Title

Symmetrized importance samplers for stochastic differential equations

Journal

Communications in Applied Mathematics and Computational Science, 13(2)

ISSN

1559-3940

Authors

Leach, Andrew
Lin, Kevin
Morzfeld, Matthias

Publication Date

2018

DOI

10.2140/camcos.2018.13.215

Peer reviewed

Symmetrized importance samplers for stochastic differential equations

Andrew Leach*, Kevin K. Lin*,†, and Matthias Morzfeld*,†,‡

October 1, 2018

Abstract

We study a class of importance sampling methods for stochastic differential equations (SDEs). A small-noise analysis is performed, and the results suggest that a simple symmetrization procedure can significantly improve the performance of our importance sampling schemes when the noise is not too large. We demonstrate that this is indeed the case for a number of linear and nonlinear examples. Potential applications, e.g., data assimilation, are discussed.

1 Introduction

Consider a stochastic differential equation (SDE)

$$dX_t = f(X_t)\,dt + \sigma\,dB_t\,,\;\; X_t \in \mathbb{R}^D, \tag{1.1}$$

where $f : \mathbb{R}^D \to \mathbb{R}^D$ and B_t is D-dimensional Brownian motion. Suppose we make noisy observations of the system at times $t = T, 2T, 3T, \cdots, JT$ ($T > 0$, fixed), obtaining a sequence of measurements $Y_j = m(X_{jT}) + \eta_j$, where $m : \mathbb{R}^D \to \mathbb{R}^d$ ($d \leqslant D$) is the quantity being measured (the "observable"), η_j are independent identically-distributed (IID) random variables modeling measurement errors and $j = 1 \ldots J$. What is the conditional distribution of X_t for $t \in [0, JT]$ given $Y_1, Y_2, \cdots, Y_J$? This problem of "nonlinear filtering" or "data assimilation" arises in many applications; see, e.g., [5, 7, 8, 27]. A variety of algorithms have been developed to address it, but efficient data assimilation, especially in high-dimensional non-gaussian problems, remains a challenge [25].

This paper concerns an approach to data assimilation known as "particle filtering" (see, e.g., [8] for more details) based on *sampling* the conditional distributions. We present an asymptotic analysis of certain sampling algorithms designed to improve the efficiency of particle filtering, and, based on this analysis, we propose a general way to improve their performance. The analysis relies on taking a small-noise limit, but the algorithms do not require a small

*Program in Applied Mathematics, University of Arizona, 617 N. Santa Rita Ave., Tucson, AZ 85721, USA.
†Department of Mathematics, University of Arizona, 617 N. Santa Rita Ave., Tucson, AZ 85721, USA.
‡Corresponding author. E-mail: mmo@math.arizona.edu

noise to operate (but may not be as efficient when the noise is not small). We focus on *one* step of the filtering problem, i.e., we set $J = 1$ in the above, as this is sufficient to capture the computational difficulty we wish to address. For simplicity, we assume $\eta \sim \mathcal{N}(0, rI)$, where $r > 0$ is a scalar and I is the $d \times d$ identity matrix; we also assume $\sigma > 0$ is a scalar. These assumptions can be relaxed if needed.

To take one step of particle filtering, one begins by discretizing Eq. (1.1) using, e.g., the Euler scheme, to obtain

$$X_{n+1} = X_n + \Delta t\ f(X_n) + \sqrt{\Delta t}\ \sigma \cdot \xi_n\ , \quad X_0 = x_0 \in \mathbb{R}^D,\ n = 0, \cdots, N-1, \tag{1.2}$$

where $N\Delta t = T$, the ξ_n are IID standard normal random variables. A straightforward application of Bayes's Theorem tells us that the conditional distribution of interest satisfies

$$p(x_1, \cdots, x_N | y) \propto \exp\left(\frac{1}{2\sigma^2 \Delta t} \sum_{n=0}^{N-1} \left\| x_{n+1} - x_n - f(x_n)\Delta t \right\|^2 + \frac{\|m(x_N) - y\|^2}{2r} \right). \tag{1.3}$$

One then tries to design a Monte Carlo algorithm to generate discrete-time sample paths $(X_1, \cdots, X_N)$ from Eq. (1.3), conditioned on the observation y. We refer to the distribution in Eq. (1.3) as the *target distribution*. They are the discrete-time analogs of the conditional distributions introduced above, with $J = 1$ observation.

Without the last term in the exponent in Eq. (1.3), the target distribution is just the distribution of the discretized SDE, and one can generate sample paths by carrying out the recursion in Eq. (1.2). When the last term is included, however, it is generally not feasible to sample directly from the target distribution. A solution to this problem is *importance sampling:* instead of drawing samples from the target distribution, we draw sample paths $(Z_1, \cdots, Z_N)$ from an approximation q, usually called the "proposal distribution". Any statistics we compute based on sample paths from q will be biased. We compensate for this bias by associating a weight $W^{(k)} > 0$ to the kth sample path $(Z_1^{(k)}, \cdots, Z_N^{(k)})$, with $\sum_k W^{(k)} = 1$, so that the *weighted* sample paths $(Z^{(k)}, W^{(k)})$ again have the correct statistics (in a sense we make precise later).

Weare and Vanden-Eijnden [28, 29] proposed an algorithm for sampling distributions like Eq. (1.3). They showed that their algorithm is efficient in the sense that in the limit of small dynamical and observation noise, the relative variance of the weights vanishes (see [29] for precise definitions and statements). The basic idea of the sampler is to look for the most likely sample path of the target distribution (1.3) and use this information to modify the dynamics so that samples from the proposal remain close to the target distribution. In this paper, by a combination of formal asymptotic analysis and numerical examples, we show that a symmetrization procedure proposed in [17] can be applied to SDEs to improve the efficiency of importance samplers. The symmetrization and "small noise analysis" has also been discussed in the context of implicit sampling [6, 23], see [17].

While our primary motivation here is data assimilation for SDEs, our symmetrization procedure may be effective for sequential Monte Carlo sampling of more general types of systems. As well, the class of importance sampling algorithms studied here are closely related to algorithms proposed in [9–13] and in [28] for sampling "rare events" in SDEs, though there are some significant differences between the two applications. We plan to explore some of these connections in future work.

Paper organization. The remainder of this paper is organized as follows. We state our main results in Section 2. Section 3 briefly reviews the linear map method and its symmetrization, as well as the small noise theory (see [17]). We explain a new sampling method, the dynamic linear map, in Section 4. We study its efficiency in the small noise regime and show how to use symmetrization to improve its efficiency in small noise problems. Several numerical examples are provided in Section 5 that illustrate our asymptotic results as well as the efficiency of our dynamic approach in multimodal problems. The continuous time limit of the dynamic linear map is discussed in Section 6 and we present conclusions in Section 7.

2 Problem statement and summary of results

We now formulate the problem more precisely and summarize our key findings. We consider a discretized SDE in the small noise regime

$$X_{n+1} = X_n + \Delta t\ \widetilde{f}(X_n, \Delta t) + \sqrt{\Delta t}\ \sqrt{\varepsilon}\ \sigma \cdot \xi_n\ , \quad X_0 = x_0 \in \mathbb{R}^D, \tag{2.1}$$

where $\widetilde{f}(x, \Delta t) = f(x) + O(\Delta t)$ corresponds to a numerical discretization of $\dot{x} = f(x)$ (for most of this paper, we assume the Euler discretization $\widetilde{f}(x, \Delta t) = f(x)$), and $\varepsilon \ll 1$ is the "small noise parameter". Throughout this paper we assume that the D-dimensional vector field $\widetilde{f}$ is smooth, and that the process starts at a given initial position x_0 and proceeds for N time steps of size Δt each. The transitions are made with independent gaussian samples $\xi_n \sim \mathcal{N}(0, I)$. We denote the path as $x_{1:N}$, a sequence of positions $x_1, \ldots x_N$, and its likelihood in the process with the path distribution $\rho(x_{1:N}|x_0)$.

The observation of the state at time $N\Delta t$ gives rise to the *likelihood*

$$\theta(x_N) := \exp\left(-\tfrac{1}{\varepsilon} g(x_N)\right), \tag{2.2}$$

where g is assumed to be a smooth, nonnegative function. For example, for observations $y = m(x_N) + \eta$, $\eta \sim \mathcal{N}(0, \varepsilon r I)$, we have $g(x_N) = (2r)^{-1} ||m(x_N) - y||^2$. Hereafter we will sometimes refer to g as the "log-likelihood," in a slight abuse of standard terminology. By Bayes's Theorem, the target distribution then has the form

$$p(x_{1:N}|x_0) \propto \rho(x_{1:N}|x_0) \cdot \theta(x_N)\ . \tag{2.3}$$

Importance sampling methods generate samples using a proposal distribution q, and attach weights

$$W^{(k)} = w(X^{(k)}_{1:N}|x_0) = p(X^{(k)}_{1:N}|x_0)/q(X^{(k)}_{1:N}|x_0) \tag{2.4}$$

to each sample, so that the weighted samples can be used to compute unbiased statistical estimates with respect to the target distribution. To measure the efficiency of the sampling methods, we evaluate the relative variance of the weights

$$Q := \frac{\mathrm{Var}[W]}{\mathrm{E}[W]^2}. \tag{2.5}$$

Here the expected values are computed with respect to the proposal distribution q. This relative variance Q is connected to a standard heuristic called the "effective sample size," defined by

$$N_{\text{eff}} := \frac{N_e}{1+Q}, \tag{2.6}$$

where N_e is the number of weighted samples (see, e.g., [4, 8, 21]). The effective sample size is meant to measure the size of an unweighted ensemble that is equivalent to the weighted ensemble of size N_e. All else being equal, the smaller the Q, the more efficient the importance sampling algorithm, and if all the samples were independent, we would have $Q = 0$ and $N_{\text{eff}} = N_e$. The quantity Q is convenient because it is not tied to any specific observable; recent work (see [1]) has also given it a more precise meaning. Other quantities that can assess effective sample sizes are discussed in [22]. We note that in practice, p and q are only known up to a constant. The algorithms we describe do not require knowing the normalization constants. Likewise, Q is invariant under rescaling of p or q by a constant.

