Unbiased Estimation using a Class of Diffusion Processes

BY HAMZA RUZAYQAT[1], ALEXANDROS BESKOS[2], DAN CRISAN[3], AJAY JASRA[1] & NIKOLAS KANTAS[3]

[1]Applied Mathematics and Computational Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, KSA. E-Mail: *hamza.ruzayqat@kaust.edu.sa, ajay.jasra@kaust.edu.sa*

[2]Department of Statistical Science, University College London, London, WC1E 6BT, UK. E-Mail: *a.beskos@ucl.ac.uk*

[3]Department of Mathematics, Imperial College London, London, SW7 2AZ, UK. E-Mail: *d.crisan@ic.ac.uk, n.kantas@ic.ac.uk*

Abstract

We study the problem of unbiased estimation of expectations with respect to (w.r.t.) π a given, general probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ that is absolutely continuous with respect to a standard Gaussian measure. We focus on simulation associated to a particular class of diffusion processes, sometimes termed the Schrödinger-Föllmer Sampler, which is a simulation technique that approximates the law of a particular diffusion bridge process $\{X_t\}_{t\in[0,1]}$ on $\mathbb{R}^d$, $d \in \mathbb{N}_0$. This latter process is constructed such that, starting at $X_0 = 0$, one has $X_1 \sim \pi$. Typically, the drift of the diffusion is intractable and, even if it were not, exact sampling of the associated diffusion is not possible. As a result, [10, 16] consider a stochastic Euler-Maruyama scheme that allows the development of biased estimators for expectations w.r.t. π. We show that for this methodology to achieve a mean square error of $\mathcal{O}(\epsilon^2)$, for arbitrary $\epsilon > 0$, the associated cost is $\mathcal{O}(\epsilon^{-5})$. We then introduce an alternative approach that provides unbiased estimates of expectations w.r.t. π, that is, it does not suffer from the time discretization bias or the bias related with the approximation of the drift function. We prove that to achieve a mean square error of $\mathcal{O}(\epsilon^2)$, the associated cost (which is random) is, with high probability, $\mathcal{O}(\epsilon^{-2}|\log(\epsilon)|^{2+\delta})$, for any $\delta > 0$. We implement our method on several examples including Bayesian inverse problems.

Keywords: Diffusions, Unbiased approximation, Schrödinger bridge, Markov chain simulation.
Corresponding author: Hamza Ruzayqat. E-mail: hamza.ruzayqat@kaust.edu.sa
AMS subject classifications: 60J60, 62D05, 65C40

1 Introduction

Let π be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $d \in \mathbb{N}$, with positive Lebesgue density – denoted also π – assumed to be known up-to a normalizing constant. In many applications in applied mathematics and statistics, it is often of interest to compute expectations of π-integrable functionals, $\varphi : \mathbb{R}^d \to \mathbb{R}$, that is compute $\pi(\varphi) := \int_{\mathbb{R}^d} \varphi(x)\pi(x)dx$, see for instance [21] and the references therein. There are numerous methodologies for the approximation of $\pi(\varphi)$ often based upon the simulation of Markov processes with the most prominent example being Markov chain Monte Carlo (MCMC). In this article we consider approximations based upon independent samples, each of the latter generated via an 'embarrassingly' parallel approach. Such schemes have become rather popular in the recent literature [9, 8, 12].

We will consider a stochastic differential equation (SDE) on the time domain $t \in [0, 1]$, starting from $X_0 = 0$ and satisfying a terminal constraint $X_1 \sim \pi$. This is an instance of a more general

problem of assigning an initial and final distribution to a Markov process, which was initially formulated by Schrödinger in [22] and later developed into a general stochastic bridge construction by Jamison in [11], whereby an additive drift function is used to ensure the terminal constraint $X_1 \sim \pi$ will be satisfied. In our case π is absolutely continuous w.r.t. a standard Gaussian measure, so this drift will be added to a d-dimensional Brownian motion, $\{W_t\}_{t\in[0,1]}$, whose terminal distribution is known: $W_1 \sim \mathcal{N}_d(0, I)$, denoting the d-dimensional Gaussian distribution of mean 0 and identity covariance. This gives the following $\mathbb{R}^d$-valued diffusion process:

