma_l$  fields have large spatial gradients, provided the landmarks are moving rapidly enough for their momenta to be non-negligible. We will see in the example that in certain cases this dependence on the momentum can significantly affect the dynamics of the landmarks.*

**3.2. The Fokker-Planck Equation.** In this section, we study the evolution of the probability density function of the stochastic landmarks by using the Fokker-Planck

equation. This study is possible in the case of landmarks because the associated phase space is finite dimensional.

We will denote the probability density by  $\mathbb{P}(\mathbf{q}, \mathbf{p}, t)$ , on the phase space  $\mathbb{R}^{2dN}$  at time  $t$ . The Fokker-Planck equation can be computed using standard procedures and is given in the following proposition.

**Proposition 3.2.** *The Fokker-Planck equation associated to the stochastic process (3.6) for the probability distribution  $\mathbb{P} : \mathbb{R}^{2dN} \times \mathbb{R} \rightarrow \mathbb{R}$  is given by*

$$\frac{d}{dt}\mathbb{P}(\mathbf{q}, \mathbf{p}, t) = \{h, \mathbb{P}\}_{\text{can}} + \frac{1}{2} \sum_l \{\phi_l, \{\phi_l, \mathbb{P}\}_{\text{can}}\}_{\text{can}} := \mathcal{L}^* \mathbb{P}, \quad (3.9)$$

where  $\{F, G\}_{\text{can}} = \nabla F^T \mathbb{J} \nabla G$  is the canonical bracket with  $\mathbb{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$  and  $\phi_l(\mathbf{q}, \mathbf{p}) = \sum_i \mathbf{p}_i \cdot \sigma_l(\mathbf{q}_i)$  are the stochastic potentials. This formula also defines the forward Kolmogorov operator  $\mathcal{L}^*$ .

*Proof.* The proof follows the standard derivation of the Fokker-Planck equation, by taking into account the geometrical structure of the stochastic process (3.6). The time evolution of an arbitrary function  $f : \mathbb{R}^{2dN} \rightarrow \mathbb{R}$  can be written

$$df(\mathbf{p}, \mathbf{q}) = \{f, h\}_{\text{can}} dt + \sum_l \{f, \phi_l\}_{\text{can}} \circ dW_t^l.$$

We then take the expectation of this stochastic process. Thus, we first compute the Itô correction, which reads  $\frac{1}{2} \sum_l \{\{f, \phi_l\}_{\text{can}}, \phi_l\}_{\text{can}} dt$ , as a double Poisson bracket. The expectation of the Itô process then removes the noise term and defines the forward Kolmogorov operator such that  $\dot{f} = \mathcal{L}^* f$ . By pairing this formula with the density function  $\mathbb{P}(\mathbf{q}, \mathbf{p}, t)$  over the phase space  $(\mathbf{q}, \mathbf{p})$  with the usual  $L^2$  pairing as

$$\int \mathbb{P}(\mathbf{q}, \mathbf{p}, t) \mathcal{L} f(\mathbf{q}, \mathbf{p}) d\mathbf{q} d\mathbf{p} = \int \mathcal{L}^* \mathbb{P}(\mathbf{q}, \mathbf{p}, t) f(\mathbf{q}, \mathbf{p}) d\mathbf{q} d\mathbf{p},$$

we obtain the Fokker-Planck equation  $\dot{\mathbb{P}} = \mathcal{L}^* \mathbb{P}$  that is explicitly given by (3.9) as the double bracket term is self-adjoint and the advection term anti-self-adjoint.  $\square$

Of course, the direct study of this equation is not possible, even numerically, due to its high dimensionality. The main use here of the Fokker-Planck equation will be to understand the time evolution of uncertainties around each landmark. Indeed, for each landmark  $\mathbf{q}_i$ , the corresponding marginal distribution (integrating  $\mathbb{P}$  over all the other variables) will represent the time evolution of the error on the mean trajectory of this landmark. We will show in the next section how to approximate the Fokker-Planck equation with a finite set of ordinary differential equations which describe the dynamics of the first moments of the distribution  $\mathbb{P}$ . This will then be used to estimate parameters of the noise fields  $\sigma_l$  for given sets of initial and final landmarks.

**Remark 3.3** (On ergodicity). *The question of ergodicity of the process (3.6) is not relevant here, as we will only consider this process for finite times, usually between  $t = 0$  and  $t = 1$ . The existence of stationary measures of the Fokker-Planck equation via Hörmander's theorem is thus not needed here. Nevertheless, we will need a notion of reachability in the landmark position in the next section, where we will show how*

to sample diffusion bridges for landmarks with fixed initial and final positions. This ensures that there exists a noise realisation which can bring any set of landmark to any other set of landmarks. This property is weaker than the Hörmander condition and was introduced by [Sus73].

**3.3. Diffusion Bridges.** The transition probability and solution to the Fokker-Planck equation  $\mathbb{P}(\mathbf{q}, \mathbf{p}, t)$  can also be estimated by Monte Carlo sampling of diffusion bridges. This approach will, in particular, be natural for maximum likelihood estimation of parameters of landmark processes using the Expectation-Maximisation (EM) algorithm that will involve expectation over unobserved landmark trajectories. This estimation approach will be used in section 4.3, and we here develop a theory of conditioned bridge processes for landmark dynamics which we will employ in the estimation. The approach is based on the method of [DH06] with two main modifications. The scheme and its modifications will be detailed after a short description of the approach of [DH06].

In [DH06], a Girsanov formula [Gir60], generalized to account for unbounded drifts, is used to show that when the diffusion field  $\Sigma(\mathbf{x}, t)$  of an  $\mathbb{R}^d$ -valued diffusion process

$$d\mathbf{x} = b(\mathbf{x}, t)dt + \Sigma(\mathbf{x}, t)dW, \quad \mathbf{x}_0 = \mathbf{u} \quad (3.10)$$

is uniformly invertible, the corresponding process conditioned on hitting a point  $\mathbf{v} \in \mathbb{R}^d$  at time  $T > 0$  are absolutely continuous with respect to an explicitly constructed unconditioned process  $\mathbf{y}$  that will hit  $\mathbf{v}$  at time  $T$  a.s.. The modified process  $\mathbf{y}$  is constructed by adding a guiding drift term that forces the process towards the target  $\mathbf{v}$ . In [DH06], this process is constructed as a modification of (3.10)

$$d\mathbf{y} = b(\mathbf{y}, t)dt - \frac{\mathbf{y} - \mathbf{v}}{T - t}dt + \Sigma(\mathbf{y}, t)dW. \quad (3.11)$$

Letting  $P_{\mathbf{x}|\mathbf{v}}$  denote the law of  $\mathbf{x}$  conditioned on hitting  $\mathbf{v}$  with corresponding expectation  $\mathbb{E}_{\mathbf{x}|\mathbf{v}}$ , the Cameron-Martin-Girsanov theorem implies that  $P_{\mathbf{x}|\mathbf{v}}$  is absolutely continuous with respect to  $P_{\mathbf{y}}$ , see for example [Øk03] and the discussion in [PR12]. An explicit expression for the Radon-Nikodym derivative  $dP_{\mathbf{x}|\mathbf{v}}/dP_{\mathbf{y}}$  can be computed, and this derivative is central for using simulations of the process  $\mathbf{y}$  to compute expectations over the conditioned process  $\mathbf{x}|\mathbf{v}$ . In particular, as shown in [DH06], the conditioned process  $\mathbf{x}|\mathbf{v}$  and the modified process  $\mathbf{y}$  are related by

$$\mathbb{E}_{\mathbf{x}|\mathbf{v}}(f(\mathbf{x})) = \frac{\mathbb{E}_{\mathbf{y}}\left(f(\mathbf{y})\varphi(\mathbf{y})\right)}{\mathbb{E}_{\mathbf{y}}(\varphi(\mathbf{y}))}, \quad (3.12)$$

where  $\varphi(\mathbf{y})$  is a correction factor applied to each stochastic bridge  $\mathbf{y}$ . Notice here that  $f$  is a real-valued function of the stochastic path from  $t = 0$  to  $t = T$  and that the formula does not depend on time.

Returning to landmark evolutions in the phase space  $\mathbb{R}^{2dN}$ , the process (3.6) has two vector variables  $(\mathbf{q}, \mathbf{p})$  that typically will be conditioned on hitting a fixed set of landmark positions  $\mathbf{v}$  at time  $T$ . The conditioning thus happens only in the  $\mathbf{q}$  variables by requiring  $\mathbf{q}_T = \mathbf{v}$ . To construct bridges with an approach similar to the scheme of [DH06], we need to find an appropriate guiding term and handle the

fact that the diffusion field may not be invertible in general. Recall first that the landmark process (3.6) has diffusion field

$$\Sigma(\mathbf{q}, \mathbf{p}) = \begin{pmatrix} \Sigma_{\mathbf{q}}(\mathbf{q}) \\ \Sigma_{\mathbf{p}}(\mathbf{q}, \mathbf{p}) \end{pmatrix} := \begin{pmatrix} \sigma_1(\mathbf{q}), \dots, \sigma_J(\mathbf{q}) \\ -\nabla_{\mathbf{q}}(\mathbf{p} \cdot \sigma_1(\mathbf{q})), \dots, -\nabla_{\mathbf{q}}(\mathbf{p} \cdot \sigma_J(\mathbf{q})) \end{pmatrix} \quad (3.13)$$

where  $\sigma_j(\mathbf{q})$  denotes the vector  $(\sigma_j(q_1), \dots, \sigma_j(q_N))^T$ . Notice that this matrix is not square and has dimension  $2dN \times J$  so that  $\Sigma(\mathbf{q}, \mathbf{p}) \circ dW_t$  with  $dW_t$  a  $J$ -vector corresponds to the stochastic term of (3.6). Though  $\Sigma(\mathbf{q}, \mathbf{p})$  couples the  $\mathbf{q}$  and  $\mathbf{p}$  equation, when the number of noise fields  $J$  is sufficiently large, the  $\mathbf{q}$  part  $\Sigma_{\mathbf{q}}(\mathbf{q})$  will often be surjective as a linear map  $\mathbb{R}^J \rightarrow \mathbb{R}^{dN}$ . In this situation, by letting  $\Sigma_{\mathbf{q}}(\mathbf{q})^\dagger$  denote the Moore-Penrose pseudo-inverse of  $\Sigma_{\mathbf{q}}(\mathbf{q})$ , we can construct a guiding drift term as

$$G(\mathbf{q}, \mathbf{p}) := -\frac{\Sigma(\mathbf{q}, \mathbf{p})\Sigma_{\mathbf{q}}(\mathbf{q})^\dagger(\mathbf{q} - \mathbf{v})}{T - t}. \quad (3.14)$$

This term, when added to the process (3.6), ensures that the modified process hits  $\mathbf{q}_T$  a.s. at time  $T$ . The drift term (3.14) is a direct generalization of the term added in (3.11). If  $\Sigma$  had been invertible then  $\Sigma\Sigma^\dagger = \text{Id}$  resulting in the guiding term of [DH06] used in equation (3.11). In the current non-invertible case,  $\Sigma\Sigma_{\mathbf{q}}^\dagger(\mathbf{q} - \mathbf{v})$  uses the difference  $\mathbf{q} - \mathbf{v}$  which only involves the landmark position but affects both the position and the momentum equations. We stress here the fact that the introduction of noise in the  $\mathbf{q}$  equation in (3.6) is essential for this approach. When conditioning on the  $\mathbf{q}$  variable, a guided process could not directly be constructed in this way if the noise was instead introduced in the  $\mathbf{p}$  equation as in [TV12, Via13, MS16]. The fact that this term is weighted by  $\Sigma\Sigma^\dagger$  is also important as it allows the guiding term to be more efficient in the noisy regions of the image, where there is more freedom to deviate from the deterministic path. The guiding term can be interpreted as originating from a time-rescaled gradient flow, and with the guiding term added, the diffusion process can be seen as a stochastically perturbed gradient flow, see [AHPS17].

The guiding term (3.14) is, in practice, not always appropriate for landmarks. Because the correction is dependent only on the difference to the target in the position equation, a phenomenon of over-shooting is often observed. In such cases, the landmarks travel too fast initially due to a large momentum, strengthened by the guiding term forcing the landmarks towards  $\mathbf{v}$ . The high initial speed is only corrected when the time approaches  $T$  and the guiding term brings the landmark back to their final position. This effect is illustrated in Figure 4 in section 5.2 and results in low values of the correction factor  $\varphi(\mathbf{q}, \mathbf{p})$  used to compute the expectation in (3.12). This results in inefficient samples when approximating (3.12) by Monte Carlo sampling. Instead, letting  $b(\mathbf{q}, \mathbf{p})$  denote the drift term of (3.6) with Itô correction, we use a guided diffusion process of the form

$$\begin{pmatrix} d\hat{\mathbf{q}} \\ d\hat{\mathbf{p}} \end{pmatrix} = b(\hat{\mathbf{q}}, \hat{\mathbf{p}})dt - \frac{\Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})\Sigma_{\mathbf{q}}(\hat{\mathbf{q}})^\dagger(\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \mathbf{v})}{T - t}dt + \Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})dW, \quad (3.15)$$

for some appropriately chosen function  $\phi_{t,T} : \mathbb{R}^{2dN} \rightarrow \mathbb{R}^{dN}$  that gives an estimate of the value of  $\hat{\mathbf{q}}_T$  using the value of the modified stochastic process  $(\hat{\mathbf{q}}_t, \hat{\mathbf{p}}_t)$  at time  $t$ . The hat denotes the solution of the process (3.15), which is different from

the original dynamics of the process (3.6) written without the hats. The choice  $\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) := \hat{\mathbf{q}}$  recovers the guiding term (3.14). It would be natural to define  $\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) := \mathbb{E}_{(\mathbf{q}, \mathbf{p})}(\mathbf{q}_T | (\mathbf{q}_t, \mathbf{p}_t) = (\hat{\mathbf{q}}, \hat{\mathbf{p}}))$ . The resulting guiding term will only be driven by the expected amount needed at the endpoint, not from the value at time  $t$ . A similar choice but easier to handle is to let  $\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}})$  be the solution at time  $T$  of the original deterministic landmark dynamics (2.3), obtained from the initial conditions  $(\hat{\mathbf{q}}_t, \hat{\mathbf{p}}_t) = (\hat{\mathbf{q}}, \hat{\mathbf{p}})$ . We will use this latter choice, which is visualised in Figure 4, in the rest of the paper. We note that to ensure convergence, a bounded approximation  $\tilde{b}$  will replace the original unbounded drift  $b$  in (3.15), but this fact has little influence in practice.

The matrix  $\Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})\Sigma_{\mathbf{q}}(\hat{\mathbf{q}})^\dagger$  in 3.15 only accounts for the  $\mathbf{q}$  dynamics in the pseudo-inverse  $\Sigma_{\mathbf{q}}(\hat{\mathbf{q}})^\dagger$ . When the momentum is high and the noise fields  $\sigma_j$  have high gradients, this fact can again lead to improbable sample paths. In such cases, the scheme can be further generalised by using a guiding term of the form

$$\frac{1}{T-t}\Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})\left(D_h(\phi_{t,T}(\Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})\mathbf{h}))|_{\mathbf{h}=\mathbf{0}}\right)^\dagger(\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \mathbf{v}). \quad (3.16)$$

The matrix  $D_h(\phi_{t,T}(\Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})\mathbf{h}))|_{\mathbf{h}=\mathbf{0}}$  is a linear approximation of the expected endpoint dynamics as a function of the noise vector  $\mathbf{h} \in \mathbb{R}^J$ . Again, with  $\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) := \hat{\mathbf{q}}$ , the original guiding term (3.14) is recovered, and the term is close to the guiding term of (3.15) when the momentum or gradients of  $\sigma_j$  are low. We use this term for the experiments in section 5.2 involving high momentum dynamics, e.g. Figure 6.

The following result is an extension of [DH06, Theorem 5] and [Mar11, Theorem 3] to the modified guided SDE (3.15). It is the basis for the EM approach for estimating the parameters of the landmark processes developed in section 4.3. We let  $W(\mathbb{R}^{2dN})$  denote the Wiener space of continuous paths  $[0, T] \rightarrow \mathbb{R}^{2dN}$  and write  $D$  and  $H$  for the component-wise Jacobian and Hessian matrices, respectively.