We study two types of importance sampling methods in this paper. The first method, called the "linear map" (LM), uses a gaussian proposal distribution centered at the most likely path. The second method, called dynamic linear map (DLM), re-applies the linear map after each time step between $t = 0$ and $t = N\Delta t$ given the previous moves. Note that the linear map can be viewed as a version of implicit sampling [6, 23] applied to the path distribution of an SDE. The dynamic linear map applies this implicit sampling step repeatedly to transition densities and is also closely linked to the continuous time control method of Weare and Vanden-Eijnden [28, 29] (see also Section 6). For each method, we perform a symmetrization and exploit symmetries of the proposal distributions to increase sampling efficiency. Symmetrization was previously studied for the LM in a more general context in [17]. Here we adapt this procedure to problems involving SDE and to the dynamic linear map. Following the approach taken in [17], we show that under suitable assumptions (see Section 4), the relative variances of the various methods are as follows:

Method	$Q(\varepsilon)$ scaling
Linear Map (LM)	$O(\varepsilon)$
Symmetrized LM	$O(\varepsilon^2)$
Dynamic LM (DLM)	$O(\varepsilon)$
Symmetrized DLM	$O(\varepsilon^2)$

We also present examples showing that the leading coefficient of the DLM can be smaller than that of LM, suggesting that DLM may be more effective in some situations (see Section 5). We discuss the continuous time limit of LM and DLM for scalar SDE, and calculate the leading coefficient of $Q(\varepsilon)$ in an asymptotic expansion in ε. In doing so, we show that, under additional assumptions, the sampling method discussed in [28] is recovered in the $\Delta t \to 0$ limit of the DLM (see Section 6).

Notes.

(i) The ε-expansions we will consider are formally justified as the relevant quantities, e.g., relative weight variance, are gaussian integrals.

(ii) The insertion of the small noise parameter ε into the problem is mainly to enable asymptotic analysis. In specific problems, there is not always an identifiable small parameter, and in any case our methods do not require a small parameter to operate.

3 Background

We simplify notation and write $x := x_{1:N}$, and $F(x) := F(x_{1:N}|x_0)$, and consider the small noise target distribution defined in (2.3) which can be written as $p(x) \propto \exp(-F(x)/\varepsilon)$, where

$$F(x) = \frac{\Delta t}{2\sigma^2} \sum_{n=0}^{N-1} \left\| \frac{x_{n+1} - x_n}{\Delta t} - \widetilde{f}(x_n, \Delta t) \right\|^2 + g(x_N) \,, \tag{3.1}$$

for g, a scalar function as in (2.2). If we assume that F has a unique, nondegenerate minimum, and let

$$\varphi = \operatorname*{argmin}_{x \in \mathbb{R}^{D \cdot N}} F(x), \tag{3.2}$$

i.e., φ is the optimal path with prescribed initial condition x_0, we can employ Laplace asymptotics to expand the target distribution around φ. (See, e.g., [24] for a general formulation of Laplace asymptotics.) After a change of variables

$$z = \varepsilon^{-1/2} \cdot (x - \varphi) \tag{3.3}$$

the expansion is

$$F(z) = F(\varphi) + z^T H z/2 + \varepsilon^{1/2} C_3(z) + \varepsilon C_4(z) + O(\varepsilon^{3/2}), \tag{3.4}$$

where H is the Hessian evaluated at φ, C_k are the higher order terms in the Taylor series. Here and below, we use the shorthand $F(z) := F(\varphi + \varepsilon^{1/2} z)$, and similarly write $w(z)$ for $w(\varphi + \varepsilon^{1/2} z)$ etc. Note that while we will continue to refer to $z := \{z_1, \ldots, z_n\}$ as a "path" after the change of coordinates, $x = \varphi + \sqrt{\varepsilon} z$ is the actual solution of Eq. (2.1).

The small noise analysis of LM, and other methods to follow will make frequent use of this expansion, as well as the "variance lemma" (see [17]).

Lemma 1. (Variance Lemma) *For a function $u(z, \varepsilon)$ that can be expanded in ε at least to the terms*

$$u(z) = 1 + \varepsilon^r u_1(z) + \varepsilon^{2r} u_2(z) + O(\varepsilon^{3r}) \tag{3.5}$$

the relative variance of u with respect to a probability density q is

$$Q = \varepsilon^{2r} \operatorname{Var}_q [u_1(z)] + O(\varepsilon^{3r}) \tag{3.6}$$

3.1 Linear map

The proposal distribution of the linear map (LM) sampling method, summarized in Algorithm 1, is gaussian and proportional to

$$q(z) \propto \exp\left(-z^T H z/2\right). \tag{3.7}$$

The weights are the ratio of target and proposal distribution, and can be expanded as

$$w(z) = 1 - \varepsilon^{1/2} C_3(z) + O(\varepsilon). \tag{3.8}$$

Using the variance lemma we thus find that

$$Q = \varepsilon \operatorname{Var}_q [C_3(z)] + O(\varepsilon^{3/2}), \tag{3.9}$$

i.e., the relative variance of the weights is linear in ε (see [17] for more details).

Algorithm 1: Linear Map

Calculate φ and H starting from x_0;
for $m = 1$ *to* M **do**
  Sample $X \sim \mathcal{N}(\varphi, \varepsilon H^{-1})$;
  Calculate $W = p(X)/q(X)$;
Return M weighted samples X, W;

3.2 Symmetrized linear map

It is shown in [17] that the linear map can be "symmetrized" to improve the scaling of Q from linear to quadratic in ε. This stems from the observation that the leading order term in the weight is an odd function with respect to the random variable z, whose probability distribution function is even. The symmetrized sampler uses a proposal distribution which reweights equally likely samples from the gaussian distribution of the linear map such that the resulting weights have even symmetry. The odd leading order terms in the weight expansions then cancel, which results in a quadratic scaling of Q in ε.

Specifically, the symmetrized linear map draws a sample z from the proposal distribution q. It returns z with probability $w^+/(w^- + w^+)$, and $-z$ with probability $w^-/(w^- + w^+)$, where

$$w^+ = \frac{p(-z)}{q(z)} \quad \text{and} \quad w^- = \frac{p(z)}{q(z)}. \tag{3.10}$$

Samples generated in this way have a non-symmetric distribution, but even weights:

$$q_s(z) = q(z)\frac{2w^+}{w^- + w^+}, \quad w_s(z) = \frac{w^- + w^+}{2}. \tag{3.11}$$

The Taylor expansion of the symmetrized weight is

$$w_s(z) = 1 + \varepsilon\left(\tfrac{1}{2}C_3(z)^2 - C_4(z)\right) + O(\varepsilon^2), \tag{3.12}$$

which, together with the variance lemma shows that

$$Q_s = \varepsilon^2 \operatorname{Var}_q\left[\tfrac{1}{2}C_3(z)^2 - C_4(z)\right] + O(\varepsilon^4). \tag{3.13}$$

The symmetrization therefore improves the linear scaling of Q in ε of LM, to a quadratic scaling of Q for SLM (see [17] for more details).

4 Dynamic linear map and its symmetrization

4.1 A multimodal example

The linear map can be efficient when the hypotheses underlying its derivation are satisfied, i.e., when the pathspace distribution is unimodal and a gaussian approximation is appropriate.

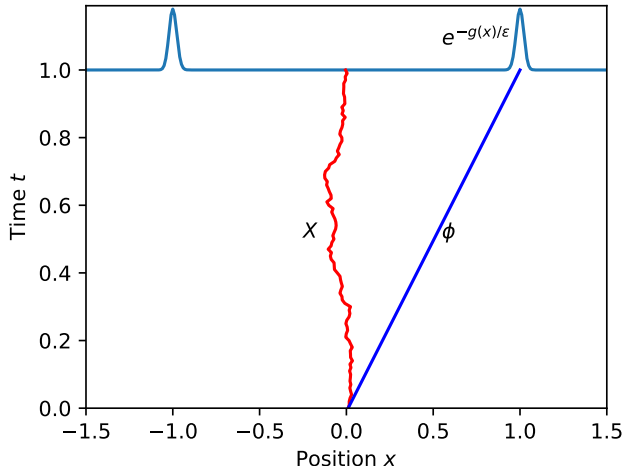


Figure 1: Brownian motion with bimodal likelihood. Here, the initial condition is $X_0 = 0.01$, and we use $\varepsilon = 0.1$. Shown are a sample path X and the optimal path φ starting from X_0.

However, when there are multiple modes, LM can become inefficient. To see how this might happen, consider the simple random walk

$$X_{n+1} = X_n + \sqrt{\Delta t}\ \sqrt{\varepsilon}\ \xi_n \tag{4.1}$$

i.e., $X_n = X_0 + \sqrt{\Delta t}\ \sqrt{\varepsilon}\ W_n$ where W_n is standard Wiener process. Suppose we have a bimodal likelihood function $e^{-g(x)/\varepsilon}$ whose graph is as shown in Figure 1; this type of situation can arise when multiple states can give the same measurement, so that observations may have ambiguous interpretation. In this case, the high probability paths will be those that reach the vicinity of $x = \pm 1$ at $t = 1$; effectively, the high probability paths are sample paths of Brownian motion, conditioned to be near $x = \pm 1$ at $t = 1$. The probability of this occurring by chance is exponentially small as $\varepsilon \to 0$, and direct sampling is unlikely to ever produce such a path.

A straightforward calculation shows that the optimal path φ approaches a straight line in the xt-plane as $\varepsilon \to 0$, going to the right bump if $X_0 > 0$, to the left if $X_0 < 0$ (and undefined if $X_0 = 0$). With a bimodal likelihood function, the target distribution $p(x)$ is bimodal as well. If the initial condition is sufficiently to the right of $x = 0$, one of the two modes will dominate, and LM can be expected to be effective. As X_0 moves closer to $x = 0$, however, the other mode will begin to make a greater contribution; at $X_0 = 0$, the two modes carry exactly the same weight. But LM will *always* pick the mode on the right when $X_0 > 0$, no matter how close X_0 is to $x = 0$. So LM will produce essentially no sample paths going to the left, leading to a large weight variance. See Section 5 for detailed numerical results.