$$dX_t = b(X_t, t)dt + dW_t, \quad X_0 = 0, \tag{1}$$

with

$$b(x, t) = \nabla \log \mathbb{E}_{x,t}[f(W_1)],$$

where for any $(x, t) \in \mathbb{R}^d \times [0, 1]$, ∇ denotes the gradient w.r.t x, $\mathbb{E}_{x,t}$ denotes the expectation w.r.t $\{W_s\}_{s\in[t,1]}$ starting at $W_t = x$ and note that $b(x, 1) = 0$. We remark that f corresponds to the analogous of a likelihood function for a standard Gaussian prior, i.e. for $z \in \mathbb{R}^d$ we have $f(z) = \pi(z)/\phi(z)$, with $\phi(z)$ the standard $d-$dimensional Gaussian density. Using standard manipulations the expression for b simplifies to

$$b(x, t) = \nabla \log \mathbb{E}[f(x + W_{1-t})] = \frac{\mathbb{E}_\phi[\nabla f(x + \sqrt{1-t}Z)]}{\mathbb{E}_\phi[f(x + \sqrt{1-t}Z)]} = \frac{1}{\sqrt{1-t}} \frac{\mathbb{E}_\phi[Z f(x + \sqrt{1-t}Z)]}{\mathbb{E}_\phi[f(x + \sqrt{1-t}Z)]},$$

with $\mathbb{E}_\phi$ denoting expectation w.r.t. a d-dimensional standard Gaussian. We refer to [11, 5] for a generalizations of (1) and more details on a general existence result obtained by means of Girsanov's theorem and more importantly establishing that $X_1 \sim \pi$. Based on [11] there is existence of a weak solution of (1) in $[0, T]$. This requires f to be bounded and $\mathbb{E}_{x,t}[f(W_1)]$ to be twice continuously differentiable in x and once in t, i.e. in $\mathcal{C}^{2,1}\left(\mathbb{R}^d \times [0, T)\right)$; see [5] for more details. To ensure the existence of a strong solution of (1) the drift b needs to satisfy certain conditions; see [16] for details.

The formulation in (1) was proposed in [10, 16] as a sampling scheme for π and as an alternative to MCMC. The authors used the name Schrödinger-Föllmer Sampler (SFS) inspired by the original Schrödinger problem in [22] and its links with the entropy based time reversal of SDEs by Föllmer [7]. The approach of [10, 16] is to discretize the process in time via an Euler-Maruyama scheme and to numerically approximate the drift using standard perfect Monte Carlo estimators. This approach can then be parallelized to produce $M \in \mathbb{N}$ independent samples of X_1 to approximate $\pi(\varphi)$. This type of embarrassingly parallel estimators are extremely attractive for their computational savings, versus conventional time averages that often appear in standard iterative (non-parallelisable) MCMC simulation. The possibility of massive parallelization in SFS is also quite competitive against various other recent MCMC schemes from 'uncorrected' discretized diffusion schemes such as the unadjusted Langevin method [6, 19]. Indeed, here we do not need to concern ourselves with long-time asymptotic behavior, as the solution of (1) at time 1 is exactly distributed according to π. In addition, the stochastic bridge in (1) and its generalizations have been to solve a certain optimal transport problem (see [5]). As a result, different sampling schemes have been proposed recently in [1, 3] using iterative proportional fitting within Sequential Monte Carlo and generative modeling respectively. Whilst these schemes are interesting and use variants of (1), they are iterative in nature and cannot be parallelised to the extent of SFS.

In this article we make several contributions to the SFS, that we now list.

1. We show that for the method in [10, 16] to obtain estimators of $\pi(\varphi)$ achieving a mean square error (MSE) of $\mathcal{O}(\epsilon^2)$, the associated cost is $\mathcal{O}(\epsilon^{-5})$ for an arbitrary $\epsilon > 0$.

2. We construct a *doubly randomized* estimator, based upon the ideas in [14]. This approach delivers unbiased estimates of finite variance (in contrast to the method in [10, 16] that is biased).

3. We show that the proposed estimator achieves an MSE of $\mathcal{O}(\epsilon^2)$ with an associated random computing cost that is, with high probability, $\mathcal{O}(\epsilon^{-2}|\log(\epsilon)|^{2+\delta})$, for any $\delta > 0$. The term high-probability simply means that to achieve the prescribed MSE with the given cost, this latter cost is achieved with probability $1-\zeta$ for some small $\zeta \in (0,1)$. We note that the expected cost of our method is infinite. Remark 3.1 explains this in detail.

4. We apply our new approach for several examples, including Bayesian inverse problems, and numerically verify the above theoretical findings.