**Theorem 3.4.** *Assume  $\Sigma_{\mathbf{q}}(\mathbf{q}) : \mathbb{R}^J \rightarrow \mathbb{R}^{Nd}$  is surjective for all  $\mathbf{q}$  with  $\Sigma_{\mathbf{q}}(\hat{\mathbf{q}})^\dagger$  bounded, and that  $\tilde{b}$  and  $\Sigma$  are  $C^{1,2}$ , bounded, and with bounded derivatives. Let  $P_{(\mathbf{q}, \mathbf{p})|\mathbf{v}}$  be the law of  $(\mathbf{q}, \mathbf{p}) | \mathbf{q}_T = \mathbf{v}$ , and let  $(\hat{\mathbf{q}}, \hat{\mathbf{p}})$  be solution to (3.15),  $(\hat{\mathbf{q}}_0, \hat{\mathbf{p}}_0) = (\mathbf{q}_0, \mathbf{p}_0)$  with  $\phi_{t,T} : \mathbb{R}^{2dN} \rightarrow \mathbb{R}^{2dN}$  given by the solution of (2.3). Then, with a positive measurable  $f : W(\mathbb{R}^{2dN}) \rightarrow \mathbb{R}$ ,*

$$\mathbb{E}_{(\mathbf{q}, \mathbf{p})|\mathbf{v}}(f(\mathbf{q}, \mathbf{p})) = \lim_{t \rightarrow T} \frac{\mathbb{E}_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}(f(\hat{\mathbf{q}}, \hat{\mathbf{p}})\varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t))}{\mathbb{E}_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}(\varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t))}, \quad (3.17)$$

with

$$\begin{aligned} \log \varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t) = & - \int_0^t \frac{\hat{\mathbf{q}}_{\mathbf{v}}^T A(\hat{\mathbf{q}})((D_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}\hat{\mathbf{q}}_{\mathbf{q}})d\hat{\mathbf{q}} - \frac{1}{2}\text{Tr}(\Sigma_{\mathbf{q}}(\hat{\mathbf{q}})^T(H_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}\hat{\mathbf{q}}_{\mathbf{q}})\Sigma_{\mathbf{q}}(\hat{\mathbf{q}}))ds)}{T-s} \\ & - \int_0^t \frac{\hat{\mathbf{q}}_{\mathbf{v}}^T d(A(\hat{\mathbf{q}}))\hat{\mathbf{q}}_{\mathbf{v}}}{2(T-s)} - \sum_{i,j} \int_0^t \frac{2\hat{\mathbf{q}}_{i,\mathbf{v}}[dA_{ij}(\hat{\mathbf{q}}), d\hat{\mathbf{q}}_j] + A_{ij}(\hat{\mathbf{q}})[d\hat{\mathbf{q}}_i, d\hat{\mathbf{q}}_j]}{2(T-s)} \\ & + \int_0^t (b(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \tilde{b}(\hat{\mathbf{q}}, \hat{\mathbf{p}}))^T \Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})^\dagger, T d(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \frac{1}{2} \|\Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})^\dagger(b(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \tilde{b}(\hat{\mathbf{q}}, \hat{\mathbf{p}}))\|^2 ds, \end{aligned}$$

where

$$\hat{\mathbf{q}}_{\hat{\mathbf{q}}} = \phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \hat{\mathbf{q}}, \quad \hat{\mathbf{q}}_{\mathbf{v}} = \phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \mathbf{v}, \quad A(\hat{\mathbf{q}}) = (\Sigma_{\mathbf{q}}(\hat{\mathbf{q}})\Sigma_{\mathbf{q}}(\hat{\mathbf{q}})^T)^{-1}.$$

In the Theorem,  $[\cdot, \cdot]$  is the quadratic variation of semi-martingales. As mentioned above, a bounded approximation  $\tilde{b}$  must be used to replace the original drift term  $b$  in (3.15). The last integral in the expression for  $\log \varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t)$  is a result of this approximation.

The result is proved in Appendix A. Using the guidance scheme (3.11), the right hand side limit of (3.17) is in [DH06] shown to equal

$$\frac{\mathbb{E}_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}(f(\hat{\mathbf{q}}, \hat{\mathbf{p}})\varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, T))}{\mathbb{E}_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}(\varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, T))},$$

when  $\Sigma$  is invertible. Extending the convergence argument to the present non-invertible case is non-trivial, and we postpone investigating this to future work.

The second order term  $-\frac{1}{2}\text{Tr}(\Sigma_{\mathbf{q}}(\hat{\mathbf{q}})^T(H_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}\hat{\mathbf{q}}_{\hat{\mathbf{q}}}))$  of the correction factor is generally small. Though it can be computed during the simulation, we ignore it in the numerical experiments. The terms

$$\hat{\mathbf{q}}_{\mathbf{v}}^T d(A(\hat{\mathbf{q}}))\hat{\mathbf{q}}_{\mathbf{v}} + 2\hat{\mathbf{q}}_{i,\mathbf{v}}[dA_{ij}(\hat{\mathbf{q}}), d\hat{\mathbf{q}}_j] + A_{ij}(\hat{\mathbf{q}})[d\hat{\mathbf{q}}_i, d\hat{\mathbf{q}}_j]$$

can be computed by finite difference approximation as noted in [DH06].

#### 4. ESTIMATING THE SPATIAL CORRELATION OF THE NOISE

We now assume a set of  $n$  observed landmark configurations  $\mathbf{q}^1, \dots, \mathbf{q}^n$  at time  $T$ , i.e. the observations are considered realisations of the stochastic process at some positive time  $T$ . From this data, we aim at inferring parameters of the model. This can be both parameters of the noise fields  $\sigma_i$  and parameters for the initial configuration  $(\mathbf{q}(0), \mathbf{p}(0))$ . The initial configuration can be deterministic with fixed known or unknown parameters, or it can be randomly chosen from a distribution with known or unknown parameters. We develop two different strategies for performing the inference. The first inference method in section 4.2 is a shooting method based on solving the evolution of the first moments of the probability distribution of the landmark positions while the second method in section 4.3 is based on the Expectation-Maximisation (EM) algorithm. The discussion is here in the context of landmarks, although these ideas may also apply in the more general context of section 2.

**4.1. The Noise Fields.** We start by discussing the form of the unknown  $J$  noise fields  $\sigma_l$ . To estimate them from a finite amount of observed data, we are forced to require the fields to be specified by a finite number of parameters. A possible choice for a family of noise fields is to select  $J$  linearly independent elements  $\sigma_1, \dots, \sigma_J$  from a dense subset of  $C^1(\mathbb{R}^d, \mathbb{R}^d)$ . We here use a kernel  $k$  with length-scale  $r_l$  and a noise amplitude  $\lambda_l \in \mathbb{R}^d$ , that is

$$\sigma_l^\alpha(\mathbf{q}_i) = \lambda_l^\alpha k_{r_l}(\|\mathbf{q}_i - \delta_l\|), \tag{4.1}$$

where  $\delta_l$  denotes the kernel positions. Possible choices for the kernel include Gaussians  $k_{r_l}(x) = e^{-x^2/(2r_l^2)}$ , or cubic B-splines  $k_{r_l}(x) = S_3(x/r_l)$ . The Gaussian kernel has the advantage of simplifying calculations of the moment equations whereas the

B-spline representation is compactly supported and gives a partition of unity when used in a regular grid. Other interesting choices may include a cosine or a polynomial basis of the image domain.

In principle, the methods below allow all parameters of the noise fields to be estimated given sufficient amount of data. However, for simplicity, we will fix the length-scale and the position of the kernels. The unknown parameters for the noise can then be specified in a single vector variable  $\theta = (\lambda_1, \dots, \lambda_K)$ . The aim of the next sections will be to estimate this vector, possibly in addition to the initial configuration  $(\mathbf{q}(0), \mathbf{p}(0))$ , from data using the method of moments in 4.2 and EM in 4.3, respectively.

**4.2. Method of moments.** We describe here our first method for estimating the parameters  $\theta$  by solving a shooting problem on the space of first and second order moments. Given an estimate of the endpoint distributions  $\mathbb{P}(\mathbf{q}, \mathbf{p}, T)$ , we will solve the inverse problem which consists in using the Fokker-Planck equation (3.9) to find the values of  $\theta$  such that we can reproduce the observed final distribution. Solving the Fokker-Planck equation directly is infeasible due to its high dimensionality, so we will derive a set of finite dimensional equations approximating the Fokker-Planck equation (3.9). One very successful method is to approximate the probability distribution  $\mathbb{P}$  by its first moments. This has been developed in the context of plasma physics for the Liouville equation, an equation similar to the Fokker-Planck equation (3.9).

**Remark 4.1** (Geometric moment equation). *As the Fokker-Planck (3.9) is written in term of the canonical bracket, we could expect to be able to apply a geometrical version of the method of moments such as the one developed by [HPT07]. Although this method seems to fit the present geometric derivation of the stochastic equations, we will not use it as it is not in our case practically useful. Indeed, it requires the expansions of the Hamiltonian functions in term of the moments, but we cannot obtain here a valid expansion with a finite number of terms. This is due to the fact that the LDDMM kernel and the noise kernels cannot generally be globally approximated by finite polynomials with bounded approximation error for large distances. This would, in turn, produce spurious strong interactions between distant landmarks.*

The method for approximating the Fokker-Planck that we will use here is the following. We first define the moments

$$\langle q_i^\alpha \rangle := \int q_i^\alpha \mathbb{P}_\theta(\mathbf{q}, \mathbf{p}, t) d\mathbf{q}d\mathbf{p} \quad (4.2)$$

$$\langle q_i^\alpha p_j^\beta \rangle := \int q_i^\alpha p_j^\beta \mathbb{P}_\theta(\mathbf{q}, \mathbf{p}, t) d\mathbf{q}d\mathbf{p}, \quad (4.3)$$

where we have written only two possible moments, although any combinations of  $p$  and  $q$  at any order are possible. In this work we will only consider moments up to the second order, that is the moments  $\langle q_i^\alpha \rangle$ ,  $\langle p_i^\alpha \rangle$ ,  $\langle q_i^\alpha q_j^\beta \rangle$ ,  $\langle q_i^\alpha p_j^\beta \rangle$  and  $\langle p_i^\alpha p_j^\beta \rangle$ . Notice that the first moment are  $(1, 1)$ -tensors, and the second moments are  $(2, 2)$ -tensors, although we will only use index notation here.



We illustrate this method with the first moment  $\langle q_i^\alpha \rangle$ , which represents the mean position of the landmarks. We compute its time derivative and use the property of the Kolmogorov operator  $\mathcal{L}$  defined in (3.9) to obtain

$$\frac{d}{dt} \langle q_i^\alpha \rangle = \int q_i^\alpha \mathcal{L}^* \mathbb{P}_\theta d\mathbf{q}d\mathbf{p} = \int \mathcal{L} q_i^\alpha \mathbb{P}_\theta d\mathbf{q}d\mathbf{p} = \langle \mathcal{L} q_i^\alpha \rangle. \quad (4.4)$$

We thus first need to apply the Kolmogorov operator  $\mathcal{L}$  to  $q_i^\alpha$  to obtain

$$\begin{aligned} \mathcal{L} q_i^\alpha &= -\{h, q_i^\alpha\}_{\text{can}} + \frac{1}{2} \sum_l \{\phi_l, \{\phi_l, q_i^\alpha\}_{\text{can}}\}_{\text{can}} \\ &= \frac{\partial h}{\partial p_i^\alpha} + \frac{1}{2} \frac{\partial \sigma_l^\alpha(\mathbf{q}_i)}{\partial q_i^\beta} \sigma_l^\beta(\mathbf{q}_i), \end{aligned} \quad (4.5)$$

which corresponds to the  $q$  part of the drift of the stochastic process with Itô correction. Similarly, for the momentum evolution, we obtain

$$\begin{aligned} \mathcal{L} p_i^\alpha &= -\{h, p_i^\alpha\}_{\text{can}} + \frac{1}{2} \sum_l \{\phi_l, \{\phi_l, p_i^\alpha\}_{\text{can}}\}_{\text{can}} \\ &= -\frac{\partial h}{\partial q_i^\alpha} + \frac{1}{2} p_i^\gamma \frac{\partial \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\beta} \frac{\partial \sigma_l^\beta(\mathbf{q}_i)}{\partial q_i^\alpha} - \frac{1}{2} p_i^\beta \frac{\partial^2 \sigma_l^\beta(\mathbf{q}_i)}{\partial q_i^\alpha \partial q_i^\gamma} \sigma_l^\gamma(\mathbf{q}_i). \end{aligned} \quad (4.6)$$

Most of the terms on the right hand side of (4.5) and (4.6) are nonlinear; so their expected value cannot be written in terms of only the first moments. This is the usual closure problem of moment equations, such as the BBGKY problem arising in many-body problems in quantum mechanics. The solution to this problem is to truncate the hierarchy of moments at a given order and consider the system of ODEs as an approximation of the complete Fokker-Planck solution. Here we will apply the so-called *cluster expansion method* described in [KK11]. We refer to the Appendix B.1 for more details about this method.

Apart from the first approximation  $\langle q_i^\alpha q_j^\beta \rangle \rightarrow \langle q_i^\alpha \rangle \langle q_j^\beta \rangle$ , the next order of approximation is to keep track of the correlations

$$\Delta_2 \langle q_i^\alpha q_j^\beta \rangle := \langle q_i^\alpha q_j^\beta \rangle - \langle q_i^\alpha \rangle \langle q_j^\beta \rangle. \quad (4.7)$$

This quantity is also called a centred second moment as for  $i = j$  it corresponds to the covariance of the probability distribution for the landmark  $i$ . In general, it corresponds to the correlation between the positions of two landmarks. The dynamical equation for this correlation is found from the equation of the second moment, which gives

$$\begin{aligned} \frac{\partial}{\partial t} \Delta_2 \langle q_i^\alpha q_j^\beta \rangle &= \frac{\partial}{\partial t} \langle q_i^\alpha q_j^\beta \rangle - \langle q_i^\alpha \rangle \frac{\partial}{\partial t} \langle q_j^\beta \rangle + T \\ &= \sum_l \left\langle \sigma_l^\alpha(\mathbf{q}_i) \sigma_l^\beta(\mathbf{q}_j) \right\rangle + \left\langle q_i^\alpha \frac{\partial h}{\partial p_j^\beta} \right\rangle - \langle q_i^\alpha \rangle \left\langle \frac{\partial h}{\partial p_j^\beta} \right\rangle \\ &\quad + \frac{1}{2} \sum_l \left( \left\langle q_i^\alpha \sigma_l^\gamma(\mathbf{q}_i) \frac{\partial \sigma_l^\beta(\mathbf{q}_j)}{\partial q_j^\gamma} \right\rangle - \langle q_i^\alpha \rangle \left\langle \sigma_l^\gamma(\mathbf{q}_i) \frac{\partial \sigma_l^\beta(\mathbf{q}_j)}{\partial q_j^\gamma} \right\rangle \right) + T, \end{aligned}$$

where  $T$  stands for the transpose of the previous term. This equation is interesting to study in more detail, as it already gives us information about the nature of the

dynamics for the spatial covariance of landmarks. Indeed, we have three types of terms with the following effects.

- (1) *The  $\sigma_l$ -dependent terms.* This first term is quadratic in the  $\sigma$ 's, not proportional to any linear or quadratic polynomial in  $q$  or  $p$ . This term is a direct contribution from the noise in the  $q$  equation and will have the effect of almost linearly increasing the centred covariance, wherever a  $\sigma_l > 0$ .
- (2) *The  $h$ -dependent terms.* From the form of this term, we expect it to be proportional to a correlation. It will thus have an exponential effect on the dynamics, triggered by the linear contribution of the first term. Notice that this term only depends on the Hamiltonian, and, thus, on the interaction between landmarks. If two landmarks interact, we expect their covariance to be averaged. This term will capture their averaged covariance.
- (3) *The  $\nabla_q \sigma_l$ -dependent terms.* These terms are related to the noise in the  $p$  equation and will account for the effect on the landmark position of the interaction of the momentum of the landmark with the gradients of the noise.