This is a well-known problem with importance sampling algorithms. Similar issues arise in rare event simulation, and a standard solution is to dynamically recompute the optimal path. See, e.g., the discussion of Siegmund's algorithm in [2]. In our context, this leads to an algorithm we call the dynamic linear map, which is similar to the algorithms proposed in [13, 28]. We will also discuss symmetrization in this context.

4.2 Dynamic linear map

Roughly speaking, the dynamic linear map (DLM) consists of computing the optimal path φ starting from the current state X_n, taking *one* step (so that $X_{n+1} = \varphi_{n+1}$), then repeating. See Algorithm 2 for details. The DLM thus requires redoing LM *at every step*, and is therefore more expensive.[1] However, it can avoid some of the issues arising from multi-modal target distributions. One can see this heuristically in the above example (Section 4.1): suppose we start with X_0 slightly to the right of $x = 0$, so that the optimal path φ goes to the right bump. After a few steps, we may end up in a state X_n closer to the left bump. At this point, the DLM would start steering the sample path towards the left bump. Unlike LM, repeated sampling using DLM would yield sample paths that end at both the left and the right bumps (see Section 5.1).

To make use of DLM, we need an expression for the associated weights. This, in turn, requires an expression for the proposal distribution q associated with DLM, which one can derive by first noting that in general, transition densities are marginals of the pathspace distribution:

$$\rho(x_{n+1}|x_n) = \int \rho(x_{n+1:N}|x_n)\, dx_{n+2:N}.$$

(Here we abuse notation slightly and use p and q to denote both pathspace distributions as well as their marginals.) The DLM transition density arises from making a gaussian approximation of the target distribution at each step, then taking its marginal. This leads to

$$\begin{aligned} q(x_{n+1}|x_n) &= \int q(x_{n+1:N}|x_n)\, dx_{n+2:N} \\ &\propto \exp\left(-(x-\varphi)^T_{n+1} \Sigma^{-1}_{n+1} (x-\varphi)_{n+1}/(2\Delta t)\right). \end{aligned} \tag{4.2}$$

Here φ is the optimal path from x_n to x_N and we omit its dependence on x_n for readability of the equations; we also remind the reader that $x = x_n, \ldots, x_N$ is a path. We denote the Hessian of $F(x)$ evaluated at the optimal path φ by H. We view a path from x_n to x_{n+k} as a point in $\mathbb{R}^{kD}$, arranged in k blocks of D entries. Accordingly, the matrix H can be viewed as an element of $\mathbb{R}^{(N-n)D\times(N-n)D}$ and can be subdivided into $(N-n)\times(N-n)$ blocks of dimension $D \times D$ each. The matrix Σ_{n+1} in Eq. (4.2) is $(H^{-1})_{1,1}/\Delta t$, the first block of the inverse of the Hessian H (after rescaling).

In Algorithm 2, going from step n to $n+1$ requires optimizing over the $(N-n)D$ remaining variables in the path. This is done independently at every step and for every sample path. The weights for the proposal distribution of DLM can be calculated as described in Algorithm 2, or as the product of the incremental weights

$$w = \prod_{n=0}^{N-1} w_n, \quad w_n \propto \frac{p(x_{n+1}|x_n)}{q(x_{n+1}|x_n)}. \tag{4.3}$$

Relation to Hamilton-Jacobi equation and regularity of "value functions". In the definitions above, it is assumed that $q(x_{n+1}|x_n)$ is well-defined for all (x_n, x_{n+1}). This is actually not always the

[1] Suppose each cost function evaluation requires CPU time $\propto N$, the number of steps, and each optimization requires k function evaluations. Then all else being equal, LM has running time $O(kN)$ and DLM $O(kN^2)$.

Algorithm 2: Dynamic Linear Map

for $m = 1$ *to* M **do**
  for $n = 0$ *to* $N-1$ **do**
    Calculate φ and H starting from X_n;
    Calculate $\Sigma_{n+1} = (H^{-1})_{1,1}/\Delta t$;
    Sample $X_{n+1} \sim \mathcal{N}(\varphi_{n+1},\ \Delta t\ \varepsilon\ \Sigma_{n+1})$;
    Calculate $W_n = p(X_{n+1}|X_n)/q(X_{n+1}|X_n)$;
Calculate $W = W_{N-1} \cdot \ldots \cdot W_0$;
Return M weighted samples X, W;

case. To see this, consider again the example from Section 4.1. If $x_n = 0$ at some n, there are two optimal paths pointing in opposite directions. At this point, because there is not a single optimal path, $q(x_{n+1}|x_n)$ is undefined. This behavior is actually rather common, and not at all confined to the Brownian motion example. It is closely connected with regularity of solutions of a partial differential equation of Hamilton-Jacobi (HJ) type. As we do not make use of the theory of HJ equations in this paper, we do not go into details here. Instead, we provide a brief summary below, and refer interested readers to, e.g., [29] or [9, 10, 12, 13], for more information.

In the DLM method, the optimal path minimizes a version of the function F in Eq. (3.1), but starting with state x_n at time n rather than always at time 0. In the limit as $\Delta t \to 0$, the *value function* $u(x,t)$ achieved with initial condition $x_n = x$ at step $n\Delta t = t$ solves a HJ equation of the form $\partial_t u = H(x, Du)$, with Hamiltonian $H(x,p) = \frac{\sigma^2}{2}|p|^2 + p \cdot f(x)$; this is the Legendre transformation of the Freidlin-Wentzell Lagrangian $L(x,v) = \frac{1}{2\sigma^2}|v - f(x)|^2$ [14]. For the HJ equation to be well-posed, one prescribes the *final condition* that $u(x,T) = g(x)$, where g is the likelihood in Eq. (2.2) and $T > 0$. The HJ equation is then solved backwards in time. The time derivative $\dot{\varphi}$ of the optimal path starting at position x and time t is given by the gradient of $u(x,t)$ where it is differentiable. At locations (x,t) where there are multiple optimal paths, the value function $u(x,t)$ is generally continuous but not differentiable. At such *singular points* x, $q(x_{n+1}, x)$ has jump disconinuities (as x varies) and is therefore undefined.

Though very much relevant to the efficacy of the type of methods discussed in this paper, the analysis of singularities of HJ equations can be highly nontrivial. As our main goal is to assess whether some version of the symmetrization procedure proposed in [17] can be extended to SDEs, we have opted to focus on the simplest possible setting, leaving more general analysis to future work. *For the remainder of the paper, we make the following* **standing assumption:**

$q(x_{n+1}|x_n)$ is defined everywhere, and is as smooth as needed.

The analytical results described below should therefore be interpreted as a *best-case scenario*. We also note that while the numerical algorithm is unlikely to produce an x_n *exactly* in the set of singular points in actual practice, the presence of singularities does mean that the performance of the algorithm may be worse than predicted by our analysis. We have therefore designed our numerical examples to test the extent to which the algorithms behave as predicted even when $q(x_{n+1}|x_n)$ is not differentiable everywhere.

4.3 Small-noise analysis

To find the scaling of the relative variance of the weights of DLM with the small noise parameter ε, we apply the same change of variables as in Eq. (3.3) to each transition density and expand the incremental weights w_n as

$$w_n \;=\; w(z_{n+1}|z_n) \;=\; 1+\varepsilon^{1/2}\cdot w_{1,n}(z_{n+1}|z_n)+\varepsilon\cdot w_{2,n}(z_{n+1}|z_n)+O(\varepsilon^{3/2}), \tag{4.4}$$

where

$$w_{1,n}(z_{n+1}|z_n)=\frac{\int C_3(z)\exp\left(-z^THz/2\right)\,dz_{n+2:N}}{\int \exp\left(-z^THz/2\right)\,dz_{n+2:N}} \tag{4.5}$$

$$\begin{aligned} w_{2,n}(z_{n+1}|z_n)=&\frac{\int (C_3(z)^2/2-C_4(z))\exp\left(-z^THz/2\right)\,dz_{n+2:N}}{\int \exp\left(-z^THz/2\right)\,dz_{n+2:N}}\\ &-\int (C_3(z)^2/2-C_4(z))\exp\left(-z^THz/2\right)\,dz_{n+1:N}, \end{aligned} \tag{4.6}$$

noting that Eq. (4.4) relies strongly on our standing assumption that $q(x_{n+1}|x_n)$ is differentiable. Since the weight of a sample is the product of the incremental weights, we have

$$w(z)=1+\varepsilon^{1/2}\cdot w_1+\varepsilon\cdot w_2+O(\varepsilon^{3/2}),$$

where

$$w_1=\sum_{n=0}^{N-1}w_{1,n},\quad w_2=\sum_{n=0}^{N-1}w_{2,n}+\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}w_{1,n}\cdot w_{1,m}. \tag{4.7}$$

The scaling of Q in ε now follows from the variance lemma:

$$Q^\varepsilon=\varepsilon\cdot \mathrm{Var}_q\left[w_1\right]+O(\varepsilon^2). \tag{4.8}$$

Thus, the relative variance of DLM scales linearly in ε, the same asymptotic scaling as LM. However, we will show in numerical examples below that the dynamic approach can be more effective in practice than LM, especially when the target distribution has multiple modes.

4.4 Symmetrization

The leading order term in the weight for DLM has an odd symmetry, just like the LM, and a symmetrization procedure can be applied to DLM to improve the scaling of Q in ε. The reason is that, at each time step, X_{n+1} is generated by a composition of the previous state X_n and a new gaussian sample ξ_n. While this procedure leads to a proposal distribution that is not necessarily even, the paths are constructed incrementally from gaussian samples which are even.