The significance of our contributions can be explained as follows. In the context of 1. we establish that due to the Monte Carlo error in the drift and the bias of the time discretization, one requires a large computational effort to approximate $\pi(\varphi)$ with high precision. Therefore, whilst the trivially parallel nature of the estimator is intuitively appealing, the associated cost can be prohibitive. In 2. we then consider a methodology to remove the time discretization bias of the Euler-Maruyama scheme, based upon the randomization schemes of [18, 20]. As the standard approach in those papers cannot be implemented, due to the fact that the drift must be approximated using Monte Carlo, we show that ideas related to [14] can be adapted in the context of SFS to overcome biases due to both the time discretisation and the drift approximation. The overall methodology delivers unbiased estimators of finite variance, using only simulation of standard Gaussian random variables. This latter aspect of the new algorithm is particularly interesting, since the approaches for instance in [9, 8, 12], also deliver unbiased estimators, but one must resort to complex coupling techniques, whereas we show here that such involved constructs are not always needed. In 3., relying on tools from the analysis of time discretized diffusions, we show that our new method provides a substantial reduction in cost over the original method in [16].

This article is structured as follows. In Section 2 we present the approach in [10, 16] and our new unbiased algorithm. In Section 3 we show that a particular version of the Algorithm provides unbiased estimators of finite variance. Section 4 contains our numerical results. Appendix A collects some of the technical results are used in Section 3.

2 Algorithm

The apparent challenges with the simulation of the diffusion process in (1) are, firstly, that the drift function $b(x,t)$ is typically intractable and, secondly, even if $b(x,t)$ is available point-wise, exact simulation from (1) is not possible.

2.1 Approximate SFS using Euler-Maruyama discretization

Let $\Delta_l = 2^{-l}$, with $l \in \mathbb{N}_0$ given. Then, the approach of [10, 16] considers the Euler-Maruyama scheme, for $k \in \{0, 1, \ldots, \Delta_l^{-1} - 1\}$:

$$\widetilde{X}^{l,N}_{(k+1)\Delta_l} = \widetilde{X}^{l,N}_{k\Delta_l} + \hat{b}(\widetilde{X}^{l,N}_{k\Delta_l}, k\Delta_l)\Delta_l + W_{(k+1)\Delta_l} - W_{k\Delta_l}, \tag{2}$$

where independently of all other random variables we have $(W_{(k+1)\Delta_l} - W_{k\Delta_l}) \sim \mathcal{N}_d(0, \Delta_l I)$. In the following, $N \in \mathbb{N}$ will associated to the accuracy of the Monte Carlo estimator of the drift b.

The quantity $\hat{b}$ is a numerical approximation of b and is defined as:

$$\hat{b}(x,t) = \frac{\frac{1}{N}\sum_{i=1}^{N} \nabla f(x+\sqrt{1-t}Z^i)}{\frac{1}{N}\sum_{i=1}^{N} f(x+\sqrt{1-t}Z^i)}, \tag{3}$$

where for $i \in \{1,\dots,N\}$, $Z^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(0,I)$ are random variables independent of the sequence $(W_{(k+1)\Delta_l} - W_{k\Delta_l})$. It should be noted that the Gaussian random variables are *updated at each simulation time*. As we will see, this re-simulation is not necessary and in some instances one can substantially improve the algorithm if such re-simulation is avoided. The exact method of [16] is given in Algorithm 1.

Algorithm 1 Biased SFS with Euler-Maruyama and i.i.d. Monte Carlo estimation for b

---

Input: number of i.i.d. replicates, $M \in \mathbb{N}$; number of samples, $N \in \mathbb{N}$, for the approximation of the drift function b; level of discretization, $l \in \mathbb{N}_0$.

1. Repeat for $i \in \{1,2,\dots,M\}$:

   a. Initialise $\widetilde{X}_0^{l,N}(i) = 0$.