Notice that the last two types of terms describe second order effects with respect to the spatial covariance of the landmarks, as they depend linearly on the correlations. In the expansion of these nonlinear terms, the other correlations involving  $p$  will appear. This means that all of the possible second order correlations must be computed. This computation is done in Appendix B, where we also approximate the expected value of the kernels as  $\langle K(\mathbf{q}) \rangle \approx K(\langle \mathbf{q} \rangle)$ . As we will see in the numerical examples in section 5, these approximations can give a reliable estimate of the landmark covariance, but this should be rigorously justified to obtain a precise estimate of the errors. Such a study is beyond the scope of this work and is left open.

Given the equations for the moment evolution, we can estimate the parameters  $\theta$  by minimising the cost function

$$C(\langle \mathbf{p} \rangle(0), \lambda_l) = \frac{1}{\gamma_1} \|\langle \mathbf{q} \rangle - \langle \mathbf{q} \rangle(T)\|^2 + \frac{1}{\gamma_2} \|\Delta_2 \langle \mathbf{q}\mathbf{q} \rangle - \Delta_2 \langle \mathbf{q}\mathbf{q} \rangle(T)\|^2, \quad (4.8)$$

where  $\gamma_1$  and  $\gamma_2$  are weights. We denote by  $\langle \mathbf{q} \rangle$  and  $\Delta_2 \langle \mathbf{q}\mathbf{q} \rangle$  the target first and second moments and by  $\langle \mathbf{q} \rangle(T)$  and  $\Delta_2 \langle \mathbf{q}\mathbf{q} \rangle(T)$  the estimated moments which implicitly depend on the parameters of the noise and the initial momentum. The choice of the norm is free here, and we chose a norm which only considers  $i = j$  and normalises each term to 1 so that all the covariance of the landmarks contribute equally to the cost. Other choices could be made, depending on applications. Also, the cost function may depend on other parameters, but this would make its minimisation more difficult.

To minimise the cost (4.8), we can use gradient based methods such as the BFGS algorithm. Such methods require the evaluation of the Jacobian of  $C$  with respect to all of its arguments. Usually, for the estimation of the initial momentum, a linear adjoint equation is used. However, the derivative with respect to the parameters of the noise cannot be computed in this way. We will evaluate the gradients symbolically by using the Theano library in Python [The16]. To improve the efficiency of the algorithm, we first match the mean final position, by only updating the initial

momentum. Then, with this initial condition, we match for both first and second moments and update the initial momentum as well as the parameters  $\lambda_t$ . As we will see in the numerical experiments in section 5, gradient-based methods are not optimal, and genetic algorithms, such as the differential evolution algorithm of [SP97] designed for global minimizations, turn out to perform better.

**4.3. Maximum Likelihood and Expectation-Maximization.** We now describe how to estimate the unknown parameters collected in the variable  $\theta$  by a maximum likelihood estimation based on the expectation-maximisation (EM) algorithm of [DLR77]. The likelihood of  $n$  independent observations  $(\mathbf{q}^1, \dots, \mathbf{q}^n)$  at time  $T$  given parameters  $\theta$  takes the form

$$\mathcal{L}(\mathbf{q}^1, \dots, \mathbf{q}^n, \theta) = \prod_{i=1}^n \mathbb{P}_\theta(\mathbf{q}^i, T) = \prod_{i=1}^n \int_{\mathbb{R}^{Nd}} \mathbb{P}_\theta(\mathbf{q}^i, \mathbf{p}, T) d\mathbf{p}. \quad (4.9)$$

The parameters  $\theta$  can be estimated by maximizing the likelihood, i.e. letting

$$\hat{\theta} \in \operatorname{argmax}_\theta \mathcal{L}(\theta; \mathbf{q}^1, \dots, \mathbf{q}^n).$$

For this, the likelihood could be directly computed by numerical approximation of  $\mathbb{P}_\theta(\mathbf{q}_i, T)$  using an approximation of the Fokker-Planck equation (3.9). Alternatively, the fact that the stochastic process is only sampled at time  $T$  suggests a missing data approach that marginalises out the unobserved trajectories up to time  $T$ . Let  $(\mathbf{q}, \mathbf{p}; \theta)$  denote the stochastic landmark process with parameters  $\theta$ , and let  $P(\mathbf{q}, \mathbf{p}; \theta)$  denote its law. Let  $\mathcal{L}(\mathbf{q}, \mathbf{p}; \theta)$  denotes the likelihood of the entire stochastic path for a given realisation of the noise, and computed with respect to the parameter  $\theta$ . Notice that this likelihood is only defined for finite time discretizations of the process and there is no notion of path density for the infinite dimensional process. We thus proceed formally, while noting that the approach can be justified rigorously, see e.g. [DS08].

The EM algorithm finds a sequence of parameter estimates  $\{\theta_l\}$  converging to a  $\hat{\theta}$  by iterating over the following two steps:

- (1) *Expectation:* Compute the expected value of the log-likelihood given the previous parameter estimate  $\theta_{l-1}$ :

$$\begin{aligned} Q(\theta|\theta_{l-1}) &:= \mathbb{E}_{\theta_{l-1}}(\log \mathcal{L}(\mathbf{q}, \mathbf{p}; \theta) | \mathbf{q}^1, \dots, \mathbf{q}^n) \\ &= \sum_{i=1}^n \mathbb{E}_{\theta_{l-1}}(\log \mathcal{L}(\mathbf{q}, \mathbf{p}; \theta | \mathbf{q}^i)). \end{aligned} \quad (4.10)$$

The expectation (4.10) over the process conditioned on the observations  $\mathbf{q}_i$  integrates the likelihood over all sample paths reaching  $\mathbf{q}_i$ . Expectation over conditioned diffusion processes can be approximated by Monte Carlo simulation of diffusion bridges, see for example [DH06, PR12, BFS16]. Here, we employ the bridge simulation approach developed in section 3.3. For each  $\mathbf{q}^i$ , we thus exchange  $(\mathbf{q}_t, \mathbf{p}_t; \theta)$  with a guided process  $(\hat{\mathbf{q}}, \hat{\mathbf{p}}; \theta, \mathbf{q}^i)$  and use the equality (3.17) from Theorem 3.4. The expectation on the right-hand side of (3.17) can be approximated by drawing samples from the guided process. Note that the correction factor  $\varphi(\mathbf{q}, \mathbf{p}|\theta_{l-1}, \mathbf{q}_i)$  makes the approach equal to

importance sampling of the conditioned process with the guided process as proposal distribution.

- (2) *Maximisation*: Find the new parameter estimate

$$\theta_l = \operatorname{argmax}_\theta Q(\theta|\theta_{l-1}). \quad (4.11)$$

The maximisation step can be approximated by updating  $\theta_l$  such that it increases  $Q(\theta|\theta_{l-1})$  instead of maximising it. This is the approach of the generalised EM algorithm [NH98]. The update of  $\theta$  is thus computed by taking a gradient step

$$\theta_l = \theta_{l-1} + \epsilon \nabla_\theta Q(\theta|\theta_{l-1}), \quad (4.12)$$

where  $\epsilon > 0$ . The gradient which is evaluated for each of the sampled paths can be computed symbolically using the Theano library of [The16].

The resulting estimation algorithm is listed in Algorithm 1. For each  $\mathbf{q}^i$ , the expectation  $\mathbb{E}_{\theta_{l-1}}(\log \mathbb{P}_\theta(\mathbf{q}, \mathbf{p}|\mathbf{q}_i))$  is estimated by sampling  $N_{\text{bridges}}$  bridges. The algorithm can perform a fixed number  $L$  of updates to the estimate  $\theta_l$  or stop at convergence.

---

**Algorithm 1:** Stochastic EM-estimation of parameters  $\theta$ .

---

```

// Initialization
 $\theta_0 \leftarrow$  initialization value
// Main loop
for  $l = 1$  to  $L$  do
  for  $i = 1$  to  $n$  do
    for  $j = 1$  to  $N_{\text{bridges}}$  do
      sample bridge  $(\hat{\mathbf{q}}(\omega_j), \hat{\mathbf{p}}(\omega_j); \theta_{l-1}, \mathbf{q}^i)$ 
      compute  $\log \mathbb{P}_{\theta_l}(\hat{\mathbf{q}}(\omega_j), \hat{\mathbf{p}}(\omega_j))$  and  $\varphi(\hat{\mathbf{q}}(\omega_j), \hat{\mathbf{p}}(\omega_j))$ 
    end
    set  $C_{\mathbf{q}_i} = \text{mean}_j(\varphi(\hat{\mathbf{q}}(\omega_j), \hat{\mathbf{p}}(\omega_j)))$ 
    set  $\mathbb{E}_{(\mathbf{q}, \mathbf{p})|\mathbf{q}^i}(\log \mathbb{P}_{\theta_{l-1}}(\mathbf{q}, \mathbf{p})) \approx$ 
       $C_{\mathbf{q}_i}^{-1} \text{mean}_j(\log \mathbb{P}_\theta(\hat{\mathbf{q}}(\omega_j), \hat{\mathbf{p}}(\omega_j)) \varphi(\hat{\mathbf{q}}(\omega_j), \hat{\mathbf{p}}(\omega_j)))$ 
    end
    set  $Q(\theta|\theta_{l-1}) = \text{mean}_i(\mathbb{E}_{(\mathbf{q}, \mathbf{p})|\mathbf{q}_i}(\log \mathbb{P}_{\theta_{l-1}}(\mathbf{q}, \mathbf{p})))$ 
    compute  $\nabla_\theta Q(\theta|\theta_{l-1})$ 
    update  $\theta$ :  $\theta_l = \theta_{l-1} + \epsilon \nabla_\theta Q(\theta|\theta_{l-1})$ 
  end
end
```

---

## 5. NUMERICAL EXAMPLES

We now present several numerical tests of the stochastic perturbation of the landmark dynamics. In particular, we want to illustrate aspects of the effect of the noise on the landmarks and test the algorithms for estimation of the spatial correlation of the noise. We will focus here on synthetic examples and refer to [AHPS17] for an application of the methods on a dataset of Corpora Callosa shapes represented by 77 landmarks. The numerical simulations of this work have been done in Python,

using the symbolic computation framework Theano [The16]. The code is available from the public repository <https://bitbucket.org/stefansommer/stochlandyn>.

**5.1. Solution of the Fokker-Planck equation.** We first consider a simple experiment with a single landmark, subjected a square array of noise fields with Gaussian noise kernel. To a first order approximation, the mean trajectory of the landmark is a straight line with constant momenta as the Hamiltonian is a pure kinetic energy.

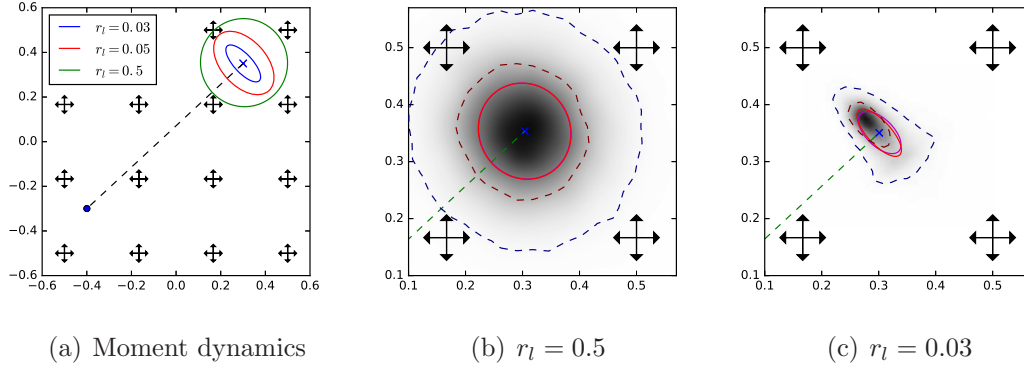


FIGURE 2. We plot on the left panel three simulations of a single landmark dynamics subject to an array of Gaussian noise fields. Their parameters are either  $\lambda_l = (0.08, 0)$  or  $\lambda_l = (0, 0.08)$ . We used three different length-scales  $r_l$  for the noise fields to analyse the effects of small or large Gaussian fields  $\sigma_l$  on the mean path of the landmark and final covariance (ellipses). We used 2000 timesteps to integrate the moment equation forwards from  $t = 0$  to  $t = 1$ . The initial momenta were found using a shooting method in the deterministic landmark equation. We display on the other two panels a zoom on two of the simulations of the left panel and compare the estimation of the final covariance from a Monte Carlo sampling (magenta) and from the solution of the moment equation (red) for two values of  $r_l$ . The black density represents the probability distribution of the landmark estimated from samples, and the dashed lines two level sets.

This experiment is displayed in Figure 2(a) where we used two arrays of four by four noise fields with either  $\lambda_l = (0.08, 0)$  or  $\lambda_l = (0, 0.08)$  and three values of the noise radius  $r_l = 0.5, 0.05, 0.03$ . For large values of  $r_l$ , the noise is mostly uniform and the gradients of the  $\sigma_l$  are negligible. The only term contributing to the final covariance of the landmark is therefore  $\langle \sigma_l^\alpha(\mathbf{q}_i) \sigma_l^\beta(\mathbf{q}_j) \rangle$ . This term only has a linear effect on the covariance which is thus an ellipse proportional to the noise fields. Here the noise has equal strength in both the  $x$  and  $y$  coordinate thus we observe a circle. For smaller values of  $r_l$ , the gradient of  $\sigma_l$  is large enough for the other term in the momentum equation which couple the momentum and the gradient of  $\sigma$  to affect the moment dynamics. This effect is shown in Figure 2(a) where the covariance has a larger value in the direction of the gradient of  $\sigma_l$  than in the other directions. This is explained by the fact that this coupling is of the form  $\frac{\partial}{\partial q_i}(\sigma_l(\mathbf{q}_i) \cdot \mathbf{p}_i)$ , thus the ellipse is in the direction of the gradient, not the momenta. Notice that there should be some noise in the direction of the momenta for this term

to have an effect. Using the same experiment, we compared the estimation of the covariance from the moment equation with a direct sampling obtained by solving the stochastic landmarks equations. We did this experiment for  $r_l = 0.5$  and  $r_l = 0.03$  in Figure 2(b),2(c). The left panel with  $r_l = 0.5$  shows an excellent agreement between the two methods but the right panel with  $r_l = 0.03$  shows differences. This type of error in the estimation of the covariance is explained by the fact that the final distribution has a large skewness. This effect is not captured by the moment equations as we neglected the effects of order higher than 2, and the skewness is a third order effect described by terms such as  $\Delta_3 \langle q_i^\alpha q_j^\beta q_k^\gamma \rangle$ . Nevertheless, the final covariance is close enough to the correct one to be able to use it in the estimation of the noise fields. This demonstrates that even in rather extreme cases, which are not realistic for applications, the second order approximation used to derive the moment equation still produces reliable results.

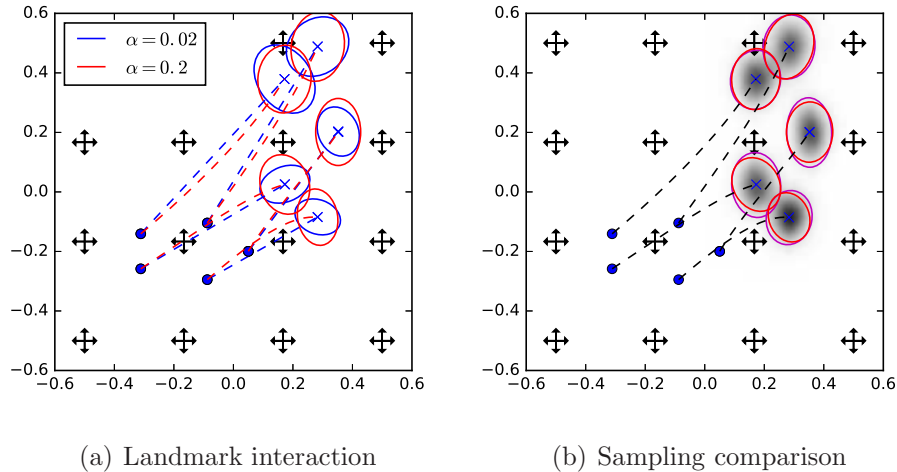


FIGURE 3. In these two panels, we present a study similar to Figure 2(a) but with 5 interacting landmarks. On the left, we illustrate the effect of varying the landmark length scale  $\alpha$ , and, on the right, we compare the result of the moment equation and a Monte Carlo simulation in the case of  $\alpha = 0.1$ . As before, the black density plot shows the probability density of the landmarks, the magenta curve the covariance from sampling and the red curve the covariance from the moment equation.