Algorithm 3: Symmetrization

for $m = 1$ *to* M **do**

Sample $\xi \sim \mathcal{N}(0, I)$;

Calculate $X^+ = h(\varepsilon^{-1/2}\xi)$ and $X^- = h(-\varepsilon^{-1/2}\xi)$;

Calculate $W^+ = p(X^+)/q(X^+)$ and $W^- = p(X^-)/q(X^-)$;

Sample $X = X^+$ with prob. $\frac{W^+}{W^+ + W^-}$ and $X = X^-$ with prob. $\frac{W^-}{W^+ + W^-}$;

Calculate $W = \frac{W^+ + W^-}{2}$;

Return M weighted samples X, W;

More specifically, the recursive composition forms a map h from the $N \cdot D$ dimensional gaussian to the path $X = h(\varepsilon^{1/2}\xi)$, and for every sampled path $X^+ = h(\varepsilon^{1/2}\xi)$, there is a path $X^- = h(-\varepsilon^{1/2}\xi)$ which is equally likely. Following the algorithm described in Algorithm 3, we sample X^+ with probability $W^+/(W^+ + W^-)$, and X^- with probability $W^-/(W^+ + W^-)$, the resulting proposal is a "symmetrized" distribution with even weights (see Eq. (3.11)).

The symmetrized weights can be written in terms of the map as

$$w_s(h(\varepsilon^{1/2}\xi)) = \frac{w(h(\varepsilon^{1/2}\xi)) + w(h(-\varepsilon^{1/2}\xi))}{2}. \tag{4.9}$$

Recall the expansion of the weights in (4.4), and note that

$$z = \varepsilon^{-1/2}(h(\varepsilon^{1/2}\xi) - h(0)),$$

since the most likely path φ can be written in terms of the map as $\varphi = h(0)$.

If φ is unique (at each time step), h can be expanded around the most likely path as

$$h(\varepsilon^{1/2}\xi) = \varphi + \varepsilon^{1/2}(Dh)(0) \cdot \xi + O(\varepsilon), \tag{4.10}$$

$$h(-\varepsilon^{1/2}\xi) = \varphi - \varepsilon^{1/2}(Dh)(0) \cdot \xi + O(\varepsilon). \tag{4.11}$$

We thus have that

$$w(h(\varepsilon^{1/2}\xi)) = 1 + \varepsilon^{1/2} w_1(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) + \varepsilon w_2(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) + O(\varepsilon^{3/2}) \tag{4.12}$$

$$w(h(-\varepsilon^{1/2}\xi)) = 1 - \varepsilon^{1/2} w_1(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) + \varepsilon w_2(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) + O(\varepsilon^{3/2}) \tag{4.13}$$

which results in the cancellation of the leading order term in ε of the symmetrized weight

$$w_s(h(\varepsilon^{1/2}\xi)) = 1 + \varepsilon w_2(\varepsilon^{1/2}(Dh)(0) \cdot \xi, \varphi) + O(\varepsilon^{3/2}) \tag{4.14}$$

Applying the variance lemma completes the proof for the quadratic scaling of Q_s in ε

$$Q_s = \varepsilon^2 \cdot \mathrm{Var}_{q_s}[w_2] + O(\varepsilon^4). \tag{4.15}$$

5 Numerical examples

We now examine a number of concrete examples, both to illustrate the scaling of the proposed algorithms and to test their limitations. The source code for all examples in this section can be found at `https://github.com/AndrewLeach/SDE_Importance_Sampling` .

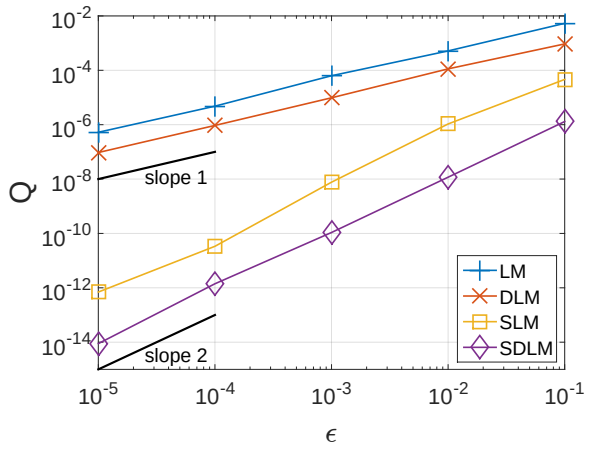


Figure 2: Brownian motion with asymmetric unimodal likelihood. The scaling of Q in ε for LM, SLM, DLM and SDLM are plotted.

5.1 Examples with linear SDE

We begin with the Brownian motion example from Section 4.1:

$$X_{n+1} = X_n + \sqrt{\Delta t}\ \sqrt{\varepsilon}\ \xi_n\ , \tag{5.1}$$

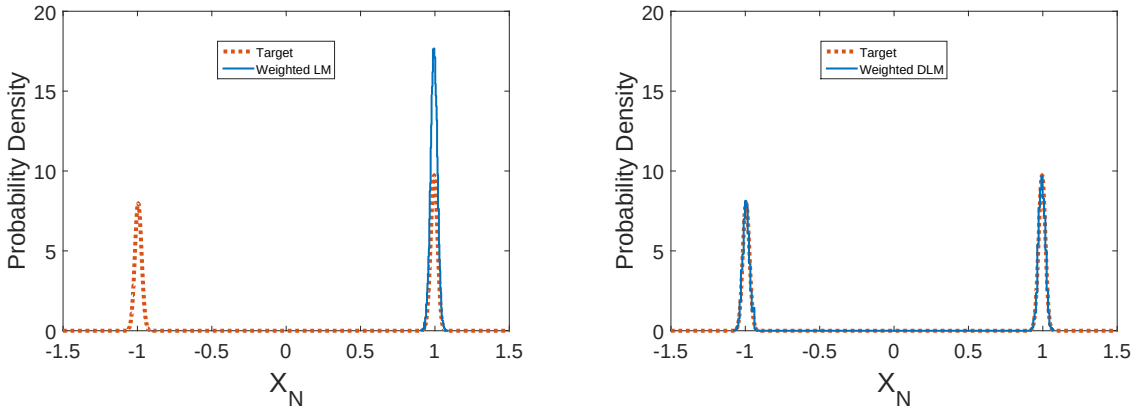
with initial condition $X_0 = x_0$ and with likelihood $\theta = e^{-g(X_N)/\varepsilon}$ for two different choices for g. We first consider the case of a unimodal target distribution for which the assumptions made during the small noise analysis are satisfied. We then violate the assumption of a unique optimal path to indicate limitations of DLM and our small noise analysis. For the examples below, the time step is $\Delta t = 10^{-2}$. The observation is collected at step $N = 100$ (i.e., $T = 1$). Computing the optimal paths is straightforward to do analytically and we use the analytic formulas in our implementation of the various samplers.

Brownian motion with unimodal likelihood. We first consider a likelihood defined by

$$g(x) = \frac{x^4}{24} + \frac{x^3}{6} + \frac{x^2}{2}.$$

The likelihood is asymmetric in x and leads to a non-gaussian and unimodal target distribution. In this example, the assumptions made in our small noise analysis are satisfied.

We apply LM, SLM, DLM, and SDLM to sample the target distribution over a wide range of ε, and compute the relative variance Q for each of these methods. For each ε and method (LM, SLM, DLM and SDLM), we draw 1200 samples . The results are shown in Figure 2. As can be seen, the results show the predicted scalings for Q for a wide range of ε for all four methods: both LM and DLM are $O(\varepsilon)$, while SLM and SDLM are both $O(\varepsilon^2)$. Perhaps this is no surprise, as all assumptions that lead to the small noise theory are valid in this example. We also see that the dynamic methods (DLM and SDLM) have smaller relative variance Q at each value of ε, though they also cost more per sample.



(a) LM (b) DLM

Figure 3: Final-time distribution for Brownian motion with bimodal likelihood. In (a), we plot the marginal distribution $p(x_N|x_0)$ estimated by weighted histograms of 12000 samples generated using LM. Also shown is the target distribution. In (b), we plot the same information for DLM.

Brownian motion with bimodal likelihood. Next, we examine

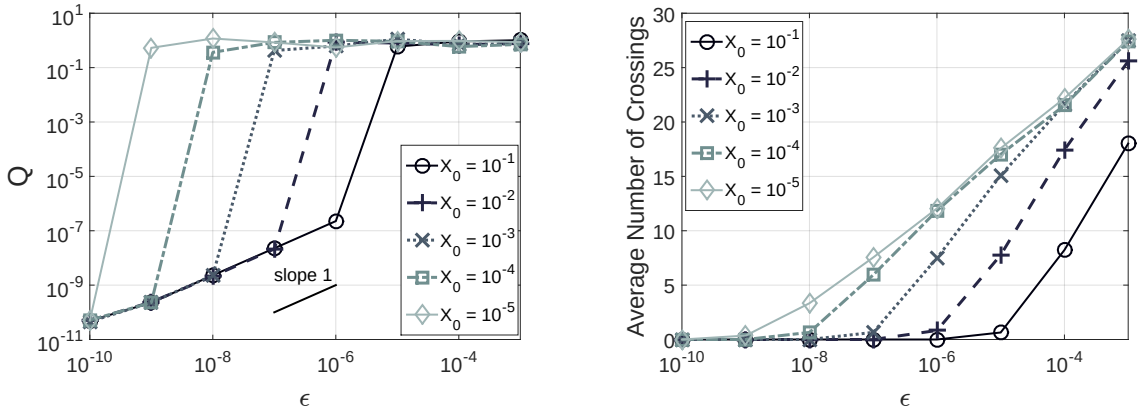
$$g(x) = 100 \cdot \left(\frac{x^4}{4} - \frac{x^2}{2} \right).$$

As explained in Section 4.1, this leads to a bimodal target distribution. We fix $\varepsilon = 10^{-1}$, and leave all other parameters as above. We apply LM and DLM to compute the final-time distribution $p(X_N|X_0)$, using 1.2×10^4 (weighted) samples. The results are shown in Figure 3, along with the target distribution $\propto e^{-(g(x)+x^2/2)/\varepsilon}$.