   b. Repeat for $k \in \{0,1,\dots,\Delta_l^{-1}-1\}$:

      i. Sample $Z_k^j(i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(0,I)$, $j \in \{1,\dots,N\}$, and compute:

$$\hat{b}(\widetilde{X}_{k\Delta_l}^{l,N}(i), k\Delta_l) = \frac{\frac{1}{N}\sum_{j=1}^{N} \nabla f(\widetilde{X}_{k\Delta_l}^{l,N}(i) + \sqrt{1-k\Delta_l}Z_k^j(i))}{\frac{1}{N}\sum_{j=1}^{N} f(\widetilde{X}_{k\Delta_l}^{l,N}(i) + \sqrt{1-k\Delta_l}Z_k^j(i))}.$$

      ii. Generate $(W_{(k+1)\Delta_l}(i) - W_{k\Delta_l}(i)) \sim \mathcal{N}_d(0,\Delta_l I)$ and set:

$$\widetilde{X}_{(k+1)\Delta_l}^{l,N}(i) = \widetilde{X}_{k\Delta_l}^{l,N}(i) + \hat{b}(\widetilde{X}_{k\Delta_l}^{l,N}(i), k\Delta_l)\Delta_l + (W_{(k+1)\Delta_l}(i) - W_{k\Delta_l}(i)).$$

2. Return $\widetilde{X}_1^{l,N}(1),\dots,\widetilde{X}_1^{l,N}(M)$.

Using Algorithm 1 one can compute Monte Carlo estimators of $\pi(\varphi)$, with $\varphi : \mathbb{R}^d \to \mathbb{R}$ a π-integrable function, simply by using the sample average:

$$\pi^M(\varphi) := \frac{1}{M}\sum_{i=1}^{M} \varphi(\widetilde{X}_1^{l,N}(i)). \tag{4}$$

Now, to study the mean square error of this method, we consider the standard Euler-Maruyama discretization:

$$\widetilde{X}_{(k+1)\Delta_l}^{l} = \widetilde{X}_{k\Delta_l}^{l} + b(\widetilde{X}_{k\Delta_l}^{l}, k\Delta_l)\Delta_l + W_{(k+1)\Delta_l} - W_{k\Delta_l}. \tag{5}$$

To assist our analysis, we make the following assumptions. For a vector $x \in \mathbb{R}^d$ (resp. matrix A) we write the j^{th}-element (resp. $(j,k)^{th}$-element) as x_j (resp. A_{jk}). Also, $\|\cdot\|_1$ is the L_1-norm.

(**A1**) a) There exist $0 < \underline{C} < \overline{C} < +\infty$ such that for any $x \in \mathbb{R}^d$

$$\underline{C} \le f(x) \le \overline{C}, \qquad \|\nabla f(x)\|_1 \le \overline{C}, \qquad \|\nabla^2 f(x)\|_1 \le \overline{C}.$$

b) There exists $C < +\infty$ such that for any $(x, y) \in \mathbb{R}^{2d}$

$$\max \left\{ |f(x) - f(y)|, \|\nabla f(x) - \nabla f(y)\|_1, \|\nabla^2 f(x) - \nabla^2 f(y)\|_1 \right\} \leq C\|x - y\|_2.$$

Below $\text{Lip}(\mathbb{R}^d)$ denotes the collection of measurable functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ so that for a $C < \infty$, for all $(x, y) \in \mathbb{R}^{2d}$ we have $|\varphi(x) - \varphi(y)| \leq C\|x - y\|_2$, with $\|\cdot\|_2$ denoting the Euclidean norm. Let $\mathcal{B}_b(\mathbb{R}^d)$ be the collection of measurable and bounded functions $\varphi : \mathbb{R}^d \to \mathbb{R}$.

We then have the result stated in Proposition 2.1 below. The associated technical result (Lemma A.3) can be found in Appendix A.

Proposition 2.1. *Assume (A1). Then for any $\varphi \in \text{Lip}(\mathbb{R}^d) \cap \mathcal{B}_b(\mathbb{R}^d) \cap \mathcal{C}^4\left(\mathbb{R}^d\right)$ there exists a $C < \infty$ such that for any $(l, N, M) \in \mathbb{N}_0 \times \mathbb{N}^2$ we have*

$$\mathbb{E}\Big[\big(\pi^M(\varphi) - \pi(\varphi)\big)^2\Big] \leq C\big(\tfrac{1}{N} + \tfrac{1}{M} + \Delta_l^2\big).$$

Proof. Throughout the proof, C is a finite, positive constant that does not depend on (l, N, M), with a value that may change from line-to-line. We have that

$$\pi^M(\varphi) - \pi(\varphi) = \tfrac{1}{M}\sum_{i=1}^{M}\big\{\varphi(\widetilde{X}_1^{l,N}(i)) - \varphi(\widetilde{X}_1^{l}(i))\big\} + \tfrac{1}{M}\sum_{i=1}^{M}\big\{\varphi(\widetilde{X}_1^{l}(i)) - \pi_l(\varphi)\big\} + \big\{\pi_l(\varphi) - \pi(\varphi)\big\}.$$