We did a similar experiment but with 5 interacting landmarks arranged in an ellipse configuration and with initial conditions obtained from the deterministic shooting method such that the endpoint of the deterministic landmark equations match another ellipse. We display these experiments in Figure 3 with the same noise as in the previous tests and with  $r_l = 0.2$ . We modified here the landmark interaction length scale  $\alpha$  from  $\alpha = 0.02$  (no-interactions) to  $\alpha = 0.2$  (neighbours interactions) to see the effect of the noise with the landmark interactions. Due to the different length scales, the trajectories to the target ellipse are slightly different so the landmarks will be subject to different noise. The larger length scale has the effect of reducing the differences between the covariances of interacting landmarks.

**5.2. Bridge Sampling.** Here, we aim at visualising the effect of the constructed bridge sampling scheme. In Figure 4, the effect of the guiding term is visualized on a sample path. At  $t = T/2$ , the predicted endpoint  $\phi_{t,T}(\hat{\mathbf{q}}(t), \hat{\mathbf{p}}(t))$  is calculated and the difference  $\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \mathbf{v}$  is used to guide the evolution of the path towards the target  $\mathbf{v}$ . The guiding term ensures that  $\hat{\mathbf{q}}$  will hit  $\mathbf{v}$  almost surely at time  $T$ . Notice that the difference  $\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \mathbf{v}$  is generally much smaller than the difference  $\hat{\mathbf{q}} - \mathbf{v}$ . The introduction of  $\phi_{t,T}$  therefore implies that the process is modified less giving more likely bridges. Without  $\phi_{t,T}$ , the process is generally attracted to quickly towards the target as can be seen by the landmarks at  $t = 0.5$  being almost at their final positions in Figure 4 (b). The path thus overshoots the target. This effect is not present when using  $\phi_{t,T}$  in Figure 4 (a).

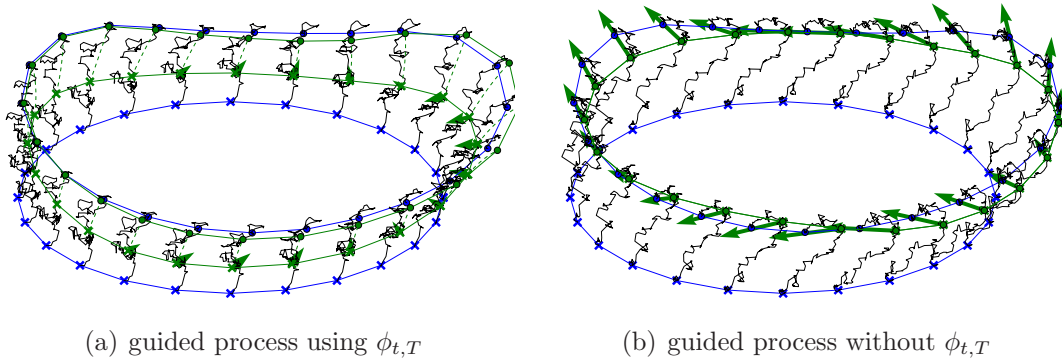


FIGURE 4. (a) Visualization of the process (3.15). From the initial landmark configuration  $\mathbf{q}(0)$  (blue crosses), the target  $\mathbf{v}$  (blue dots) is hit using the modified process  $(\hat{\mathbf{q}}, \hat{\mathbf{p}})$  (black lines:  $\hat{\mathbf{q}}$ ). At time  $t = T/2$ ,  $\phi_{t,T}(\hat{\mathbf{q}}(t), \hat{\mathbf{p}}(t))$  is calculated (green dots) and the process is guided by  $-(T-t)^{-1}\Sigma\Sigma_{\mathbf{q}}^{\dagger}(\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \mathbf{v})$  ( $\mathbf{q}$  part: green arrows, length doubled for visualization). The use of  $\phi_{t,T}$  implies small guiding and high probability sample bridges. (b) Similar setup but using the guiding term  $-(T-t)^{-1}\Sigma\Sigma_{\mathbf{q}}^{\dagger}(\hat{\mathbf{q}} - \mathbf{v})$  without  $\phi_{t,T}$ . The momentum couples with the guiding term, and, intuitively, the path travels too fast towards the target ( $\mathbf{q}$  at  $t = T/2$  much closer than halfway towards  $\mathbf{v}$ ) and overshoots. This effect gives low probability sample bridges and the guiding term (green arrows) is much larger than in (a).

**5.3. Estimating the noise amplitudes.** We here aim at estimating the noise amplitude from sampled data using both the method of moments and maximum likelihood.

We first use the genetic algorithm of [SP97] called differential evolution algorithm to minimise the cost function  $C$  in (4.8). This algorithm has in experiments proven successful in avoiding local minima during the optimisation. We compared it with the standard BFGS gradient descent algorithm with a single landmark in Figure 5. This algorithm relies on the Jacobian of the cost functional computed symbolically using the Theano package of [The16]. It is able to estimate the noise amplitude along the trajectory of the landmark where the signal from the gradient of  $C$  is

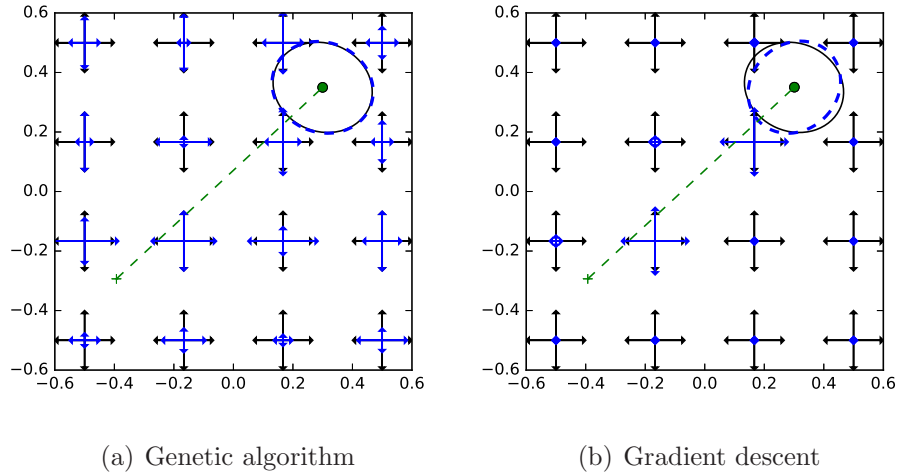


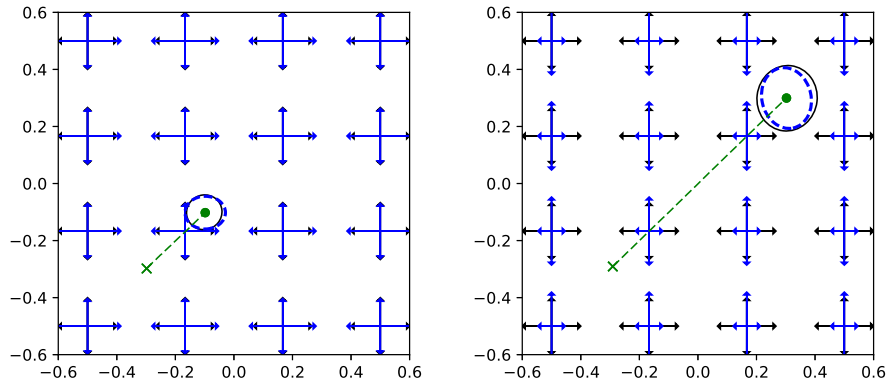
FIGURE 5. This figure shows the results of estimating parameters of the  $\sigma_l$  fields with the moment equation. Black arrows: The original  $\sigma_l$ . Blue arrows: The estimated  $\sigma_l$ . The error in the final covariance for the differential equation genetic algorithm is of the order of  $10^{-10}$  and for the BFGS algorithm, it is of the order of  $5 \cdot 10^{-2}$ .

the strongest. For the other regions of the image, the algorithm cannot detect any signal to update the noise fields. The genetic algorithm can overcome this issue as it is based on evolving a population of solution which randomly spans the entire parameter space. In this way, the solution obtained is a better approximation of the global minimum of  $C$ . It is interesting that even if the final moment of figure 5 are well matched with the genetic algorithm, but the noise amplitude is not perfectly recovered. This illustrates the expected degeneracy of this model for a low number of landmarks. When more landmarks are added, the noise amplitude estimation is closer to the expected one, see figure 7 below.

In Figure 6, the same experiment is performed with MLE and the bridge sampling scheme. The noise kernels are in this experiment cubic B-splines placed in a grid providing a partition of unity. In the optimisation,  $\lambda_l$  are fixed to be equal for all  $l = 1, \dots, J$  implying that the total noise variance will be uniform at each point of the domain. The figure shows the experiment performed with low momentum (Figure 6 (a)) and high momentum (Figure 6 (b)). In the low momentum case, the noise parameters are estimated correctly and the sample covariance with the estimated parameters matches the covariance of the original samples. The SDE (3.15) is here used for the bridge sampling scheme. In contrast to the previous method, the algorithm is now optimising for the maximum likelihood of the samples and not directly for matching the final covariance. A higher difference in the endpoint covariance is, therefore, to be expected.

With higher initial momentum, the coupling between the guidance and noise makes the scheme (3.15) overestimate the variance. Instead, the guidance term (3.16) is used. Notice that even though the sample covariance with the estimated parameters matches the covariance of the original samples, the estimated  $\lambda_l$  are





(a) Landmark and estimated noise, low momentum (b) Landmark and estimated noise, high momentum

FIGURE 6. Similarly to Figure 5 but the noise amplitudes are here estimated using maximum likelihood with the bridge sampling scheme. (a) The parameters are estimated correctly in the low momentum setting. (b) While the sample covariance matches the covariance of the original data in the high momentum case, the estimated parameters are different from the original.

different than the original values. This indicates that the maximum likelihood estimate of the parameters may not match the original setting in the highly nonlinear case occurring when the coupling between noise and momentum is high. Because of the nonlinearity, the noise is able to generate horizontal variation in the position of the final the landmark even though the variation with the estimated parameters are mainly vertical along the trajectory.

Figure 7, Figure 8 and Figure 9 show the result of noise estimation using different configurations of the ellipse using both the method of moments and MLE. The noise parameters  $\lambda_l$  are allowed to vary with  $l$  in both cases giving spatially non-uniform noise amplitude. The algorithms find the correct noise parameters in the areas covered by the landmark trajectories.

## 6. DISCUSSION AND OUTLOOK

As the first topic of this work, we attempted to answer the question of how to include stochasticity and uncertainty in the framework of large deformation matching in a systematic and geometrically consistent way. In section 2, we exposed a general theory of stochastic deformations in the LDDMM framework, based on the momentum map representation of images of [BGBHR11], by introducing spatially correlated time-dependent noise in the reconstruction relation used to compute the deformation map from its velocity field. By doing so, most of the advantages of the theory of reduction by symmetry remain, in particular, the possibility of applying this generic stochastic model to many data structures. The general dynamical equation is the stochastic EPDiff equation, and the noise appears in a particular

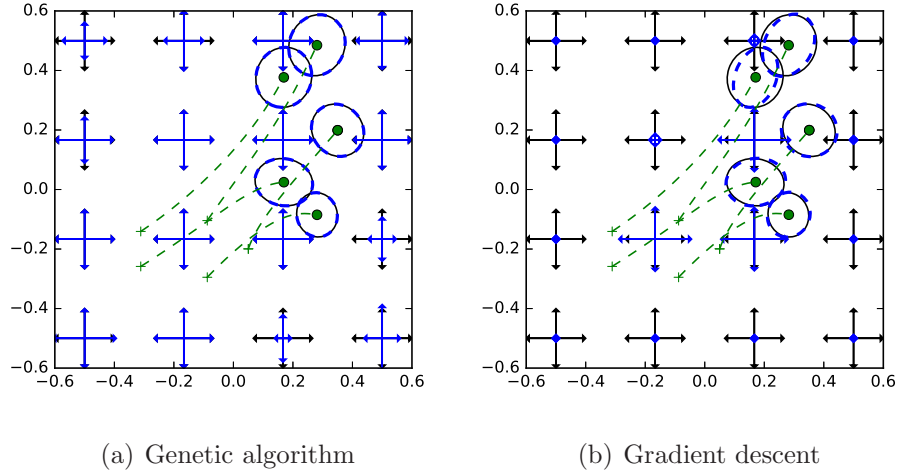


FIGURE 7. This figure shows the result of noise estimation using the moment equation as in Figure 5 but for the ellipse experiment. The error in the final covariance for the differential equation genetic algorithm is of the order of  $10^{-9}$  and for the BFGS algorithm it is of the order of  $5 \cdot 10^{-3}$ .

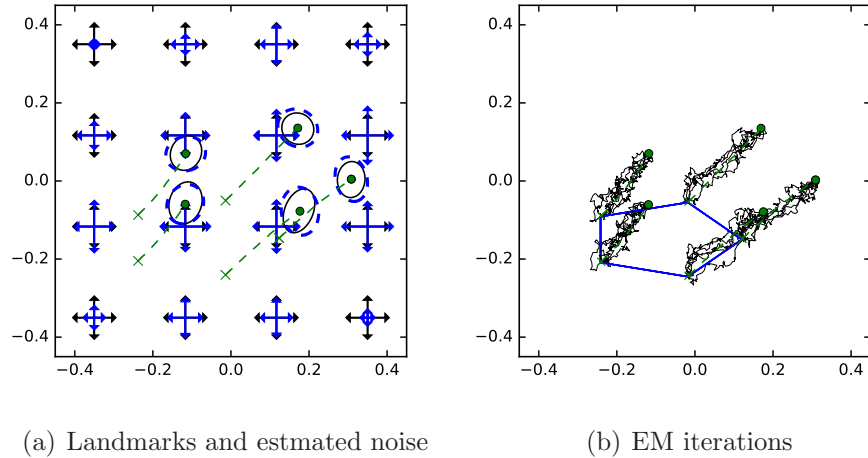


FIGURE 8. (a) Setup as Figure 6 but with five landmarks in an ellipsis configuration. (b) Examples of simulated bridges as used in the approximation of the  $Q$  function in the EM procedure.

multiplicative form with spatial correlation encoded in a set of spatially dependent functions  $\sigma_l$ . The key feature of this noise is that it preserves the structure of the original equation provided by the reduction by symmetry and in particular the momentum map allowing for both exact and inexact matching.

The question of local in time existence and uniqueness of this equation is important but not treated in this work. We refer to [BFM16] for such a study for the 2d Euler equation and to [CFH17] for the 3d case. Another possible extension is to consider an infinite number of  $\sigma_l$  fields with an infinite dimensional Wiener process

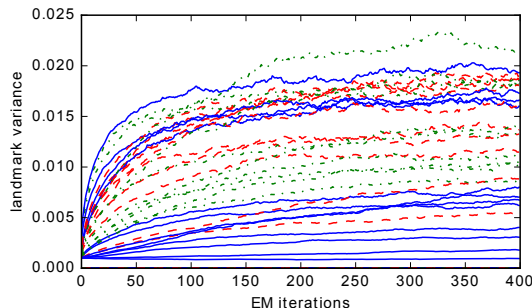


FIGURE 9. For the experiment in Figure 8, evolution of the estimated parameters  $\lambda_l$  in the EM procedure as a function of iterations.

for the stochastic EPDiff equation as investigated by [Via13], also in the context of stochastic shape analysis. We considered time-independent  $\sigma_l$  fields. However, there are several approaches for making these fields time dependent beside simply prescribing them as functions of time. Some of these other approaches were derived by [GBH17] in the context of stochastic fluid dynamics. In particular, the idea of having the noise fields being carried by the deformation could be of interest in this context as well. Yet another possibility could be to have two different types of noise fields, one modelling small-scale noise correlation and the other larger scale noise correlations. In this case, it would make sense for the small scale variability to be advected by the large-scale deformation, similarly to the multi-scale model of [HT12].