As expected, LM essentially ignores one of the two modes, while DLM captures both modes. As explained before, even though both samplers should reproduce the target distribution in the large-sample-size limit, in practice LM produces almost no sample paths that go to the left bump. In contrast, DLM readily generates sample paths ending at both bumps, leading to a more effective sampling of the target distribution. We have experimented with increasing the sample size for LM, but even the largest sample sizes we consider did not lead to weighted samples that represent both modes.

Finally, note that empirical estimates of Q are insufficient to detect this problem: even though the true value of Q for LM should be quite large in this case, empirical estimates of Q for LM are actually quite small because none of the sample paths go to the left bump. Indeed, for Figure 3, the empirical Q for LM is $\sim 3 \times 10^{-3}$, while that of DLM is ~ 1. The example thus shows that for non-gaussian and possibly multimodal distributions, DLM can be more reliable despite the same scaling of Q.

Overdamped Langevin equation with bimodal likelihood. The scaling arguments for DLM and its symmetrized version rely on the assumption that the most likely path φ is unique at every time step. We now consider an example for the DLM in which we deliberately violate this



(a) (b)

Figure 4: DLM applied to the overdamped Langevin equation with bimodal likelihood. Panel (a) shows the scaling of Q vs. ε for x_0 approaching $x = 0$. In (b), we plot the average number of $x = 0$ crossings against ε.

assumption. The model is

$$X_{n+1} = X_n - \Delta t \ \alpha \cdot X_n + \sqrt{\Delta t} \ \sqrt{\varepsilon} \ \xi_n \ . \tag{5.2}$$

This is the Euler discretization of the overdamped Langevin equation $\dot{X} = -\alpha X + \sqrt{\varepsilon}\ \dot{B}$. We use the log-likelihood

$$g(x) = 10 \cdot \left(\frac{x^4}{4} - \frac{x^2}{2} \right).$$

As in the previous example, the optimal path goes to the right bump when $X_0 > 0$ and to the left when $X_0 < 0$. At $X_0 = 0$ there is no unique optimal path.

The linear drift makes it likely that DLM sample paths encounter the $x = 0$ line and the small noise results may not hold in this case. To illustrate the behavior and efficiency of the methods in this situation we perform experiments with varying values of ε and x_0. Specifically, for a fixed ε, we take $N = 10^3$ time steps with DLM, starting from initial conditions ranging from $x_0 = 10^{-1}$ to $x_0 = 10^{-5}$. We compute the averaging number of $x = 0$ crossings for each experiment. Figure 4 shows the results as well as the computed values of Q.

As can be seen in Figure 4(a), the predicted asymptotic scaling of Q only emerges for small ε; the critical value of ε at which the Q curve crosses over into the asymptotic regime decreases as x_0 approaches 0, making crossings more likely. Comparing Figures 4(a) and 4(b), we see that the asymptotic regime corresponds to values of ε small enough that the average number of crossings per sample is near zero. Closer examination of the data suggests that this critical ε scales roughly linearly with distance of the initial condition x_0 to $x = 0$. The example thus suggests that the efficiency of DLM may suffer if one encounters non-unique optimal paths while constructing the proposal distribution q sequentially, but the predicted Q scaling again holds if ε is small enough.

Finally, we note that even in the pre-asymptotic regime, the value of Q are $O(1)$, meaning the effective number of samples is $\approx N_e/2$, which is still a significant improvement over direct

sampling.

5.2 Example with a nonlinear SDE

Our second example is a stochastic version of an idealized geomagnetic pole reversal model due to Gissinger [16]:

$$\begin{array}{rcrcrcl} \dot{x}^1 &=& 0.119x^1 &-& x^2x^3 &+& \sqrt{\varepsilon}\dot{B}^1 \\ \dot{x}^2 &=& -0.1x^2 &+& x^1x^3 &+& \sqrt{\varepsilon}\dot{B}^2 \\ \dot{x}^3 &=& 0.9-x^3 &+& x^1x^2 &+& \sqrt{\varepsilon}\dot{B}^3 \end{array} \quad . \tag{5.3}$$

(In this section, x^k refers to the kth component of a vector x.) The $\varepsilon = 0$ system of ordinary differential equations has 3 unstable fixed points: $(0, 0, 0.9)$ and $p_\pm \approx (\mp 0.96, \pm 1.05, -0.109)$. It has a chaotic attractor on which trajectories circulate around either p_+ or p_- many times before making a quick transition to the other fixed point. See Figure 5. Following [16], we refer to these transitions as "pole reversals," since the second component $x^2(t)$ can be thought of as a proxy for the geomagnetic dipole field, and it changes signs at these transitions.

Here, we consider Eq. (5.3) with $\varepsilon > 0$. We start with an initial condition near p_+, and after $N = 100$ steps make an observation with log-likelihood $g(x) = ||x - y||^2/2$, where $x = (x^1, x^2, x^3)$. We view $y \in \mathbb{R}^3$ as the outcome of a "measurement" made at step N.

We consider two cases:

Case (a): The measured value y is near p_-, i.e., on the opposite "lobe" from the initial condition;

Case (b): y is near p_+, i.e., on the same "lobe" as the initial condition.

Figure 5 illustrates the initial conditions, data, and optimal paths for the two cases. Shown are trajectories of the deterministic model (light gray), representing the chaotic attractor. The dashed line is the most likely path with initial condition marked by "•" and with measured state at time $t = 10$ marked by "+"; this trajectory undergoes a "pole reversal" (Case (a)). The solid blue line represents the most likely path with initial condition "○" and observation "×," and does not exhibit a pole reversal (Case (b)).

To see how the two cases differ, we fix $\varepsilon = 10^{-2}$ and apply the LM and DLM to generate 1200 sample paths in each case and plot marginals of the proposal distributions at two different times. In Case (a), we plot histograms of the marginal distributions at time $j\Delta t$ as marked by ◇ in Figure 5; in Case (b), we plot histograms of the marginal distributions at time $j\Delta t$ as marked by □. For each method, the resulting "triangle plot" consists of histograms of the one-dimensional marginals, $q(X_j^k|X_0)$ for $k \in \{1, 2, 3\}$, and the two-dimensional marginals, $q(X_j^k, X_j^\ell|X_0)$, $k \neq \ell$, of the proposal distributions. The triangle plots are shown in Figure 6. In each panel, the diagonal plots are the one-dimensional marginal distributions. The lower-triangular parts of each panel are the two-dimensional marginal distributions generated by LM, while the upper-triangular parts show marginals generated by DLM.

In Case (a), the marginal distributions of the DLM proposal are multimodal, possibly related to the underlying geometry of the strange attractor. In contrast, the LM proposal distribution misses this complexity altogether (as one might expect). Moving now to Case (b), which involves starting and end points on the same lobe connected by a shorter optimal path, the marginals are unimodal, and LM and DLM give more similar answers (though there is still significant deviation from gaussianity in the DLM proposal distribution).

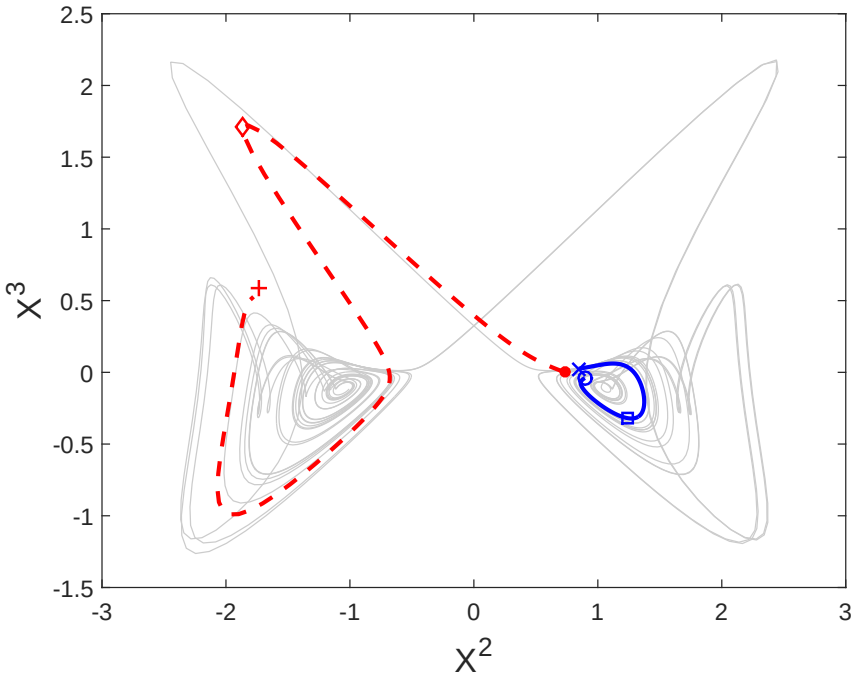


Figure 5: The Gissinger model and its phase space geometry. Shown are trajectories of the deterministic model (light gray) projected to the x^2-x^3 plane. The dashed line is the most likely path with initial condition marked by "•" and measured state at time $t = 10$ marked by "+"; this trajectory undergoes a "pole reversal" (Case (a)). The solid blue line represents the most likely path with initial condition "○" and observation "×" at $t = 10$, and does not exhibit a pole reversal (Case (b)). The symbols □ and ◇ are the times at which we computed the histograms in Figure 6.

Finally, we vary ε in Cases (a) and (b) and apply LM, SLM, DLM and SDLM. For each value of ε, we estimate Q for each of the 4 methods. The results are shown in Figure 7. Not surprisingly, LM breaks down for Case (a), in which the target distribution is likely multimodal. In contrast, both DLM and SDLM exhibit the predicted scaling. For Case (b), because the target distribution is unimodal, all four methods behave as predicted by the small noise theory.