Here, $\widetilde{X}_1^l(i)$, $i \in \{1, \ldots, M\}$, are i.i.d. samples obtained via recursion (5), starting at $X_0^l = 0$, up until time instance 1. Also, $\pi_l(\varphi)$ is the expectation of φ w.r.t. the law of $\widetilde{X}_1^l$. Via the C_2-inequality ($\mathbb{E}[|a + b|^2] \leq 2\mathbb{E}[|a|^2] + 2\mathbb{E}[|b|^2]$, where a, b are random variables of finite second moments), we have the upper-bound

$$\begin{aligned}\mathbb{E}\Big[\big(\pi^M(\varphi) - \pi(\varphi)\big)^2\Big] \leq C\Bigg(&\mathbb{E}\Big[\Big(\tfrac{1}{M}\sum_{i=1}^{M}\big\{\varphi(\widetilde{X}_1^{l,N}(i)) - \varphi(\widetilde{X}_1^{l}(i))\big\}\Big)^2\Big] \\ &+ \mathbb{E}\Big[\Big(\tfrac{1}{M}\sum_{i=1}^{M}\big\{\varphi(\widetilde{X}_1^{l}(i)) - \pi_l(\varphi)\big\}\Big)^2\Big] + \big\{\pi_l(\varphi) - \pi(\varphi)\big\}^2\Bigg). \qquad (6)\end{aligned}$$

For the first-term on the R.H.S. of (6) one can use the conditional Jensen inequality, the Lipschitz property of φ followed by Lemma A.3 (note that in the latter result, the fact that the $Z^1, \ldots, Z^N$ are refreshed at each time, does not affect the proof, so the result still holds for the recursion used in Algorithm 1), to obtain

$$\mathbb{E}\Big[\Big(\tfrac{1}{M}\sum_{i=1}^{M}\big\{\varphi(\widetilde{X}_1^{l,N}(i)) - \varphi(\widetilde{X}_1^{l}(i))\big\}\Big)^2\Big] \leq \frac{C}{N}.$$

For the second term on the R.H.S. of (6), one can use standard results for i.i.d. variables to yield

$$\mathbb{E}\Big[\Big(\tfrac{1}{M}\sum_{i=1}^{M}\big\{\varphi(\widetilde{X}_1^{l}(i)) - \pi_l(\varphi)\big\}\Big)^2\Big] \leq \frac{C}{M}.$$

For the third term on the R.H.S. of (6), standard weak error results for the Euler discretization of diffusions ([17, Theorem 14.1.5]) give

$$\big\{\pi_l(\varphi) - \pi(\varphi)\big\}^2 \leq C\Delta_l^2.$$

The proof is now complete. □

As a result of Proposition 2.1, to achieve a mean square error of $\mathcal{O}(\epsilon^2)$, for some given $\epsilon > 0$, one must choose $N = \mathcal{O}(\epsilon^{-2})$, $M = \mathcal{O}(\epsilon^{-2})$ and $l = \mathcal{O}(|\log(\epsilon)|)$, yielding a cost of $\mathcal{O}(\epsilon^{-5})$.

Remark 2.1. *The smoothness requirement of the test function, $\varphi \in \mathcal{C}^4(\mathbb{R}^d)$, is only used in the final step to get a weak error of order 1 for the Euler approximation. Less smoothness will result in lower order, for instance measurable and bounded Lipschitz derivatives would result to the relevant term in the upper bound of Proposition 2.1 to be $C\Delta_l$ instead.*

Remark 2.2. *Compared to [5] we impose the more restrictive assumption (A1) as the analysis here and in the subsequent results in Section 3 is clear this way. In more details, the assumption on f in terms of the upper and lower bounds is one used often in the importance sampling literature (see for example [13, Assumption 2.2]) and relates essentially to the ability of the Gaussian to mimic the target π. This boundedness condition is purely qualitative and can be relaxed at quite considerable complications to the proof. The closer π is to a Gaussian, the more likely one can verify this assumption. Recall the drift coefficient can be written as $b(x,t) = \nabla \log h$, where $h(x,t) = \mathbb{E}[f(x + W_{1-t})]$. The function h is smooth (because of the mollification by the heat kernel), but in general f may not satisfy the upper and lower bounded condition required by (A1). Moreover, as $t \to 1$, the smoothness will vanish if f is not smooth. Nevertheless, we can treat the general case by working with a "proxy" of X, in other words, we apply the algorithm and the theoretical convergence argument to a process $\widetilde{X}$ that satisfies the equation*