After defining the general model in section 2, we applied it to exact landmark matching in section 3, which is the simplest, yet non-trivial application of the LD-DMM framework. This approach allowed investigation of the effects of the noise on large deformation matching in a finite dimensional model. Introducing the noise in both the momentum and the position equations of the landmarks made the landmark trajectories rougher than they would have been, had the noise been only in the momentum equation. The noise in the position equation also increased the flexibility for controlling the landmark trajectories. This property was used to derive a scheme for simulating diffusion bridges with corresponding sampling correction factor that allowed evaluation of expectations with respect to the original conditioned landmark dynamics. In addition, we used the finite dimensionality of the system to derive the Fokker-Planck equation.

Some modifications to the standard theory of diffusion bridges were made to fit the case of landmark dynamics and to improve the speed and accuracy of the estimation of expectations over conditioned landmarks trajectories. The landmarks represent the simplest cases for numerical shape analysis, especially in the context of stochastic systems. We used a simple Heun method to solve the stochastic landmark equations. Higher order integration schemes could be used such as the stochastic variational integrators of [HT16a]. The next step in extending the landmark example is to allow for inexact matching and to study the trade-off between the effect of noise and the tolerance of the matching. Several issues regarding ergodicity and other properties of the Kolmogorov operator were left open in this paper, whose future treatments

could add to the theoretical understanding of the model. Finally, the stochastic LDDMM framework can be applied to other types of data structures, in particular to images with inexact matching as originally done in [BMTY05]. Studying the effects of the stochastic model on other nonlinear data structures such as curves or surfaces would also be of great interest for future works.

As a second topic, we raised the question of determining the noise correlation from data sets which would allow the theory of stochastic deformations to be used with observed data. We developed two independent methods which we implemented and applied to several test examples. First, the moment equation allows matching of the sample moments. It is deterministic, making optimisation of the noise parameters stable and efficient, and it does not require special conditions on the noise fields. Its accuracy depends on the approximation order in the moment equation. Scaling the moment equation to a large number of landmarks or continuous shapes such as curves may be challenging as well as optimising for a high number of unknown parameters. In the presented landmark experiments, the approach allowed us to reliably estimate the underlying noise.

The second method is the MLE optimisation, a Monte Carlo method which evaluates expectations over conditioned stochastic trajectories. The bridge sampling scheme we used requires the noise fields to span the entire  $\mathbf{q}$ -space to allow guiding the landmarks towards their target. With high nonlinearity as may happen with large initial momentum and high gradients of the noise fields, guiding the trajectories towards their target with high-probability bridges can be challenging. In general, the stochastic nature of the algorithm makes it harder to control than the matching provided by the moment equation. The bridge sampling scheme can be interpreted as a gradient flow, as discussed in [AHPS17], when applied to images. It allows the likelihood of observed images to be evaluated without a prior image registration step. The method may thus be applicable to image analysis problems, and more generally for inexact matching of shapes in which case the requirement of the noise to span the  $\mathbf{q}$ -space may be relaxed.

The inference of noise parameters treated here can be extended to more general statistical inference problems on shape spaces. Inferring the initial  $\mathbf{q}_0$  positions can be regarded as estimating a most-likely mean, thereby drawing similarities to the Fréchet mean [Fré48] and to means defined by maximum likelihood of probability distributions in nonlinear spaces [Som15]. When generalised to images, the approach can be used for simultaneous estimation of template images [JDJG04], possible time-dependent transformations in the momentum as caused for example by disease processes [MF12], and population variation in the spatial noise correlation.

*Acknowledgements.* We are grateful to M. Bruveris, M. Bauer, A. Pai, N. Ganaba, C. Tronci and T. Tyranowski for helpful discussions of this material. AA acknowledges partial support from an Imperial College London Roth Award. AA and DH are partially supported by the European Research Council Advanced Grant 267382 FCCA held by DH. DH is also grateful for support from EPSRC Grant EP/N023781/1. SS is partially supported by the CSGB Centre for Stochastic Geometry and Advanced Bioimaging funded by a grant from the Villum foundation. The research was partially completed while the

authors were visiting the Institute for Mathematical Sciences, National University of Singapore in 2016.

## REFERENCES

- [AAT07] Stephanie Allasonnière, Yali Amit, and Alain 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.
- [ACH16] Alexis Arnaudon, Alex L Castro, and Darryl D Holm. Noise and dissipation on coadjoint orbits. *arXiv preprint arXiv:1601.02249*, 2016.
- [AHPS17] Alexis Arnaudon, Darryl D Holm, Akshay Pai, and Stefan Sommer. A stochastic large deformation model for computational anatomy. In *Information Processing for Medical Imaging (IPMI)*, 2017.
- [BFM16] Zdzisław Brzeźniak, Franco Flandoli, and Mario Aureliani. Existence and uniqueness for stochastic 2D Euler flows with bounded vorticity. *Archive for Rational Mechanics and Analysis*, 221(1):107–142, 2016.
- [BFS16] Mogens Bladt, Samuel Finch, and Michael Sørensen. Simulation of multivariate diffusion bridges. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):343–369, 2016.
- [BGBHR11] Martins Bruveris, François Gay-Balmaz, Darryl D Holm, and Tudor S Ratiu. The momentum map representation of images. *Journal of Nonlinear Science*, 21(1):115–150, 2011.
- [BH15] Martins Bruveris and Darryl D Holm. Geometry of image registration: The diffeomorphism group and momentum maps. In *Geometry, Mechanics, and Dynamics*, pages 19–56. Springer, 2015.
- [BMTY05] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.
- [CFH17] Dan Crisan, Franco Flandoli, and Darryl D Holm. Local well posedness for 3D stochastic Euler equations. *in preparation*, 2017.
- [CH93] Roberto Camassa and Darryl D Holm. An integrable shallow water equation with peaked solitons. *Physical Review Letters*, 71:1661–1664, Sep 1993.
- [CRM96] Gary E. Christensen, Richard Rabbitt, and Michael I. Miller. Deformable templates using large deformation kinematics. *Image Processing, IEEE Transactions on*, 5(10), 1996.
- [DGM98] Paul Dupuis, Ulf Grenander, and Michael I. Miller. Variational Problems on Flows of Diffeomorphisms for Image Matching. *Quarterly of applied mathematics*, 1998.
- [DH06] Bernard Delyon and Ying Hu. Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Processes and their Applications*, 116(11):1660 – 1675, 2006.
- [DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society, series B*, 39(1):1–38, 1977.
- [DS08] Sophie Donnet and Adeline Samson. Parametric inference for mixed models defined by stochastic differential equations. *ESAIM: Probability and Statistics*, 12:196–218, 2008.
- [Fré48] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10:215–310, 1948.
- [GBH17] François Gay-Balmaz and Darryl D Holm. Stochastic geometric models with non-stationary spatial correlations in lagrangian fluid flows. *arXiv preprint arXiv:1609.00463*, 2017.
- [Gir60] I. Girsanov. On Transforming a Certain Class of Stochastic Processes by Absolutely Continuous Substitution of Measures. *Theory of Probability & Its Applications*, 5(3):285–301, 1960.

- [Gre94] Ulf Grenander. *General Pattern Theory: A Mathematical Study of Regular Structures*. Oxford University Press, USA, February 1994.
- [HLS90] DD Holm, WP Lysenko, and JC Scovel. Moment invariants for the vlasov equation. *Journal of mathematical physics*, 31(7):1610–1615, 1990.
- [HM05] Darryl D. Holm and Jerrold E Marsden. Momentum maps and measure-valued solutions (peakons, filaments, and sheets) for the epdiff equation. In *The breadth of symplectic and Poisson geometry*, pages 203–235. Springer, 2005.
- [HMR98] Darryl D Holm, Jerrold E Marsden, and Tudor S Ratiu. The Euler–Poincaré equations and semidirect products with applications to continuum theories. *Advances in Mathematics*, 137(1):1 – 81, 1998.
- [Hol11] Darryl D. Holm. *Geometric Mechanics - Part I: Dynamics and Symmetry*. Imperial College Press, London : Hackensack, NJ, 2 edition edition, 2011.
- [Hol15] Darryl D Holm. Variational principles for stochastic fluid dynamics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2176):20140963, 2015.
- [HPT07] Darryl D. Holm, Vakhtang Putkaradze, and Cesare Tronci. Geometric dissipation in kinetic equations. *Comptes Rendus Mathématique*, 345(5):297 – 302, 2007.
- [HPT10] Darryl D Holm, Vakhtang Putkaradze, and Cesare Tronci. Double-bracket dissipation in kinetic theory for particles with anisotropic interactions. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 2010.
- [HT12] Darryl D Holm and Cesare Tronci. Multiscale turbulence models based on convected fluid microstructure. *Journal of Mathematical Physics*, 53(11):115614, 2012.
- [HT16a] Darryl D Holm and Tomasz M Tyranowski. Stochastic discrete hamiltonian variational integrators. *arXiv preprint arXiv:1609.00463*, 2016.
- [HT16b] Darryl D Holm and Tomasz M Tyranowski. Variational principles for stochastic soliton dynamics. *Proc. R. Soc. A*, 472(2187):20150827, 2016.
- [JDJG04] Sarang Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:151–160, 2004.
- [JS14] Henry O Jacobs and Stefan Horst Sommer. Higher-order spatial accuracy in diffeomorphic image registration. *Geometry, Imaging and Computing*, 2014.
- [KK11] Mackillo Kira and Stephan W Koch. *Semiconductor quantum optics*. Cambridge University Press, 2011.
- [KSPR17] Line Kühnel, Stefan Sommer, Akshay Pai, and Lars Lau Raket. Most likely separation of intensity and warping effects in image registration. *SIAM Journal on Imaging Sciences*, *in press*, 2017.
- [Mar11] Jean-Louis Marchand. Conditioning diffusions with respect to partial observations. *arXiv*, 2011. arXiv: 1105.1608.
- [MBC<sup>+</sup>97] M. Miller, A. Banerjee, G. Christensen, S. Joshi, N. Khaneja, U. Grenander, and L. Matejic. Statistical methods in computational anatomy. *Statistical Methods in Medical Research*, 6(3):267–299, 1997.
- [MF12] Prasanna Muralidharan and P. Thomas Fletcher. Sasaki Metrics for Analysis of Longitudinal Data on Manifolds. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012:1027–1034, 2012.
- [MMTY08] Jun Ma, Michael I. Miller, Alain Trouvé, and Laurent Younes. Bayesian template estimation in computational anatomy. *NeuroImage*, 42(1):252–261, 2008.
- [MR99] Jerrold E. Marsden and Tudor S. Ratiu. *Introduction to Mechanics and Symmetry*, volume 17 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 1999.
- [MS16] Stephen Marsland and Tony Shardlow. Langevin equations for landmark image registration with uncertainty. *arXiv preprint arXiv:1605.09276*, 2016.
- [MTY02] Michael I Miller, Alain Trouvé, and Laurent Younes. On the metrics and Euler–Lagrange equations of computational anatomy. *Annual Review of Biomedical Engineering*, 4:375–405, 2002.

- [NH98] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.
- [Øk03] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, 2003.
- [PR12] Omiros Papaspiliopoulos and Gareth O. Roberts. Importance sampling techniques for estimation of diffusion models. In *Statistical Methods for Stochastic Differential Equations*. Chapman & Hall/CRC Press, 2012.
- [RSM14] Lars Lau Raket, Stefan Sommer, and Bo Markussen. A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data. *Pattern Recognition Letters*, 38:1–7, 2014.
- [Som15] Stefan Sommer. Anisotropic Distributions on Manifolds: Template Estimation and Most Probable Paths. In *Information Processing in Medical Imaging*, volume 9123 of *Lecture Notes in Computer Science*, pages 193–204. Springer Berlin Heidelberg, 2015.
- [SP97] Rainer Storn and Kenneth Price. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optim.*, 11(4):341–359, 1997.
- [Sus73] Héctor J Sussmann. Orbits of families of vector fields and integrability of distributions. *Transactions of the American Mathematical Society*, 180:171–188, 1973.
- [The16] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, may 2016.
- [Tro95] Alain Trouvé. An infinite dimensional group approach for physics based models in pattern recognition. *preprint*, 1995.
- [TV12] Alain Trouvé and François-Xavier Vialard. Shape splines and stochastic shape evolutions: a second order point of view. *Quarterly of Applied Mathematics*, 70(2):219–251, 2012.
- [Via13] François-Xavier Vialard. Extension to infinite dimensions of a stochastic second-order model associated with shape splines. *Stochastic Processes and their Applications*, 123(6):2110–2157, 2013.
- [VMYT04] M. Vaillant, M.I. Miller, L. Younes, and A. Trouvé. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23(Supplement 1):S161–S169, 2004.
- [YAM09] Laurent Younes, Felipe Arrate, and Michael I. Miller. Evolutions equations in computational anatomy. *NeuroImage*, 45(1, Supplement 1):S40–S50, 2009.
- [You10] Laurent Younes. *Shapes and Diffeomorphisms*. Springer, 2010.
- [ZSF13] Miaomiao Zhang, Nikhil Singh, and P Thomas Fletcher. Bayesian estimation of regularization and atlas building in diffeomorphic image registration. In *International Conference on Information Processing in Medical Imaging*, pages 37–48. Springer, 2013.

## APPENDIX A. BRIDGE SAMPLING

We here follow parts of the program in [DH06] and the later paper [Mar11] to argue for the a.s. hitting of a target  $\mathbf{v}$  for the guided process (3.15) and to find the correction term  $\varphi(\mathbf{q}, \mathbf{p})$ . We will explicitly derive the correction term following [DH06, Theorem 5]. The guided SDE (3.15) differs from the previous schemes in using the function  $\phi_{t,T} : \mathbb{R}^{2dN} \rightarrow \mathbb{R}^{dN}$  to predict the endpoint, and, importantly, in that the diffusion field  $\Sigma$  is not invertible resulting in a hypoelliptic diffusion. We handle the first issue by writing the SDE in a general form presented in [Mar11] that also deals with the unboundedness of the drift term  $b(\hat{\mathbf{q}}, \hat{\mathbf{p}})$  coming from the momentum of the landmarks. The result in [Mar11] however does not handle the hypoellipticity of the process. We do not here argue for the  $t \rightarrow T$  limit that in the elliptic case follows from [DH06, Mar11].

We let the function  $\phi_{t,T} : \mathbb{R}^{2dN} \rightarrow \mathbb{R}^{dN}$  be the solution of the deterministic landmark dynamics (2.3) at time  $T$  and use the notation of Theorem 3.4. Because the drift  $b(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t)$  in (3.6) is unbounded, we let  $\tilde{b}$  be an approximation of  $b$  so that the  $\mathbf{q}$ -part  $\tilde{b}_{\mathbf{q}}$  is bounded on  $\mathbb{R}^{dN}$  and  $\tilde{b}$  is in the range of  $\Sigma(\mathbf{q}, \mathbf{p})$  for all  $(\mathbf{q}, \mathbf{p})$ . We let  $\tilde{\phi}_{t,T}$  be given by the corresponding deterministic solution using  $\tilde{b}$  in 2.3. Then  $\partial_t \tilde{\phi}_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) = -\tilde{b}(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t)$  is bounded and the process  $\frac{\phi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \hat{\mathbf{q}}}{T-t}$  is defined, bounded and continuous on  $[0, T]$ . The SDE

$$\begin{pmatrix} d\hat{\mathbf{q}} \\ d\hat{\mathbf{p}} \end{pmatrix} = \tilde{b}(\hat{\mathbf{q}}, \hat{\mathbf{p}})dt - \frac{\Sigma(\mathbf{q}, \mathbf{p})\Sigma_{\mathbf{q}}(\mathbf{q})^\dagger(\hat{\mathbf{q}} - \mathbf{v})}{T-t}dt + \Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})dW \quad (\text{A.1})$$

differs from the SDE (3.15) by

$$(b(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \tilde{b}(\hat{\mathbf{q}}, \hat{\mathbf{p}}))dt - \frac{\Sigma(\mathbf{q}, \mathbf{p})\Sigma_{\mathbf{q}}(\mathbf{q})^\dagger(\varphi_{t,T}(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \hat{\mathbf{q}})}{T-t}dt. \quad (\text{A.2})$$

As argued for [Mar11], A.1 has a unique solution satisfying  $\lim_{t \rightarrow T} \hat{\mathbf{q}} = v$  a.s., and the processes A.1 and (3.15) are absolutely continuous with respect to each other.