Numerical details. The Gissinger model requires attention to numerical implementation when we compute its statistics.We describe our numerical implementation in detail.

(i) *Time-stepping.* The Euler scheme for the Gissinger model requires small time steps because of numerical instabilities. To improve stability, we discretize the drift part of Eq. (5.3) using a standard 4th-order Runge-Kutta (RK4) method, then adding IID $\mathcal{N}(0, \sqrt{\varepsilon}\sqrt{\Delta t}\, I)$ normal random vectors at each step. This yields a model of the form (2.1), where $\widetilde{f}(x, \Delta t)$ now represents one step of the RK4 scheme. In all the examples shown above, the time step is $\Delta t = 10^{-1}$.

(ii) *Estimation of Q.* In Figure 7, because of their different variances, we use 1200 sample

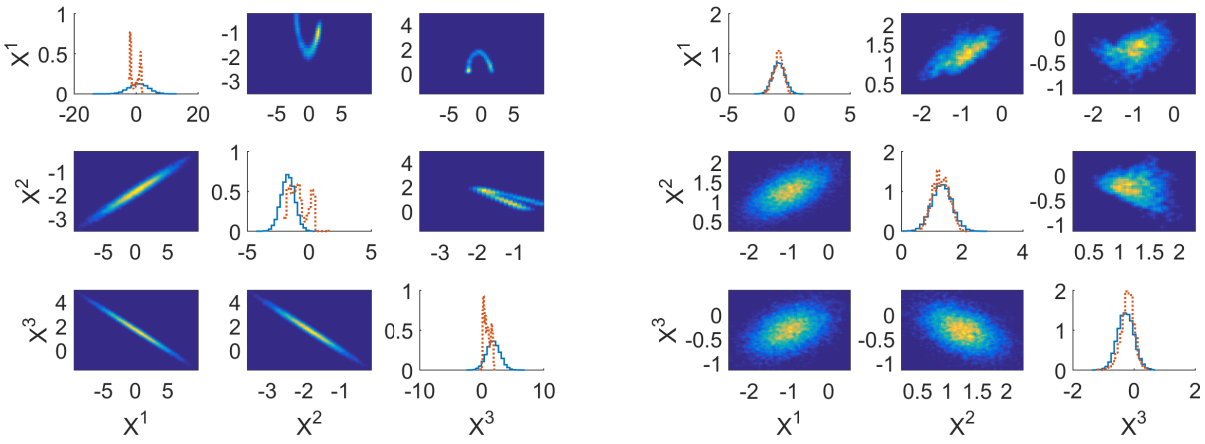


Figure 6: Final-time marginal distributions for the Gissinger model. In each panel, the diagonal plots are histograms for the final-time marginal proposal distributions for of x^1, x^2, and x^3 (solid = LM, dashed = DLM). The times at which the marginals are computed are marked by $\diamond$ in Figure 5 for Case (a), and $\square$ for Case (b). Plots on the lower-triangular submatrix are two-dimensional marginal proposal distributions computed by LM, while two-dimensional marginal proposal distributions computed by DLM form the upper-triangle (see text for details).

paths to estimate Q for DLM and for SDLM, and 12000 paths for LM and for SLM.

(iii) *Computing optimal paths.* Our methods requires computing optimal paths. For the Gissinger model, we use Newton's method. Since explicit analytical expressions for the gradient and the Hessian are available, this is relatively straightforward to program. To reduce the (fairly significant) computational cost of computing φ at each time step, we "guess" a good initialization for the optimization procedure using the solution from the previous time step using the linearized dynamics. See [20] for details.

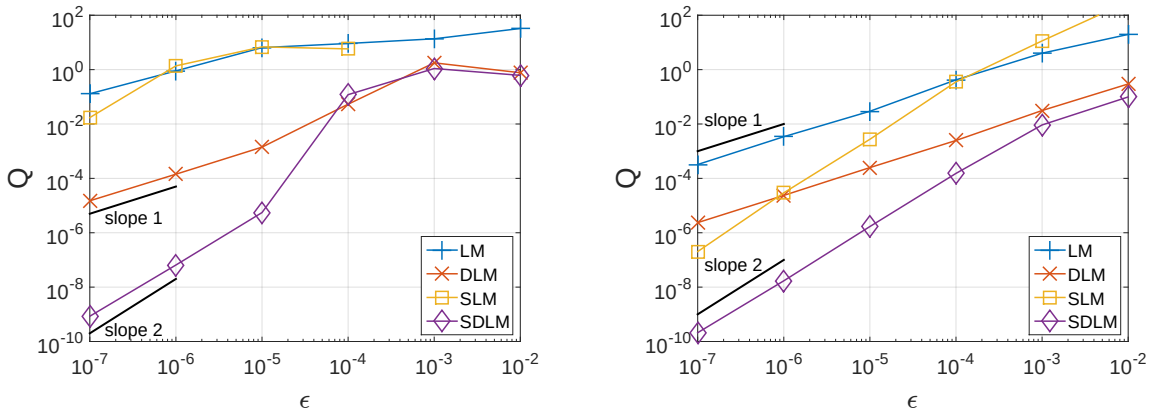
6 Continuous-time limit of dynamic linear map

So far, we have focused on time discretizations of SDEs. A natural question is what happens to the proposed algorithms in the limit $\Delta t \to 0$. In this section, we sketch some analytical arguments aimed at addressing these questions for scalar SDE. Though restrictive, we believe these results yield useful insights. A more complete and rigorous analysis is left for future work, as it is expected to be more involved.

6.1 Dynamic linear map

For scalar SDE, the DLM can be defined through the recursion

$$X_{n+1}^{\Delta t} = \varphi_{n+1}^{\Delta t}(X_n^{\Delta t}, n) + \sqrt{\Delta t}\ \sqrt{\varepsilon}\ \sqrt{\Sigma_{n+1}^{\Delta t}(X_n^{\Delta t}, n)}\ \xi_n, \tag{6.1}$$



Case (a) Case (b)

Figure 7: Relative variance Q as a function of ε for the Gissinger model. Case (a) involves a pole reversal, whereas Case (b) does not.

where $\varphi_n^{\Delta t}(x_0, m)$, $n \in \{m, m+1, \cdots, N\}$, is the optimal path (3.2) with prescribed initial condition $x_m = x_0 \in \mathbb{R}$, $\Sigma_{n+1}^{\Delta t}(X_n^{\Delta t}, n)$ is the $(1,1)$th entry of the Hessian of $F^{\Delta t}$ (see Eqs. (3.1) and (4.2)), and the ξ_n are independent standard normal random variables. Keeping in mind that $\varphi_n(x, n) = x$ for all n, the above can be written as

$$X_{n+1}^{\Delta t} = X_n^{\Delta t} + \Delta t \, \frac{\varphi_{n+1}^{\Delta t}(X_n^{\Delta t}, n) - \varphi_n^{\Delta t}(X_n^{\Delta t}, n)}{\Delta t} + \sqrt{\Delta t} \, \sqrt{\varepsilon} \, \sqrt{\Sigma_{n+1}^{\Delta t}(X_n^{\Delta t}, n)} \, \xi_n \, . \tag{6.2}$$

Our goal in this subsection is to sketch an argument suggesting that as $\Delta t \to 0$, solutions of (6.2) converge weakly [19] to those of

$$dX_t = \dot{\varphi}_t(X_t, t) \, dt + \sqrt{\varepsilon} \, \sigma \cdot dB_t \tag{6.3}$$

with $X_0 = x_0$. Since we consider "continuous time" and "discrete time" cases, we mark the discrete time case by a Δt superscript (i.e., in this section, the function in Eq. (3.1) is called $F^{\Delta t}$). In Eq. (6.3), "$\dot{\varphi}_s(x,t)$" denotes $\partial_s(\varphi_s(x,t))$, and the path $s \mapsto \varphi_s(x_0, t)$ $(t \leqslant s \leqslant T)$ minimizes the *action functional* [14]

$$F(x_{t:T} | x_t = x_0) = \frac{1}{2\sigma^2} \int_t^T (\dot{x}_s - f(x_s))^2 \, ds + g(x_T) \, , \quad \varphi_t(x_0, t) = x_0 \, . \tag{6.4}$$

This is the continuous-time analog of Eq. (3.1).

Eq. (6.3) was derived in [28] as the proposal for an importance sampling algorithm. This was later used in [29] for data assimilation in the small-noise regime. We assume minimizers φ of the action functional are twice-differentiable in the time parameter and satisfy the Euler-Lagrange equations; this can be justified via standard results from the calculus of variations (see, e.g., Section 3.1 of [15]). In what follows, we also assume that the action functional has a single global minimum for all initial positions x and initial time $t \in [0, T]$. This *unique optimal paths* assumption (the continuous-time analog of the unimodality of $p(x)$) implies that

$\dot{\varphi}_t(x,t)$ is defined everywhere. Without unique optimal paths, any analysis will require more care; see, e.g., [28] and references therein for a discussion of these and related issues. The assumption is natural for linear systems with unimodal likelihood functions $e^{-g/\varepsilon}$, and may hold (approximately) in nonlinear systems when T is small.

We now sketch our argument. We begin by recalling that a numerical approximation of an SDE *converges weakly with weak order k* if for all test functions $\psi \in C^{k+1}$ with at most polynomial growth,

$$\left| \mathrm{E}\left(\psi(X_N^{\Delta t})\middle|X_0\right) - \mathrm{E}\left(\psi(X_T)\middle|X_0\right)\right| = O(\Delta t^k) \tag{6.5}$$

as $\Delta t \to 0$. By standard results in the numerical analysis of SDEs, weak convergence is implied by "weak consistency" plus some mild polynomial growth conditions; see, e.g., Section 14.5 in [19] for details.

In the present context, consistency means that the factors $\left(\varphi_{n+1}^{\Delta t}(x,n) - \varphi_n^{\Delta t}(x,n)\right)/\Delta t$ and $\Sigma_{n+1}^{\Delta t}(x,n)$ in Eq. (6.2) approximate the corresponding factors in Eq. (6.3) ($\dot{\varphi}_t(x,n\Delta t)$ and σ^2, respectively). These we now prove.