$$d\widetilde{X}_t = \widetilde{b}(\widetilde{X}, t)dt + dW_t, \qquad \widetilde{X}_0 = 0,$$

where

$$\widetilde{b}(\widetilde{X}, t) = \nabla \log \mathbb{E}\left[\tilde{f}(x + W_{1+\epsilon-t})\right]$$

where $\tilde{f}$ is the original f "clipped" from above and below: $\tilde{f} = \max\{\alpha, \min\{f, 1/\alpha\}\}$ and we choose ϵ and α sufficiently small to assume that the law of X and the law of $\widetilde{X}$ are sufficiently close or that the boundedness deduced in Proposition 2.1 remain the same.

2.2 Unbiased Estimation

[18, 20] present a methodology developed in the setting that $Y_n \to Y_\infty$ w.r.t. L_2-norm, for squared integrable random variables $\{Y_n\}_{n\geq 1}$, Y_∞, and the objective is the unbiased estimation of $\mathbb{E}[Y_\infty]$. Extensions of the initial approach developed in [14] will prove useful in the context of the current work. Before we continue, we shall denote a consistent Monte Carlo based estimator of $b(x,t)$ with N samples as $\hat{b}_N(x,t)$. Examples include (3) or a convergent MCMC algorithm with target probability $\pi_{x,t}$, as we now explain. For $(x,t) \in \mathbb{R}^d \times [0,1]$ we have

$$b(x,t) = \mathbb{E}_{\pi_{x,t}}\left[\frac{\nabla f(x + \sqrt{1-t}Z)}{f(x + \sqrt{1-t}Z)}\right],$$

where $\mathbb{E}_{\pi_{x,t}}$ denotes expectation w.r.t. the probability measure

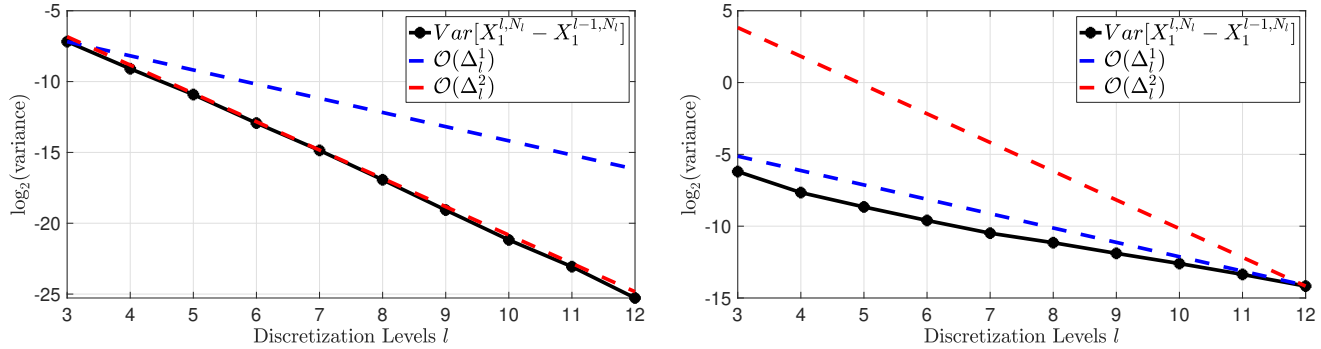
$$\pi_{x,t}(dz) \propto f(x + \sqrt{1-t}z)\phi(z)dz.$$

Thus, one can obtain a consistent estimator $\hat{b}(x,t)$ of $b(x,t)$ using e.g. MCMC methods. The exact form of the estimator is not specified for now.