In fact, the correction term  $\varphi(\mathbf{q}, \mathbf{p})$  can be derived from [Mar11, Theorem 3] and the difference A.2. Instead, we below give a direct derivation in the landmark case that proves Theorem 3.4.

*Proof of Theorem 3.4.* Let  $f : W(\mathbb{R}^{2dN}) \rightarrow \mathbb{R}$  be positive and measurable. Following [DH06] and using the notation of Theorem 3.4, define

$$h(\mathbf{q}, \mathbf{p}, t) := -\frac{\Sigma_{\mathbf{q}}(\mathbf{q})^\dagger(\varphi_{t,T}(\mathbf{q}, \mathbf{p}) - \mathbf{v})}{T-t} = -\frac{\Sigma_{\mathbf{q}}(\mathbf{q})^\dagger \mathbf{q}_{\mathbf{v}}}{T-t},$$

noting that in the present case, we use the pseudo-inverse  $\Sigma_{\mathbf{q}}(\mathbf{q})^\dagger$  to define  $h$  since  $\Sigma_{\mathbf{q}}(\mathbf{q})$  is not invertible. If  $f$  vanishes on  $(t, T]$ ,  $t < T$ , we have from [DH06, Thm. 1]

$$\mathbb{E}_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})} (f(\hat{\mathbf{q}}, \hat{\mathbf{p}})\varphi^b(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t)) = \mathbb{E}_{(\mathbf{q}, \mathbf{p})} (f(\mathbf{q}, \mathbf{p}, t)\tilde{\varphi}(\mathbf{q}, \mathbf{p}, t)) \quad (\text{A.3})$$

where

$$\log \tilde{\varphi}(\mathbf{q}, \mathbf{p}, t) := \int_0^t h^T(\mathbf{q}, \mathbf{p}, s)dW - \frac{1}{2} \int_0^t \|h(\mathbf{q}, \mathbf{p}, s)\|^2 ds, \quad (\text{A.4})$$

$$\begin{aligned} \log \varphi^b(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t) &:= \int_0^t (b(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \tilde{b}(\hat{\mathbf{q}}, \hat{\mathbf{p}}))^T \Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})^\dagger d(\hat{\mathbf{q}}, \hat{\mathbf{p}}) \\ &\quad - \int_0^t \frac{1}{2} \|\Sigma(\hat{\mathbf{q}}, \hat{\mathbf{p}})^\dagger (b(\hat{\mathbf{q}}, \hat{\mathbf{p}}) - \tilde{b}(\hat{\mathbf{q}}, \hat{\mathbf{p}}))\|^2 ds. \end{aligned} \quad (\text{A.5})$$

We now define an intermediate function

$$g(\mathbf{q}, t) := \frac{\mathbf{q}_{\mathbf{v}}^T A(\mathbf{q}) \mathbf{q}_{\mathbf{v}}}{T-t}, \quad (\text{A.6})$$

and compute

$$dg(\mathbf{q}, t) = \frac{\mathbf{q}_{\mathbf{v}}^T A(\mathbf{q}) \mathbf{q}_{\mathbf{v}}}{(T-t)^2} dt + \frac{d(\mathbf{q}_{\mathbf{v}}^T A(\mathbf{q}) \mathbf{q}_{\mathbf{v}})}{T-t}.$$



Applying the product rule, we obtain for the second term

$$\begin{aligned} d(\mathbf{q}_v^T A(\mathbf{q}) \mathbf{q}_v) &= 2\mathbf{q}_v^T A(\mathbf{q}) d\mathbf{q}_v + \mathbf{q}_v^T (dA(\mathbf{q})) \mathbf{q}_v \\ &\quad + \sum_{i,j} 2\mathbf{q}_{v,i} [dA_{ij}(\mathbf{q}), d\mathbf{q}_j] + A_{ij}(\mathbf{q}) [d\mathbf{q}_i, d\mathbf{q}_j]. \end{aligned}$$

Writing (2.3) in integral form  $\phi_{t,T}(\mathbf{q}, \mathbf{p}) = (\mathbf{q}, \mathbf{p})^T + \int_t^T b(\mathbf{q}(s), \mathbf{p}(s)) ds$  and letting  $\mathbf{q}_q := \phi_{t,T}(\mathbf{q}, \mathbf{p})_q - \mathbf{q}$ , we have, using Itô calculus,

$$\begin{aligned} d\mathbf{q}_v &= d\mathbf{q} + \partial_t \mathbf{q}_q dt + \frac{1}{2} \sum_k \Sigma_q(\mathbf{q})_{ik} \Sigma_q(\mathbf{q})_{jk} (\partial_i \partial_j \mathbf{q}_q) dt + (D_q \mathbf{q}_q) d\mathbf{q} \\ &= \Sigma_q(\mathbf{q}) dW + (D_q \mathbf{q}_q) d\mathbf{q} + \frac{1}{2} \text{Tr} (\Sigma_q(\mathbf{q})^T (H_q(\mathbf{q}_q)) \Sigma_q(\mathbf{q})) dt, \end{aligned}$$

with  $H_q(\mathbf{q}_q)$  denoting the component-wise Hessian. By writing the process  $g(\mathbf{q}, t)$  in integral form, we get

$$\begin{aligned} \frac{1}{2} \int_0^t dg(\mathbf{q}, s) &= \int_0^t \frac{\mathbf{q}_v^T A(\mathbf{q}) \mathbf{q}_v}{2(T-s)^2} ds + \int_0^t \frac{\mathbf{q}_v^T \Sigma_q(\mathbf{q})^{\dagger,T} dW}{T-s} + \int_0^t \frac{\mathbf{q}_v^T (dA(\mathbf{q})) \mathbf{q}_v}{2(T-s)} \\ &\quad + \int_0^t \frac{\mathbf{q}_v^T A(\mathbf{q}) ((D_q \mathbf{q}_q) d\mathbf{q} + \frac{1}{2} \text{Tr} (\Sigma_q(\mathbf{q})_q^T (H_q(\mathbf{q}_q)) \Sigma_q(\mathbf{q}_q)) ds)}{T-s} \\ &\quad + \sum_{i,j} \int_0^t \frac{2\mathbf{q}_{v,i} [dA_{ij}(\mathbf{q}), d\mathbf{q}_v] + A_{ij}(\mathbf{q}) [d\mathbf{q}_v, d\mathbf{q}_v]}{2(T-s)} \end{aligned}$$

using that  $A(\mathbf{q}) \Sigma_q(\mathbf{q}) = \Sigma_q(\mathbf{q})^{\dagger,T}$ . Note that the first term  $-\int_0^t \frac{\mathbf{q}_v^T A(\mathbf{q}) \mathbf{q}_v}{2(T-s)^2} ds$  of the right hand side equals the term equals the term  $-\frac{1}{2} \int_0^t \|h(\mathbf{q}, \mathbf{p}, s)\|^2 ds$  in (A.4). Similarly, the second term  $-\int_0^t \frac{\mathbf{q}_v^T \Sigma_q(\mathbf{q})^{\dagger,T} dW}{T-s}$  equals  $\int_0^t h^T(\mathbf{q}, \mathbf{p}, s) dW$ . Rearranging terms and inserting in (A.4),

$$\begin{aligned} \log \tilde{\varphi}(\mathbf{q}, \mathbf{p}, t) &= \int_0^t h^T(\mathbf{q}, \mathbf{p}, s) dW - \frac{1}{2} \int_0^t \|h(\mathbf{q}, \mathbf{p}, s)\|^2 ds \\ &= -\frac{1}{2} g(\mathbf{q}(t), t) + \frac{1}{2} g(\mathbf{q}(0), 0) \\ &\quad + \int_0^t \frac{\mathbf{q}_v^T A(\mathbf{q}) ((D_q \mathbf{q}_q) d\mathbf{q} + \frac{1}{2} \text{Tr} (\Sigma_q(\mathbf{q})_q^T (H_q(\mathbf{q}_q)) \Sigma_q(\mathbf{q}_q)) ds)}{T-s} \\ &\quad + \int_0^t \frac{\mathbf{q}_v^T (dA(\mathbf{q})) \mathbf{q}_v}{2(T-s)} + \sum_{i,j} \int_0^t \frac{2\mathbf{q}_{v,i} [dA_{ij}(\mathbf{q}), d\mathbf{q}_v] + A_{ij}(\mathbf{q}) [d\mathbf{q}_v, d\mathbf{q}_v]}{2(T-s)}. \end{aligned}$$

Define the function  $\varphi(\mathbf{q}, \mathbf{p}, t)$  such that

$$\log \varphi(\mathbf{q}, \mathbf{p}, t) = -\log \tilde{\varphi}(\mathbf{q}, \mathbf{p}, t) - \frac{1}{2} g(\mathbf{q}(t), t) + \frac{1}{2} g(\mathbf{q}(0), 0) + \log \varphi(\mathbf{q}, \mathbf{p}, t)^b.$$

Then (A.3) gives

$$\mathbb{E}_{P_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}} (f(\hat{\mathbf{q}}, \hat{\mathbf{p}}) \varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t)) = e^{\frac{1}{2} g(\mathbf{q}(0), 0)} \mathbb{E}_{P_{(\mathbf{q}, \mathbf{p})}} \left( f(\mathbf{q}, \mathbf{p}) e^{-\frac{1}{2} g(\mathbf{q}(t), t)} \right).$$

Thus

$$\lim_{t \rightarrow T} \frac{\mathbb{E}_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}(f(\hat{\mathbf{q}}, \hat{\mathbf{p}})\varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t))}{\mathbb{E}_{(\hat{\mathbf{q}}, \hat{\mathbf{p}})}(\varphi(\hat{\mathbf{q}}, \hat{\mathbf{p}}, t))} = \lim_{t \rightarrow T} \frac{\mathbb{E}_{(\mathbf{q}, \mathbf{p})}\left(f(\mathbf{q}, \mathbf{p})e^{-\frac{1}{2}g(\mathbf{q}(t), t)}\right)}{E_{P_{(\mathbf{q}, \mathbf{p})}}\left(e^{-\frac{1}{2}g(\mathbf{q}(t), t)}\right)} = \mathbb{E}_{(\mathbf{q}, \mathbf{p})|v}(f(\mathbf{q}, \mathbf{p}))$$

where convergence of the right-most limit to the conditioned process follows from [DH06, Lemma 7].  $\square$

## APPENDIX B. MOMENT EQUATION FOR STOCHASTIC LANDMARK

**B.1. Cluster expansion method.** We explain the basics of this method, which can be found in more details in, for example, [KK11] with application in the context of semiconductor physics. This method is used when one seeks the dynamics of the expected value of  $N$  particles that we will write here  $\langle N \rangle$ . One cannot solve the complete system, especially if the number of particles is large, thus we want to approximate the expected value of products in terms of only a few independent variables. For this we apply the cluster expansion, which begins by writing

$$\langle 2 \rangle = \langle 2 \rangle_s + \Delta_2 \langle 2 \rangle := \langle 1 \rangle \langle 1 \rangle + \Delta_2 \langle 2 \rangle, \quad (\text{B.1})$$

The next decomposition is

$$\langle 3 \rangle = \langle 3 \rangle_s + \langle 1 \rangle \Delta_2 \langle 2 \rangle + \Delta_3 \langle 3 \rangle, \quad (\text{B.2})$$

and so on and so forth. We then only compute the dynamics for the singlets  $\langle 1 \rangle$  and the correlations, up to some chosen order. In the sequel, we will only consider the doublet correlations  $\Delta_2$ , and in this case we have the general decomposition

$$\langle N \rangle = \langle N \rangle_s + \langle N - 2 \rangle_s \Delta_2 \langle 2 \rangle + \langle N - 4 \rangle_s \Delta_2^2 \langle 2 \rangle + \sum_i \langle N - 2i \rangle_s \Delta_2^i \langle 2 \rangle + O(\Delta_3).$$

In the context of quantum mechanics, where the particle operators do not commute, extra care is needed especially for the sign of the term. Here we will consider  $q_i^\alpha$  and  $p_i^\alpha$  as our particles, and as they commute, the expansions are simpler than in [KK11]. We directly compute two of them for illustration, up to quadratic order,

$$\begin{aligned} \langle q_i^\alpha p_j^\beta \rangle &= \langle q_i^\alpha \rangle \langle p_j^\beta \rangle + \Delta_2 \langle q_i^\alpha p_j^\beta \rangle \\ \langle q_i^\alpha q_j^\beta p_k^\gamma \rangle &\approx \langle q_i^\alpha \rangle \langle q_j^\beta \rangle \langle p_k^\gamma \rangle + \langle q_i^\alpha \rangle \Delta_2 \langle q_j^\beta p_k^\gamma \rangle + \langle q_j^\beta \rangle \Delta_2 \langle q_i^\alpha p_k^\gamma \rangle + \langle p_k^\gamma \rangle \Delta_2 \langle q_i^\alpha q_j^\beta \rangle. \end{aligned}$$

This sort of expansion can fit a more geometrical framework, where the final equations for the first moment will preserve the original structure of the equations. This was developed first in [HLS90] and later in [HPT07, HPT10]. We will not use this method here for a good reason related to the form of the equations. A key step in these papers is to expand the expected value of the Hamiltonian in terms of a finite number of moments, to enable computation of the equation of motion. In our case, the Hamiltonian has a kernel function, which generally cannot be expanded in a finite sum of polynomial terms. By doing the computations directly, we will be able to do another approximation for the kernels, that is, we will assume that they commute with the operation of expectation. A more subtle approximation can be done using the heaviside function, but will give a much larger number of terms in the expansion, see appendix B.6.

To perform this expansion on the Fokker-Planck equation associated to the landmark dynamics we will use several simplifications:

- Gaussian noise fields  $\sigma_l$  in (4.1),
- for a kernel  $K(x)$ , we will assume that  $\langle K(x) \rangle \approx K(\langle x \rangle)$  and
- only the second order correlations  $\Delta_2$  will be considered in this expansion.

These assumptions can be relaxed but the resulting equation may be difficult to compute.