Proposition. *Under the unique optimal path assumption, we have*

(a)

$$\frac{\varphi_{n+1}^{\Delta t}(x,n) - \varphi_n^{\Delta t}(x,n)}{\Delta t} = \dot{\varphi}_{n\Delta t}(x,n\Delta t) + O(\Delta t) \tag{6.6}$$

for all $n = 1,\ldots,N$ and $x \in \mathbb{R}$, and

(b)

$$\Sigma_{n+1}^{\Delta t}(x,n) = \sigma^2 + O(\Delta t). \tag{6.7}$$

Proof of (a). We begin by proving that φ and $\varphi^{\Delta t}$ satisfy the first variational equations for F and $F^{\Delta t}$, respectively (see Eqs. (6.4) and (3.1)). Without loss of generality, set $t = 0$ and $n = 0$, and write $\varphi(s) := \varphi_s(x_0, 0)$ for a given x_0. Then the first variational equation of F is the boundary value problem

$$-\ddot{\varphi}(s) + f'(\varphi(s))f(\varphi(s)) = 0 \tag{6.8}$$

$$\varphi(0) - x(0) = 0 \tag{6.9}$$

$$\dot{\varphi}(T) - f(\varphi(T)) + \sigma\ g'(\varphi(T)) = 0 \tag{6.10}$$

and the first variational equation for $F^{\Delta t}$ is

$$-\frac{\varphi_{k-1}^{\Delta t} - 2\varphi_k^{\Delta t} + \varphi_{k+1}^{\Delta t}}{\Delta t^2} + f'(\varphi_k^{\Delta t})f(\varphi_k^{\Delta t}) + \frac{f(\varphi_k^{\Delta t}) - f(\varphi_{k-1}^{\Delta t})}{\Delta t} - f'(\varphi_k^{\Delta t})\frac{\varphi_{k+1}^{\Delta t} - \varphi_k^{\Delta t}}{\Delta t} = 0 \tag{6.11}$$

$$\varphi_0^{\Delta t} - x_0 = 0 \tag{6.12}$$

$$\frac{\varphi_N^{\Delta t} - \varphi_{N-1}^{\Delta t}}{\Delta t} - f(\varphi_{N-1}^{\Delta t}) + \sigma\ g'(\varphi_N^{\Delta t}) = 0 \tag{6.13}$$

By the unique optimal path assumption, Eq. (6.8) is well-posed. Eq. (6.8) is equivalent to the system

$$-\dot{v} + f'(\varphi)f(\varphi) = 0 \qquad \text{and} \qquad \dot{\varphi} = v \tag{6.14}$$

with boundary conditions $\varphi(0) = 0$ and $v(T) - f(\varphi(T)) + \sigma\ g'(\varphi(T)) = 0$, and Eq. (6.11) is equivalent to the first-order-accurate finite difference approximation

$$-\frac{v_k - v_{k-1}}{\Delta t} + f'(\varphi_k^{\Delta t}) f(\varphi_k^{\Delta t}) + \frac{f(\varphi_k^{\Delta t}) - f(\varphi_{k-1}^{\Delta t})}{\Delta t} - f'(\varphi_k^{\Delta t})\ v_k = 0 \qquad \text{and} \qquad v_k = \frac{\varphi_{k+1}^{\Delta t} - \varphi_k^{\Delta t}}{\Delta t} \tag{6.15}$$

Convergence results for numerical approximations of two-point boundary value problems tell us that for first-order accurate finite difference schemes, point-wise errors are uniformly bounded by $C\Delta t$ for some $C > 0$ (see, e.g., [18] and references therein). In particular, we have $(\varphi_{n+1}^{\Delta t} - \varphi_n^{\Delta t})/\Delta t = v_n = \dot{\varphi}(n\Delta t) + O(\Delta t)$ for each n, as claimed. □

Proof of (b). To prove (6.7), we consider the second variational equations of F and $F^{\Delta t}$. For F, we obtain a Sturm-Liouville boundary value problem

$$\begin{aligned} (Lu)(s) &= 0 \\ u(0) &= 0 \\ u'(T) + (-f'(\varphi(s)) + \sigma\ g''(\varphi(T)))u(T) &= 0 \end{aligned}$$

where the operator L is defined by

$$Lu = -u''(s) + (f'(\varphi(s))^2 + f''(\varphi(s)) f(\varphi(s)))u(s),$$

φ is the solution to the first variational equation, and u is a test function. The second variational equation for $F^{\Delta t}$ is

$$\begin{aligned} (H/\Delta t)u^{\Delta t} &= 0 \\ u_0^{\Delta t} &= 0 \\ \frac{u_N^{\Delta t} - u_{N-1}^{\Delta t}}{\Delta t} - f'(\varphi_{N-1}^{\Delta t})u_{N-1}^{\Delta t} + \sigma\ g''(\varphi_N^{\Delta t})u_N^{\Delta t} &= 0 \end{aligned}$$

where H is the Hessian of $F^{\Delta t}$, and

$$\begin{aligned} (H/\Delta t)u^{\Delta t} = &-\frac{u_{k-1}^{\Delta t} - 2u_k^{\Delta t} + u_{k+1}^{\Delta t}}{\Delta t^2} + (f'(\varphi_k^{\Delta t})^2 + f(\varphi_k^{\Delta t}) \cdot f''(\varphi_k^{\Delta t}))u_k^{\Delta t} \\ &+ \frac{f'(\varphi_k^{\Delta t})u_k^{\Delta t} - f'(\varphi_{k-1}^{\Delta t})u_{k-1}^{\Delta t}}{\Delta t} - \frac{u_{k+1}^{\Delta t} - u_k^{\Delta t}}{\Delta t} \cdot f'(\varphi_k^{\Delta t}) - \frac{\varphi_{k+1}^{\Delta t} - \varphi_k^{\Delta t}}{\Delta t} \cdot f''(\varphi_k^{\Delta t}))u_k^{\Delta t}. \end{aligned}$$

Note that the discrete equations can also be obtained by applying a first order discretization scheme to the continuous equations.

The differential operator L has an associated Green's function

$$K(t,s) = \frac{1}{y_1'(0)y_2(0)} \begin{cases} y_1(t)y_2(s) & 0 < t < s \\ y_2(t)y_1(s) & 0 < s \leqslant t \end{cases}$$

where y_1 is a solution that satisfies the left Dirichlet boundary condition, while the solution y_2 satisfies the mixed boundary condition on the right. The analog of the Green's function for

the discretized problem is H^{-1}. Specifically, the first element of the first row of H^{-1} is a second order approximation of the Green's function $K(\Delta t, \Delta t)$:

$$(H^{-1})_{1,1} = K(\Delta t, \Delta t) + O(\Delta t^2)\,. \tag{6.16}$$

A Taylor expansion of K at the origin gives

$$K(\Delta t, \Delta t) = \sigma^2 \Delta t + \Delta t^2 \frac{y_2'(0)}{y_2(0)} + O(\Delta t^3)\,, \tag{6.17}$$

Combined, we thus have

$$(H^{-1})_{1,1} = \sigma^2 \Delta t + O(\Delta t^2) \tag{6.18}$$

Since $\Sigma_{n+1}^{\Delta t}(x,n) = (H^{-1})_{1,1}/\Delta t$, this shows that $\Sigma_{n+1}^{\Delta t}(x,n) = \sigma^2 + O(\Delta t)$. □

6.2 Small noise analysis for the continuous-time limit of DLM

We investigate how the efficiency of the dynamic linear map, as measured by the quantity Q (see Eq. (2.5)), is affected by taking the $\Delta t \to 0$ limit, and apply the theory presented in [26] to show that Q scales linearly in the small noise parameter ε even as $\Delta t \to 0$.

First, we note that the weights of the continuous limit of the DLM follow from the Cameron-Martin-Girsanov Theorem [14]

$$w(X) \propto \exp\Big(-\frac{1}{\sqrt{\varepsilon}} \int_0^T v(X_s, s) \cdot dB_s - \frac{1}{2\varepsilon} \int_0^T v(X_s, s)^2 ds - \frac{1}{\varepsilon} g(X_T) \Big) \tag{6.19}$$

where $v(x,t) = \sigma^{-1} \cdot (\varphi_t'(x,t) - f(x))$. The relative variance of the weights can be written as

$$Q = e^{-(V(0,x_0) - 2G(0,x_0))/\varepsilon} - 1, \tag{6.20}$$

where

$$G(x,t) = -\varepsilon \log(\mathrm{E}_q[w | x_t = x])$$
$$V(x,t) = -\varepsilon \log(\mathrm{E}_q[(w)^2 | x_t = x])$$

In [26], it was shown that V can be expanded in powers of ε when the minimizer φ of (6.4) is unique for all (x,t) in the domain. A calculation shows that G can also be expanded in powers of ε, with similar coefficients. In summary, we have

$$G(x,t) = G_0(x,t) + \varepsilon \cdot G_1(x,t) + \varepsilon^2 \cdot G_2(x,t) + O(\varepsilon^3) \tag{6.21}$$
$$V(x,t) = V_0(x,t) + \varepsilon \cdot V_1(x,t) + \varepsilon^2 \cdot V_2(x,t) + O(\varepsilon^3) \tag{6.22}$$

where the coefficients G_i, V_i, $i = 0, 1, 2$, satisfy the following system of PDEs:

$$\partial_t G_0 + f\,\partial_x G_0 - \frac{\sigma^2}{2}(\partial_x G_0)^2 = 0, \quad G_0(x,T) = g(x)$$

$$\partial_t V_0 + (f + \sigma^2 \partial_x G_0)\cdot \partial_x V_0 - \frac{\sigma^2}{2}(\partial_x V_0)^2 - \sigma^2(\partial_x G_0)^2 = 0, \quad V_0(x,T) = 2g(x)$$

$$\partial_t G_1 + f \cdot \partial_x G_1 + \frac{\sigma^2}{2}\partial_{xx} G_0 - \sigma^2 \partial_x G_0 \cdot \partial_x G_1 = 0, \quad G_1(x,T) = 0$$

$$\partial_t V_1 + (f + \sigma^2 \partial_x G_0)\cdot \partial_x V_1 + \frac{\sigma^2}{2}\partial_{xx} V_0 - \sigma^2 \partial_x V_0 \cdot \partial_x V_1 = 0, \quad V_1(x,T) = 0$$

$$\partial_t G_2 + f \cdot \partial_x G_2 + \frac{\sigma^2}{2}\partial_{xx} G_1 - \sigma^2 \partial_x G_0 \cdot \partial_x G_2 - \frac{\sigma^2}{2}(\partial_x G_1)^2 = 0, \quad G_2(x,T) = 0$$

$$\partial_t V_2 + (f + \sigma^2 \partial_x G_0)\cdot \partial_x V_2 + \frac{\sigma^2}{2}\partial_{xx} V_1 - \sigma^2 \partial_x V_0 \cdot \partial_x V_2 - \frac{\sigma^2}{2}(\partial_x V_1)^2 = 0, \quad V_2(x,T) = 0.$$