We first assume access to two integer valued probability distribution $\mathbb{P}_R$ and $\mathbb{P}_P$ on $\mathbb{N}_0$ both on $\mathbb{N}_0$. Further let $1 \leq N_0 < N_1 < \cdots$ be a sequence of integers such that $N_p \to \infty$, as $p \to \infty$. Higher values of N_p will mean higher accuracy in estimation of b and at the limit this will lead to a perfect estimator. We will use samples of $\mathbb{P}_R$ and $\mathbb{P}_P$ to set the number of discretization levels via l and accuracy of $\hat{b}$ via N_p respectively. The objective is to develop an unbiased estimator of $\pi(\varphi)$. Following ideas in [14, Algorithm 5], we will now specify a method that aims to overcome both sources of bias we are confronted with, in a way that computing costs are reduced. We achieve debiasing via the 'single term estimator' approach, see [20]. The core idea of our method is to work with the random variable:

$$\widehat{\pi(\varphi)} = \frac{\left(\varphi(X_1^L[N_P]) - \varphi(X_1^{L-1}[N_P])\right) - \left(\varphi(X_1^L[N_{P-1}]) - \varphi(X_1^{L-1}[N_{P-1}]\right)}{\mathbb{P}_R(L)\mathbb{P}_P(P)} \tag{7}$$

with $L \sim \mathbb{P}_R$, $P \sim \mathbb{P}_P$. Also, $X_1^l[N_p]$, for $l, p \in \mathbb{N}_0$, denotes the approximation of X_1 obtained via recursion (2) for time-step $\Delta_l = 2^{-l}$ and the *same* N_p Gaussian variates for the estimation of the drift b at all locations and time instances where it is needed. Critically, the four terms in the nominator of (7) are carefully coupled. Also, simple conventions apply in the event that $L = 0$ or $P = 0$. The detailed approach is described in Algorithm 2. Note that in Step 1b. we assume that the computation of $\hat{b}$ is *dependent* across levels, at coinciding time points. One way to achieve this is to sample N_p Gaussians and use an estimator of the type (3) at both levels with the *same* Gaussians fixed once and for all – we believe this point is critical as illustrated in Figure 1. The figure shows that the variance of the increments $X_1^L[N] - X_1^{L-1}[N]$ decays much faster when the sample $\{Z_i\}_{i=1}^N$ are fixed. In Step 1b(iii), the term 'concatenated Wiener increment' means that the Wiener increment from time $k\Delta_{l-1}$ to $(k+1)\Delta_{l-1}$ at the coarser level $l-1$, for $l \in \mathbb{N}$, is the sum of the two Wiener increments from time $2k\Delta_l$ to $(2k+1)\Delta_l$ and from time $(2k+1)\Delta_l$ to $(2k+2)\Delta_l$ sampled at the finer level l, where $k \in \{0, \cdots, \Delta_{l-1}^{-1} - 1\}$. We note that an alternative to the single term estimator is the independent sum estimator, see [20], that often performs better in simulations; this latter estimator can be used with little extra difficulty in implementation.



(a) Fixed Gaussian Samples

(b) Unfixed Gaussian Samples

Figure 1: The log-variance of the level differences estimates for two cases: (a) The samples $\{Z_i\}_{i=1}^N$ are fixed for both levels l & $l-1$ and all time points. (b) $\{Z_i\}_{i=1}^N$ are sampled at every time point for both levels l & $l-1$. For simplicity we use a one-dimensional Gaussian density π.

The estimator $\widehat{\pi(\varphi)}$ developed in Algorithm 2 is unbiased. To show that it has finite variance, one strategy is to establish a bound of the type, for fixed $l, p \in \mathbb{N}$:

$$\mathbb{E}\left[\left(\varphi(X_1^l[N_p]) - \varphi(X_1^{l-1}[N_p]) - \varphi(X_1^l) + \varphi(X_1^{l-1})\right)^2\right] \leq \frac{C\Delta_l}{N_p}, \tag{9}$$

Algorithm 2 Unbiased Estimator of $\pi(\varphi)$.

Input: number of replicates, $M \in \mathbb{N}$; sequence $(N_p)_{p\in\mathbb{N}_0}$ and two positive probability mass functions, $\mathbb{P}_R$ and $\mathbb{P}_P$, on $\mathbb{N}_0$.

1. Repeat for $i \in \{1, 2, \ldots, M\}$:

   a. Sample $L^i \sim \mathbb{P}_R$ and $P^i \sim \mathbb{P}_P$.

   b. Sample the following variables.

      i. Sample N_{P^i} Gaussians $\{Z^j(i)\}_{j=1}^{N_{P^i}} \sim \mathcal{N}_d(0, I)$. Sample the Wiener increments $\{W_{(k+1)\Delta_{L^i}} - W_{k\Delta_{L^i}}\}_{k=0}^{\Delta_{L^i}^{-1}-1} \sim \mathcal{N}_d(0, \Delta_l I)$. Then generate $X_1^{L^i}[N_{P^i}]$ from recursion (2) with $l = L^i$ and $\hat{b} = \hat{b}_{N_{P^i}}$ using the same Gaussian variates, $\{Z^j(i)\}_{j=1}^{N_{P^i}}$, at every time instance.