**B.2. First moments.** Recall the backward Kolmogorov operator on  $q_i^\alpha$

$$\mathcal{L}q_i^\alpha = \frac{\partial h}{\partial p_i^\alpha} + \frac{1}{2} \sum_{l,\gamma,\delta} \frac{\partial \sigma_l^\alpha(\mathbf{q}_i)}{\partial q_i^\gamma} \sigma_l^\gamma(\mathbf{q}_i), \quad (\text{B.3})$$

which is used to compute the time evolution of the singlet

$$\frac{d}{dt} \langle q_i^\alpha \rangle = \langle p_j^\alpha \rangle K(\langle q_i \rangle - \langle q_j \rangle) - \sum_{l,\gamma,\delta} \frac{1}{2\sigma_l^2} \sigma_l^\alpha(\langle \mathbf{q}_i \rangle) (\langle q_i^\gamma \rangle - \delta_l^\gamma) \sigma_l^\gamma(\langle \mathbf{q}_i \rangle). \quad (\text{B.4})$$

In this case, the equation only depends on the singlet of the momentum variable. We thus compute

$$\mathcal{L}p_i^\alpha = -\frac{\partial h}{\partial q_i^\alpha} + \frac{1}{2} \sum_{l,\gamma,\delta} p_j^\gamma \frac{\partial \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\delta} \frac{\partial \sigma_l^\delta(\mathbf{q}_i)}{\partial q_i^\alpha} - \frac{1}{2} \sum_{l,\gamma,\delta} p_i^\gamma \frac{\partial^2 \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha \partial q_i^\delta} \sigma_l^\delta(\mathbf{q}_i). \quad (\text{B.5})$$

which similarly gives the time evolution of the momentum singlet in two terms as

$$\frac{d}{dt} \langle p_i^\alpha \rangle = A_p + B_p,$$

where

$$A_p = \frac{1}{\alpha^2} \sum_{j,\gamma} K(\langle q_i \rangle - \langle q_j \rangle) \langle p_i^\gamma p_j^\gamma (q_i^\alpha - q_j^\alpha) \rangle$$

$$B_p = \sum_{l,\gamma} \frac{1}{2\sigma_l^2} \langle p_i^\gamma \rangle \sigma_l^\gamma(\langle \mathbf{q}_i \rangle) \sigma_l^\alpha(\langle \mathbf{q}_i \rangle).$$

We then expand  $A_p$  further using the cluster-expansion method on the triplet to get

$$A_p = \frac{1}{\alpha^2} \sum_{j,\gamma} K(\langle \mathbf{q}_i \rangle - \langle \mathbf{q}_j \rangle) (\langle p_i^\gamma p_j^\gamma q_i^\alpha \rangle - \langle p_i^\gamma p_j^\gamma \rangle \langle q_i^\alpha \rangle)$$

$$\approx \frac{1}{\alpha^2} \sum_{j,\gamma} K(\langle \mathbf{q}_i \rangle - \langle \mathbf{q}_j \rangle) \left( \langle p_i^\gamma \rangle \langle p_j^\gamma \rangle \langle q_i^\alpha \rangle + \Delta_2 \langle p_i^\gamma p_j^\gamma \rangle \langle q_i^\alpha \rangle + \langle p_i^\gamma \rangle \Delta_2 \langle p_j^\gamma q_i^\alpha \rangle \right.$$

$$\quad \left. + \Delta_2 \langle p_i^\gamma q_i^\alpha \rangle \langle p_j^\gamma \rangle - \langle p_i^\gamma \rangle \langle p_j^\gamma \rangle \langle q_j^\alpha \rangle - \Delta_2 \langle p_i^\gamma p_j^\gamma \rangle \langle q_j^\alpha \rangle \right.$$

$$\quad \left. - \langle p_i^\gamma \rangle \Delta_2 \langle p_j^\gamma q_j^\alpha \rangle - \Delta_2 \langle p_i^\gamma q_j^\alpha \rangle \langle p_j^\gamma \rangle \right).$$

Already this term depends on the mixed correlations which we will compute shortly, but we first compute the position correlation.

B.3.  **$\langle qq \rangle$  correlation.** Recall the formula of the Kolmogorov operator applied to  $q_i^\alpha q_j^\beta$ ,

$$\mathcal{L}(q_i^\alpha q_j^\beta) = q_i^\alpha \frac{\partial h}{\partial p_j^\beta} + \sum_l \sigma_l^\alpha(\mathbf{q}_i) \sigma_l^\beta(\mathbf{q}_j) + \frac{1}{2} \sum_l q_i^\alpha \sigma_l^\gamma(\mathbf{q}_j) \frac{\partial \sigma_l^\beta(\mathbf{q}_j)}{\partial q_j^\gamma} + T, \quad (\text{B.6})$$

which together with (4.5) gives the time evolution of the position correlation in the form

$$\frac{d}{dt} \Delta_2 \langle q_i^\alpha q_j^\beta \rangle = A_{qq} + B_{qq} + C_{qq},$$

where

$$\begin{aligned} A_{qq} &= \left\langle q_i^\alpha \frac{\partial h}{\partial p_j^\beta} \right\rangle - \langle q_i^\alpha \rangle \left\langle \frac{\partial h}{\partial p_j^\beta} \right\rangle + T \\ B_{qq} &= \sum_l \left\langle \sigma_l^\alpha(\mathbf{q}_i) \sigma_l^\beta(\mathbf{q}_j) \right\rangle \\ C_{qq} &= \frac{1}{2} \sum_l \left\langle q_i^\alpha \sigma_l^\gamma(\mathbf{q}_j) \frac{\partial \sigma_l^\beta(\mathbf{q}_j)}{\partial q_j^\gamma} \right\rangle - \frac{1}{2} \sum_l \langle q_i^\alpha \rangle \left\langle \sigma_l^\gamma(\mathbf{q}_j) \frac{\partial \sigma_l^\beta(\mathbf{q}_j)}{\partial q_j^\gamma} \right\rangle + T. \end{aligned}$$

We will denote by  $A$  the terms corresponding to the drift, by  $B$  the terms which are not present in the first moments equation, and by  $C$  the other terms which only depend on the noise and the derivative of the noise fields. We proceed by first approximating the expectation of the kernels to get

$$\begin{aligned} B_{qq} &\approx \sum_l \sigma_l^\alpha(\langle \mathbf{q}_i \rangle) \sigma_l^\beta(\langle \mathbf{q}_j \rangle) \\ C_{qq} &\approx -\frac{1}{2\alpha_l^2} \sum_{l,\gamma} \Delta_2 \langle q_i^\alpha q_j^\gamma \rangle \sigma_l^\gamma(\langle \mathbf{q}_j \rangle) \sigma_l^\beta(\langle \mathbf{q}_j \rangle) + T. \end{aligned}$$

where we also used the explicit form of  $\sigma_l$  as a Gaussian and its derivative. We will now approximate the  $A_{qq}$  term to get

$$\begin{aligned} A_{qq} &= \sum_k \left\langle q_i^\alpha p_k^\beta K(\mathbf{q}_j - \mathbf{q}_k) \right\rangle - \sum_k \langle q_i^\alpha \rangle \left\langle p_k^\beta K(\mathbf{q}_j - \mathbf{q}_k) \right\rangle + T \\ &\approx \sum_k \Delta_2 \left\langle q_i^\alpha p_k^\beta \right\rangle K(\langle \mathbf{q}_j \rangle - \langle \mathbf{q}_k \rangle) + T. \end{aligned}$$

It is now clear that the  $B$  term will linearly increase the position correlation, which will then exponentially increase by the  $C$  term and be affected by the momentum-position correlation by the  $A$  term. We now proceed by computing the momentum correlation.

B.4.  **$\langle pp \rangle$  correlation.** We compute the Kolmogorov operator on  $p_o^\alpha p_j^\beta$  to get

$$\begin{aligned} \mathcal{L}(p_i^\alpha p_j^\beta) &= -p_i^\alpha \frac{\partial h}{\partial q_j^\beta} - \frac{1}{2} \sum_{l,\gamma,\delta} p_i^\alpha p_j^\gamma \sigma_l^\delta(\mathbf{q}_j) \frac{\partial^2 \sigma_l^\gamma(\mathbf{q}_j)}{\partial q_j^\beta \partial q_j^\delta} + T \\ &\quad + \frac{1}{2} \sum_{l,\gamma,\delta} p_i^\delta p_j^\gamma \frac{\partial \sigma_l^\delta(\mathbf{q}_i)}{\partial q_i^\alpha} \frac{\partial \sigma_l^\gamma(\mathbf{q}_j)}{\partial q_j^\beta} + \frac{1}{2} \sum_{l,\gamma,\delta} p_i^\alpha p_j^\delta \frac{\partial \sigma_l^\delta(\mathbf{q}_j)}{\partial q_j^\gamma} \frac{\partial \sigma_l^\gamma(\mathbf{q}_j)}{\partial q_j^\beta} + T, \end{aligned}$$

and using (B.5), we obtain the time evolution of the correlation in three terms as

$$\frac{d}{dt} \Delta_2 \langle p_i^\alpha p_j^\beta \rangle = A_{pp} + B_{pp} + C_{pp},$$

where

$$\begin{aligned} A_{pp} &= - \left\langle p_i^\alpha \frac{\partial h}{\partial q_j^\beta} \right\rangle + \langle p_i^\alpha \rangle \left\langle \frac{\partial h}{\partial q_j^\beta} \right\rangle + T \\ B_{pp} &= \frac{1}{2} \sum_{l,\gamma,\delta} \left\langle p_i^\delta p_j^\gamma \frac{\partial \sigma_l^\delta(\mathbf{q}_i)}{\partial q_i^\alpha} \frac{\partial \sigma_l^\gamma(\mathbf{q}_j)}{\partial q_j^\beta} \right\rangle + T \\ C_{pp} &= -\frac{1}{2} \sum_{l,\gamma,\delta} \left\langle p_i^\alpha p_j^\gamma \sigma_l^\delta(\mathbf{q}_j) \frac{\partial^2 \sigma_l^\gamma(\mathbf{q}_j)}{\partial q_j^\beta \partial q_j^\delta} \right\rangle + \frac{1}{2} \sum_{l,\gamma,\delta} \langle p_i^\alpha \rangle \left\langle p_j^\gamma \sigma_l^\delta(\mathbf{q}_j) \frac{\partial^2 \sigma_l^\gamma(\mathbf{q}_j)}{\partial q_j^\beta \partial q_j^\delta} \right\rangle \\ &\quad + \frac{1}{2} \sum_{l,\gamma,\delta} \left\langle p_i^\alpha p_j^\delta \frac{\partial \sigma_l^\delta(\mathbf{q}_j)}{\partial q_j^\gamma} \frac{\partial \sigma_l^\gamma(\mathbf{q}_j)}{\partial q_j^\beta} \right\rangle - \frac{1}{2} \sum_{l,\gamma,\delta} \langle p_i^\alpha \rangle \left\langle p_j^\delta \frac{\partial \sigma_l^\delta(\mathbf{q}_j)}{\partial q_j^\gamma} \frac{\partial \sigma_l^\gamma(\mathbf{q}_j)}{\partial q_j^\beta} \right\rangle + T. \end{aligned}$$

We first approximate

$$\begin{aligned} C_{pp} &\approx \frac{1}{2} \sum_{l,\gamma,\delta} \frac{1}{\alpha_l^2} \sigma_l^\beta(q_j) \sigma_l^\gamma(q_j) \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \\ &\quad - \frac{1}{2} \sum_{l,\gamma,\delta} \sigma_l^\delta(q_j) \sigma_l^\gamma(q_j) \left( \langle p_i^\alpha p_j^\gamma (q_j^\beta - \delta_l^\beta)(q_j^\delta - \delta_l^\delta) \rangle - \langle p_i^\alpha \rangle \langle p_j^\gamma (q_j^\beta - \delta_l^\beta)(q_j^\delta - \delta_l^\delta) \rangle \right) \\ &\quad + \frac{1}{2} \sum_{l,\gamma,\delta} \sigma_l^\delta(q_j) \sigma_l^\gamma(q_j) \left( \langle p_i^\alpha p_j^\delta (q_j^\gamma - \delta_l^\gamma)(q_j^\beta - \delta_l^\beta) \rangle + \langle p_i^\alpha \rangle \langle p_j^\delta (q_j^\gamma - \delta_l^\gamma)(q_j^\beta - \delta_l^\beta) \rangle \right). \end{aligned}$$

The last two terms cancel as they are symmetric under the transpose operation because of the sum on the free indices, thus the  $C$  term is

$$C_{pp} \approx \frac{1}{2} \sum_{l,\gamma,\delta} \frac{1}{\alpha_l^2} \sigma_l^\beta(q_j) \sigma_l^\gamma(q_j) \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle + T. \quad (\text{B.7})$$

We proceed with the  $B_{pp}$  term, which is also symmetric under the transpose operation, thus giving the approximation

$$\begin{aligned}
B_{pp} &= \sum_{l,\gamma,\delta} \frac{1}{\alpha_l^4} \sigma_l^\delta(q_i) \sigma_l^\gamma(q_j) \left\langle p_i^\delta p_j^\gamma (q_i^\alpha - \delta_l^\alpha) (q_j^\beta - \delta_l^\beta) \right\rangle \\
&\approx \sum_{l,\gamma,\delta} \frac{1}{\alpha_l^4} \sigma_l^\delta(q_i) \sigma_l^\gamma(q_j) \left( \langle p_i^\delta \rangle \langle p_j^\gamma \rangle \langle q_i^\alpha \rangle \langle q_j^\beta \rangle + \Delta_2 \langle p_i^\delta p_j^\gamma \rangle \langle q_i^\alpha \rangle \langle q_j^\beta \rangle \right. \\
&\quad + \langle p_i^\delta \rangle \Delta_2 \langle p_j^\gamma q_i^\alpha \rangle \langle q_j^\beta \rangle + \langle p_i^\delta \rangle \langle p_j^\gamma \rangle \Delta_2 \langle q_i^\alpha q_j^\beta \rangle \\
&\quad + \Delta_2 \langle p_i^\delta q_i^\alpha \rangle \langle p_j^\gamma \rangle \langle q_j^\beta \rangle + \Delta_2 \langle p_i^\delta q_j^\beta \rangle \langle p_j^\gamma \rangle \langle q_i^\alpha \rangle \\
&\quad + \langle p_i^\delta \rangle \langle q_i^\alpha \rangle \Delta_2 \langle p_j^\gamma q_j^\beta \rangle + \Delta_2 \langle p_i^\delta p_j^\gamma \rangle \Delta_2 \langle q_i^\alpha q_j^\beta \rangle \\
&\quad + \Delta_2 \langle p_i^\delta q_i^\alpha \rangle \Delta_2 \langle p_j^\gamma q_j^\beta \rangle + \Delta_2 \langle p_i^\delta q_j^\beta \rangle \Delta_2 \langle p_j^\gamma q_i^\alpha \rangle \\
&\quad - \langle p_i^\delta \rangle \langle p_j^\gamma \rangle \langle q_i^\alpha \rangle \delta_l^\beta - \Delta_2 \langle p_i^\delta p_j^\gamma \rangle \langle q_i^\alpha \rangle \delta_l^\beta \\
&\quad - \langle p_i^\delta \rangle \Delta_2 \langle p_j^\gamma q_i^\alpha \rangle \delta_l^\beta - \Delta_2 \langle p_i^\delta q_i^\alpha \rangle \langle p_j^\gamma \rangle \delta_l^\beta \\
&\quad - \langle p_i^\delta \rangle \langle p_j^\gamma \rangle \langle q_j^\beta \rangle \delta_l^\alpha - \Delta_2 \langle p_i^\delta p_j^\gamma \rangle \langle q_j^\beta \rangle \delta_l^\alpha \\
&\quad - \langle p_i^\delta \rangle \Delta_2 \langle p_j^\gamma q_j^\beta \rangle \delta_l^\alpha - \Delta_2 \langle p_i^\delta q_j^\beta \rangle \langle p_j^\gamma \rangle \delta_l^\alpha \\
&\quad \left. + \langle p_i^\delta \rangle \langle p_j^\gamma \rangle \delta_l^\alpha \delta_l^\beta + \Delta_2 \langle p_i^\delta p_j^\gamma \rangle \delta_l^\alpha \delta_l^\beta \right).
\end{aligned}$$

We treat the two Hamiltonian terms separately by first writing them explicitly as

$$\begin{aligned}
A_{pp} &= \sum_{k,\gamma} \frac{1}{\alpha^2} \left\langle p_i^\alpha p_j^\gamma p_k^\gamma (q_j^\beta - q_k^\beta) K(q_j - q_k) \right\rangle \\
&\quad - \frac{1}{\alpha^2} \sum_{k,\gamma} \langle p_i^\alpha \rangle \left\langle p_j^\gamma p_k^\gamma (q_j^\beta - q_k^\beta) K(q_j - q_k) \right\rangle + T \\
&=: \sum_{k,\gamma} \frac{1}{\alpha^2} (A_{pp}^1 - A_{pp}^2) + T.
\end{aligned}$$