(These equations are similar in structure to those of the WKB approximation [3], with the leading order term given by a nonlinear PDE of Hamilton-Jacobi type and a hierarchy of linear transport equations for the higher-order terms.) One can check that $V_0 = 2G_0$ and $V_1 = 2G_1$, but $V_2 \neq 2G_2$. Combining the expansions (6.21) and (6.22) we thus have

$$V(x_0, 0) - 2G(x_0, 0) = \varepsilon^2 K_2 + O(\varepsilon^3), \tag{6.23}$$

where $K_2 = V_2 - 2G_2$ satisfies

$$\partial_t K_2 + f \cdot \partial_x K_2 - \sigma^2 \partial_x G_0 \cdot \partial_x K_2 - \sigma^2 (\partial_x G_1)^2 = 0, \quad K_2(x,T) = 0. \tag{6.24}$$

Using (6.23) in the expression of the relative variance Q in (3), and expanding in ε results in

$$Q = \varepsilon \cdot K_2(x_0, 0) + O(\varepsilon^2). \tag{6.25}$$

Thus, the performance criterion Q for this continuous time method scales linearly with ε.

7 Concluding discussion

In this paper, we study a class of importance samplers for SDEs designed for data assimilation tasks in the small (observation and dynamic) noise regime. We have extended a small noise analysis for implicit samplers [17] to importance sampling for SDEs. We have also shown that a symmetrization procedure, originally proposed in [17], can be applied effectively to obtain higher-order samplers for SDEs. Moreover, we have shown that a dynamic version of the importance sampler retains the same asymptotic performance but is more robust in problems with multimodal distributions.

Our work also points to a number of directions for future research:

(i) *Multimodal distributions.* Our analysis is limited to unimodal target distributions, but multimodal distributions do occur in practice. We believe an analysis for such problems (which necessarily means dealing with $q(x_{n+1}|x_n)$ with jump discontinuities), possibly

on concrete examples, would yield useful insights into the performance of DLM in more general situations than the ones examined here. One use for such an analysis is to compare DLM with other data assimilation methods, e.g., the ensemble Kalman filter, which may require less computation in nearly gaussian problems.

(ii) *Continuous time limits.* In discrete time, the dimension of the sampling problem we consider is equal to the dimension of a discretized path of an SDE and, thus, equal to the product of the state dimension and the number of time steps of the path. Our continuous time limit of the DLM for scalar SDE indicate that a large dimension due to a small time step is unproblematic, but our results do not indicate how the efficiency of DLM degrades when the dimension of the SDE is large.

(iii) *Symmetrization in continuous time.* Our results with symmetrized methods in discrete time are encouraging, but we currently do not have theoretical results on symmetrization in continuous time.

(iv) *Long timescales.* As mentioned in the Introduction, the methods discussed in this paper bear a close resemblance to methods proposed in [28] and [13] for rare event simulation. However, in this paper we have assumed a fixed final time T, whereas for many (if not most) rare event problems of interest, the relevant timescale tends to ∞ as $\varepsilon \to 0$ (e.g., $T = O(1/\varepsilon)$), and our methods are not expected to perform well on such long time scales. It would be of theoretical and practical interest to extend the ideas described here to the setting of rare event simulation, particularly the idea of symmetrization.

(v) *Problems that do not come from SDEs.* Also mentioned in the Introduction is the possibility of extending the methods proposed here, in particular symmetrization, to more general sequential Monte Carlo sampling problems.

8 Acknowledgments

KL and AL were supported in part by NSF grant DMS-1418775. MM was supported by NSF grant DMS-1619630, the Office of Naval Research (grant number N00173-17-2-C003), and by the Alfred P. Sloan Foundation. The authors thank Profs. Jonathan Goodman, Jonathan Weare, and Kostas Spiliopoulos for many helpful conversations and some of the references.

References

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. Stuart, *Importance sampling: computational complexity and intrinsic dimension*, Statistical Science, 32 (2017), pp. 405–431.

[2] S. Asmussen and P. W. Glynn, *Stochastic Simulation: Algorithms and Analysis*, Springer Science & Business Media, July 2007.

[3] C. M. BENDER AND S. A. ORSZAG, *Advanced Mathematical Methods for Scientists and Engineers. I*, Springer-Verlag, New York, 1999. Asymptotic methods and perturbation theory, Reprint of the 1978 original.

[4] N. BERGMAN, *Recursive Bayesian Estimation: Navigation and Tracking Applications*, Ph.D Dissertation, Linkoping University, Linkoping, Sweden, 1999.

[5] M. BOCQUET, C. PIRES, AND L. WU, *Beyond Gaussian statistical modeling in geophysical data assimilation*, Monthly Weather Review, 138 (2010), pp. 2997–3023.

[6] A. CHORIN, M. MORZFELD, AND X. TU, *Implicit particle filters for data assimilation*, Communications in Applied Mathematics and Computational Science, 5 (2010), pp. 221–240.

[7] A. J. CHORIN AND O. H. HALD, *Stochastic Tools in Mathematics and Science*, vol. 58, Springer, 2013.

[8] A. DOUCET, N. DE FREITAS, AND N. GORDON, *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, Springer New York, 2001.

[9] P. DUPUIS, K. SPILIOPOULOS, AND H. WANG, *Rare event simulation for rough energy landscapes*, in Proceedings of the 2011 Winter Simulation Conference (WSC), IEEE, 2011, pp. 504–515.

[10] ——, *Importance sampling for multiscale diffusions*, Multiscale Modeling & Simulation, 10 (2012), pp. 1–27.

[11] P. DUPUIS, K. SPILIOPOULOS, X. ZHOU, ET AL., *Escaping from an attractor: Importance sampling and rest points I*, The Annals of Applied Probability, 25 (2015), pp. 2909–2958.

[12] P. DUPUIS AND H. WANG, *Importance sampling, large deviations, and differential games*, Stochastics: An International Journal of Probability and Stochastic Processes, 76 (2004), pp. 481–508.

[13] ——, *Subsolutions of an Isaacs equation and efficient schemes for importance sampling*, Mathematics of Operations Research, 32 (2007), pp. 723–757.

[14] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer, 1984.

[15] M. GIAQUINTA AND S. HILDEBRANDT, *Calculus of Variations. I*, vol. 310 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, 1996. The Lagrangian formalism.

[16] C. GISSINGER, *A new deterministic model for chaotic reversals*, The European Physical Journal B, 85 (2012), pp. 1–12.

[17] J. GOODMAN, K. K. LIN, AND M. MORZFELD, *Small-noise analysis and symmetrization of implicit monte carlo samplers*, Communications on Pure and Applied Mathematics, 69 (2016), pp. 1924–1951.

[18] H. B. KELLER, *Numerical Methods for Two-Point Boundary Value Problems*, Dover Publications, Inc., New York, 1992. Corrected reprint of the 1968 edition.

[19] P. KLOEDEN AND E. PLATEN, *Numerical Solution of Stochastic Differential Equations*, Applications of Mathematics, Springer-Verlag, 1992.

[20] A. LEACH, *Monte Carlo Methods for Stochastic Differential Equations and their Applications*, PhD thesis, University of Arizona, 2017.

[21] J. S. LIU AND R. CHEN, *Sequential Monte Carlo methods for dynamical systems*, J. Amer. Statist. Assoc., 93 (1998), pp. 1032–1044.

[22] L. MARTINO, V. ELVIRA, AND F. LOUZADA, *Effective sample size for importance sampling based on the discrepancy measures*", Signal Processing, 131 (2017), pp. 386–401.

[23] M. MORZFELD, X. TU, E. ATKINS, AND A. J. CHORIN, *A random map implementation of implicit filters*, Journal of Computational Physics, 231 (2012), pp. 2049–2066.

[24] E. I. OSTROVSKII, *Exact Asymptotics of Laplace Integrals for Nonsmooth Functions*, Mathematical Notes, 73 (2003), pp. 838–842.

[25] C. SNYDER, T. BENGTSSON, AND M. MORZFELD, *Performance bounds for particle filters using the optimal proposal*, Monthly Weather Review, 143 (2015), pp. 4750–4761.

[26] K. SPILIOPOULOS, *Nonasymptotic performance analysis of importance sampling schemes for small noise diffusions*, Journal of Applied Probability, 52 (2015), pp. 797–810.

[27] P. VAN LEEUWEN, *Particle filtering in geophysical systems*, Monthly Weather Review, 137 (2009), pp. 4089–4144.

[28] E. VANDEN-EIJNDEN AND J. WEARE, *Rare event simulation of small noise diffusions*, Communications on Pure and Applied Mathematics, 65 (2012), pp. 1770–1803.

[29] ——, *Data assimilation in the low noise regime with application to the kuroshio*, Monthly Weather Review, 141 (2013), pp. 1822–1841.