We expand the first term to arrive at

$$\begin{aligned}
A_{pp}^1 &\approx K(\langle q_j \rangle - \langle q_k \rangle) \left( \left\langle p_i^\alpha p_j^\gamma p_k^\gamma q_j^\beta \right\rangle - \left\langle p_i^\alpha p_j^\gamma p_k^\gamma q_k^\beta \right\rangle \right) \\
&\approx K(\langle q_j \rangle - \langle q_k \rangle) \left( \langle p_i^\alpha \rangle \langle p_j^\gamma \rangle \langle p_k^\gamma \rangle \langle q_j^\beta \rangle + \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \langle p_k^\gamma \rangle \langle q_j^\beta \rangle \right. \\
&\quad + \langle p_i^\alpha \rangle \Delta_2 \langle p_j^\gamma p_k^\gamma \rangle \langle q_j^\beta \rangle + \langle p_i^\alpha \rangle \langle p_j^\gamma \rangle \Delta_2 \langle p_k^\gamma q_j^\beta \rangle \\
&\quad + \Delta_2 \langle p_i^\alpha p_k^\gamma \rangle \langle p_j^\gamma \rangle \langle q_j^\beta \rangle + \Delta_2 \langle p_i^\alpha q_j^\beta \rangle \langle p_j^\gamma \rangle \langle p_k^\gamma \rangle \\
&\quad + \langle p_i^\alpha \rangle \langle p_k^\gamma \rangle \Delta_2 \langle p_j^\gamma q_j^\beta \rangle + \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \Delta_2 \langle p_k^\gamma q_j^\beta \rangle \\
&\quad + \Delta_2 \langle p_i^\alpha p_k^\gamma \rangle \Delta_2 \langle p_j^\gamma q_j^\beta \rangle + \Delta_2 \langle p_i^\alpha q_j^\beta \rangle \Delta_2 \langle p_k^\gamma p_j^\gamma \rangle \\
&\quad - \langle p_i^\alpha \rangle \langle p_j^\gamma \rangle \langle p_k^\gamma \rangle \langle q_k^\beta \rangle - \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \langle p_k^\gamma \rangle \langle q_k^\beta \rangle \\
&\quad - \langle p_i^\alpha \rangle \Delta_2 \langle p_j^\gamma p_k^\gamma \rangle \langle q_k^\beta \rangle - \langle p_i^\alpha \rangle \langle p_j^\gamma \rangle \Delta_2 \langle p_k^\gamma q_k^\beta \rangle \\
&\quad - \Delta_2 \langle p_i^\alpha p_k^\gamma \rangle \langle p_j^\gamma \rangle \langle q_k^\beta \rangle - \Delta_2 \langle p_i^\alpha q_k^\beta \rangle \langle p_j^\gamma \rangle \langle p_k^\gamma \rangle \\
&\quad - \Delta_2 \langle p_k^\gamma q_k^\beta \rangle \langle p_i^\alpha \rangle \langle p_k^\gamma \rangle - \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \Delta_2 \langle p_k^\gamma q_k^\beta \rangle \\
&\quad \left. - \Delta_2 \langle p_i^\alpha p_k^\gamma \rangle \Delta_2 \langle p_j^\gamma q_k^\beta \rangle - \Delta_2 \langle p_i^\alpha q_k^\beta \rangle \Delta_2 \langle p_k^\gamma p_j^\gamma \rangle \right),
\end{aligned}$$

and the second term

$$\begin{aligned}
A_{pp}^2 &\approx K(\langle q_j \rangle - \langle q_k \rangle) \left( \langle p_i^\alpha \rangle \langle p_j^\gamma p_k^\gamma q_j^\beta \rangle - \langle p_i^\alpha \rangle \langle p_j^\gamma p_k^\gamma q_k^\beta \rangle \right) \\
&\approx K(\langle q_j \rangle - \langle q_k \rangle) \left( \langle p_i^\alpha \rangle \langle p_j^\gamma \rangle \langle p_k^\gamma \rangle \langle q_j^\beta \rangle + \langle p_i^\alpha \rangle \Delta_2 \langle p_j^\gamma p_k^\gamma \rangle \langle q_j^\beta \rangle \right. \\
&\quad + \langle p_i^\alpha \rangle \Delta_2 \langle p_j^\gamma p_j^\beta \rangle \langle q_k^\gamma \rangle \langle p_i^\alpha \rangle \langle p_j^\gamma \rangle \Delta_2 \langle p_k^\gamma q_j^\beta \rangle \\
&\quad - \langle p_i^\alpha \rangle \langle p_j^\gamma \rangle \langle p_k^\gamma \rangle \langle q_k^\beta \rangle - \langle p_i^\alpha \rangle \Delta_2 \langle p_j^\gamma p_k^\gamma \rangle \langle q_k^\beta \rangle \\
&\quad \left. - \langle p_i^\alpha \rangle \langle p_j^\gamma \rangle \Delta_2 \langle p_k^\gamma q_k^\beta \rangle - \langle p_i^\alpha \rangle \langle p_k^\gamma \rangle \Delta_2 \langle p_j^\gamma q_k^\beta \rangle \right).
\end{aligned}$$

This term cancels the terms of the  $A_{pp}^1$  proportional to  $\langle p_i^\alpha \rangle$  to give the approximation

$$\begin{aligned}
A_{pp} &\approx \frac{1}{\alpha^2} \sum_{k,\gamma} K(\langle q_j \rangle - \langle q_k \rangle) \left( \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \langle p_k^\gamma \rangle \langle q_j^\beta \rangle + \Delta_2 \langle p_i^\alpha p_k^\gamma \rangle \langle p_j^\gamma \rangle \langle q_j^\beta \rangle \right. \\
&\quad + \Delta_2 \langle p_i^\alpha q_j^\beta \rangle \langle p_j^\gamma \rangle \langle p_k^\gamma \rangle + \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \Delta_2 \langle p_k^\gamma q_j^\beta \rangle \\
&\quad + \Delta_2 \langle p_i^\alpha p_k^\gamma \rangle \Delta_2 \langle p_j^\gamma q_j^\beta \rangle + \Delta_2 \langle p_i^\alpha q_j^\beta \rangle \Delta_2 \langle p_k^\gamma p_j^\gamma \rangle \\
&\quad - \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \langle p_k^\gamma \rangle \langle q_k^\beta \rangle - \Delta_2 \langle p_i^\alpha p_k^\gamma \rangle \langle p_j^\gamma \rangle \langle q_k^\beta \rangle \\
&\quad - \Delta_2 \langle p_i^\alpha q_k^\beta \rangle \langle p_j^\gamma \rangle \langle p_k^\gamma \rangle - \Delta_2 \langle p_i^\alpha p_j^\gamma \rangle \Delta_2 \langle p_k^\gamma q_k^\beta \rangle \\
&\quad \left. - \Delta_2 \langle p_i^\alpha p_k^\gamma \rangle \Delta_2 \langle p_j^\gamma q_k^\beta \rangle - \Delta_2 \langle p_i^\alpha q_k^\beta \rangle \Delta_2 \langle p_k^\gamma p_j^\gamma \rangle \right) + T.
\end{aligned}$$

We end this computation by approximating the dynamics of the mixed correlation.

**B.5.  $\langle pq \rangle$  correlation.** We compute

$$\begin{aligned}
\mathcal{L}(p_i^\alpha q_j^\beta) &= -q_j^\beta \frac{\partial h}{\partial q_i^\alpha} + p_i^\alpha \frac{\partial h}{\partial p_j^\beta} \\
&\quad + \frac{1}{2} \sum_{l,\gamma} p_i^\alpha \sigma_l^\gamma(\mathbf{q}_j) \frac{\partial \sigma_l^\beta(\mathbf{q}_j)}{\partial q_j^\gamma} - \sum_{l,\gamma} p_i^\gamma \sigma_l^\beta(\mathbf{q}_j) \frac{\partial \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha} \\
&\quad - \frac{1}{2} \sum_{l,\gamma,\delta} q_j^\beta p_i^\gamma \frac{\partial^2 \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha \partial q_i^\delta} \sigma_l^\delta(\mathbf{q}_i) + \frac{1}{2} \sum_{l,\gamma,\delta} p_i^\delta q_j^\beta \frac{\partial \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha} \frac{\partial \sigma_l^\delta(\mathbf{q}_i)}{\partial q_i^\gamma}.
\end{aligned}$$

Then, using (4.5) and (B.5) we obtain the time evolution of  $\Delta \langle p_i^\alpha q_j^\beta \rangle$  as

$$\frac{d}{dt} \Delta \langle p_i^\alpha q_j^\beta \rangle = A_{pq} + B_{pq} + C_{pq},$$

where

$$\begin{aligned}
A_{pq} &= - \left\langle q_j^\beta \frac{\partial h}{\partial q_i^\alpha} \right\rangle + \left\langle q_j^\beta \right\rangle \left\langle \frac{\partial h}{\partial q_i^\alpha} \right\rangle + \left\langle p_i^\alpha \frac{\partial h}{\partial p_j^\beta} \right\rangle - \langle p_i^\alpha \rangle \left\langle \frac{\partial h}{\partial p_j^\beta} \right\rangle \\
B_{pq} &= - \sum_{l,\gamma} \left\langle p_i^\gamma \sigma_l^\beta(\mathbf{q}_j) \frac{\partial \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha} \right\rangle \\
C_{pq} &= \frac{1}{2} \sum_{l,\gamma} \left\langle p_i^\alpha \sigma_l^\gamma(\mathbf{q}_j) \frac{\partial \sigma_l^\beta(\mathbf{q}_j)}{\partial q_j^\gamma} \right\rangle - \frac{1}{2} \sum_{l,\gamma} \langle p_i^\alpha \rangle \left\langle \sigma_l^\gamma(\mathbf{q}_j) \frac{\partial \sigma_l^\beta(\mathbf{q}_j)}{\partial q_j^\gamma} \right\rangle \\
&\quad - \frac{1}{2} \sum_{l,\gamma,\delta} \left\langle q_j^\beta p_i^\gamma \frac{\partial^2 \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha \partial q_i^\delta} \sigma_l^\delta(\mathbf{q}_i) \right\rangle + \frac{1}{2} \sum_{l,\gamma,\delta} \left\langle q_j^\beta \right\rangle \left\langle p_i^\gamma \frac{\partial^2 \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha \partial q_i^\delta} \sigma_l^\delta(\mathbf{q}_i) \right\rangle \\
&\quad + \frac{1}{2} \sum_{l,\gamma,\delta} \left\langle p_i^\delta q_j^\beta \frac{\partial \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha} \frac{\partial \sigma_l^\delta(\mathbf{q}_i)}{\partial q_i^\gamma} \right\rangle - \frac{1}{2} \sum_{l,\gamma,\delta} \langle q_j^\beta \rangle \left\langle p_i^\delta \frac{\partial \sigma_l^\gamma(\mathbf{q}_i)}{\partial q_i^\alpha} \frac{\partial \sigma_l^\delta(\mathbf{q}_i)}{\partial q_i^\gamma} \right\rangle.
\end{aligned}$$

We first approximate

$$\begin{aligned}
B_{pq} &\approx \frac{1}{\alpha_l^2} \sum_{l,\gamma} \sigma_l^\beta(\mathbf{q}_j) \sigma_l^\gamma(\mathbf{q}_i) (\langle p_i^\gamma q_i^\alpha \rangle - \langle p_i^\gamma \rangle \delta_l^\alpha) \\
C_{pq} &\approx - \sum_{l,\gamma} \frac{1}{2\alpha_l^2} \sigma_l^\beta(\mathbf{q}_j) \sigma_l^\gamma(\mathbf{q}_j) \Delta_2 \langle p_i^\alpha q_j^\gamma \rangle + \sum_{l,\gamma} \frac{1}{2\alpha_l^2} \Delta_2 \langle q_j^\beta p_i^\gamma \rangle \sigma_l^\gamma(\langle \mathbf{q}_i \rangle) \sigma_l^\alpha(\langle \mathbf{q}_i \rangle).
\end{aligned}$$

For the Hamiltonian term we obtain

$$\begin{aligned}
A_{pq} &\approx \frac{1}{\alpha^2} \sum_{k,\gamma} K(\langle \mathbf{q}_i \rangle - \langle \mathbf{q}_k \rangle) \left( \left\langle q_j^\beta p_i^\gamma p_k^\gamma q_i^\alpha \right\rangle - \left\langle q_j^\beta \right\rangle \langle p_i^\gamma p_k^\gamma q_i^\alpha \rangle \right. \\
&\quad \left. - \left\langle q_j^\beta p_i^\gamma p_k^\gamma q_k^\alpha \right\rangle + \left\langle q_j^\beta \right\rangle \langle p_i^\gamma p_k^\gamma q_k^\alpha \rangle \right) + \sum_{k,\gamma} K(\langle \mathbf{q}_j \rangle - \langle \mathbf{q}_k \rangle) \Delta_2 \langle p_i^\alpha p_k^\beta \rangle \\
&\approx \frac{1}{\alpha^2} \sum_{k,\gamma} K(\langle \mathbf{q}_i \rangle - \langle \mathbf{q}_k \rangle) \left( \Delta_2 \langle p_i^\gamma q_j^\beta \rangle \langle p_k^\gamma \rangle \langle q_i^\alpha \rangle + \Delta_2 \langle p_k^\gamma q_j^\beta \rangle \langle p_i^\gamma \rangle \langle q_i^\alpha \rangle \right. \\
&\quad + \Delta_2 \langle q_i^\alpha q_j^\beta \rangle \langle p_i^\gamma \rangle \langle p_k^\gamma \rangle + \Delta_2 \langle p_i^\gamma q_j^\beta \rangle \Delta_2 \langle p_k^\gamma q_i^\alpha \rangle \\
&\quad + \Delta_2 \langle p_k^\gamma q_j^\beta \rangle \Delta_2 \langle p_i^\gamma q_i^\alpha \rangle + \Delta_2 \langle q_i^\alpha q_j^\beta \rangle \Delta_2 \langle p_k^\gamma p_i^\gamma \rangle \\
&\quad - \Delta_2 \langle p_i^\gamma q_j^\beta \rangle \langle p_k^\gamma \rangle \langle q_k^\alpha \rangle - \Delta_2 \langle p_k^\gamma q_j^\beta \rangle \langle p_i^\gamma \rangle \langle q_k^\alpha \rangle \\
&\quad - \Delta_2 \langle q_j^\beta q_k^\alpha \rangle \langle p_i^\gamma \rangle \langle p_k^\gamma \rangle - \Delta_2 \langle p_i^\gamma q_j^\beta \rangle \Delta_2 \langle p_k^\gamma q_k^\alpha \rangle \\
&\quad \left. - \Delta_2 \langle p_k^\gamma q_j^\beta \rangle \Delta_2 \langle p_i^\gamma q_k^\alpha \rangle - \Delta_2 \langle q_j^\beta q_k^\alpha \rangle \Delta_2 \langle p_k^\gamma p_i^\gamma \rangle \right) \\
&\quad + \sum_{k,\gamma} K(\langle \mathbf{q}_j \rangle - \langle \mathbf{q}_k \rangle) \Delta_2 \langle p_i^\alpha p_k^\beta \rangle.
\end{aligned}$$

**B.6. Higher Order Approximations of the Moment Equations.** One could consider the following higher order approximation of the expected value of the kernel  $K(\mathbf{q}_i - \mathbf{q}_j) = e^{-\|\mathbf{q}_i - \mathbf{q}_j\|^2/(2\alpha^2)}$ , as

$$K(\mathbf{q}_i - \mathbf{q}_j) = e^{-\|\mathbf{q}_i - \mathbf{q}_j\|^2/(2\alpha^2)} \approx \theta(\mathbf{q}_i - \mathbf{q}_j) \left( 1 - f_\alpha \frac{1}{2\alpha^2} \|\mathbf{q}_i - \mathbf{q}_j\|^2 \right), \quad (\text{B.8})$$



where the function  $\theta(\mathbf{x})$  is given by

$$\theta(\mathbf{x}) = \begin{cases} 1 & \text{if } 1 - f_\alpha \frac{\|\mathbf{x}\|^2}{2\alpha^2} \geq 0 \\ 0 & \text{if } 1 - f_\alpha \frac{\|\mathbf{x}\|^2}{2\alpha^2} < 0, \end{cases} \quad (\text{B.9})$$

and the coefficient  $f_\alpha$  is found such that this approximation is the best fit to the Gaussian. In practice, we have  $f_\alpha \approx 0.6$ , but this value depends on  $\alpha$  in general. This cutoff function  $\theta$  is necessary here, otherwise this approximation will not be bounded, leading to large errors in the dynamics.

AA, DH: DEPARTMENT OF MATHEMATICS, IMPERIAL COLLEGE, LONDON SW7 2AZ, UK

SS: DEPARTMENT OF COMPUTER SCIENCE (DIKU), UNIVERSITY OF COPENHAGEN, DK-2100 COPENHAGEN E, DENMARK