      ii. Sample N_{P^i-1} Gaussians $\{Z^j(i)\}_{j=1}^{N_{P^i-1}} \sim \mathcal{N}_d(0, I)$. Generate $X_1^{L^i}[N_{P^i-1}]$ from recursion (2) using the same Wiener increments as in (i.) with $l = L^i$ and $\hat{b} = \hat{b}_{N_{P^i-1}}$ using the same Gaussian variates, $\{Z^j(i)\}_{j=1}^{N_{P^i-1}}$, at every time instance.

      iii. Generate $X_1^{L^i-1}[N_{P^i}]$ from recursion (2) with $l = L^i - 1$ and $\hat{b} = \hat{b}_{N_{P^i}}$ – use the same N_{P^i} Gaussian variates in (i.) and concatenated Wiener increments produced by the ones used in (i.) via the identity:

$$W_{(k+1)\Delta_{L^i-1}} - W_{k\Delta_{L^i-1}} = (W_{(2k+1)\Delta_{L^i}} - W_{2k\Delta_{L^i}}) + (W_{2(k+1)\Delta_{L^i}} - W_{(2k+1)\Delta_{L^i}}), \tag{8}$$

      with $k \in \{0, \ldots, \Delta_{L^i-1}^{-1} - 1\}$.

      iv. Generate $X_1^{L^i-1}[N_{P^i-1}]$ from recursion (2) with $l = L^i - 1$ and $\hat{b} = \hat{b}_{N_{P^i-1}}$ – use the same N_{P^i} Gaussian variates in (ii.) and the same concatenated Wiener increments as in (iii.).

   c. Set:

$$\widehat{\pi(\varphi)}(i) = \frac{\left(\varphi(X_1^{L^i}[N_{P^i}]) - \varphi(X_1^{L^i-1}[N_{P^i}])\right) - \left(\varphi(X_1^{L^i}[N_{P^i-1}]) - \varphi(X_1^{L^i-1}[N_{P^i-1}])\right)}{\mathbb{P}_R(L^i)\mathbb{P}_P(P^i)}.$$

   Apply the conventions:
   If $L^i = 0$ then set $\varphi(X_1^{L^i-1}[N_{P^i}]) = \varphi(X_1^{L^i-1}[N_{P^i-1}]) = 0$.
   If $P^i = 0$ then set $\varphi(X_1^{L^i}[N_{P^i-1}]) = \varphi(X_1^{L^i-1}[N_{P^i-1}]) = 0$.

2. Return $\widehat{\pi(\varphi)}(i)$, $i \in \{1, 2, \ldots, M\}$.

where C does not depend upon l, N_p and (X_1^l, X_1^{l-1}) are sampled from the exact Euler discretization under the same coupling procedure as the one described in Algorithm 2. Then, as in [14], setting $N_p = 2^p$ and $\mathbb{P}_P(l) = \mathbb{P}_R(l) \propto 2^{-l}(l+1)\log_2(l+2)^2$, $M = \epsilon^{-2}$, to achieve a variance (the estimator is unbiased) of $\mathcal{O}(\epsilon^2)$, for some $\epsilon > 0$ given, the cost with high probability (the cost is random) is $\mathcal{O}(\epsilon^{-2}|\log(\epsilon)|^{2+\delta})$, for any $\delta > 0$. The main challenge from here is to ascertain the bound (9).

3 Theoretical Results

3.1 Verifying the Bound (9)

We now consider a proof of (9) for a particular example. To simplify the notations, we will set for $l \in \{0, 1, \dots\}$ and $k \in \{0, 1, \dots, \Delta_l^{-1} - 1\}$

$$\widetilde{X}^{l,N}_{(k+1)\Delta_l} = \widetilde{X}^{l,N}_{k\Delta_l} + \hat{b}(\widetilde{X}^{l,N}_{k\Delta_l}, k\Delta_l)\Delta_l + W_{(k+1)\Delta_l} - W_{k\Delta_l},$$

where, for $x \in \mathbb{R}^d$

$$\hat{b}(x, k\Delta_l) = \frac{\frac{1}{N}\sum_{i=1}^N \nabla f(x + \sqrt{1 - k\Delta_l} Z^i)}{\frac{1}{N}\sum_{i=1}^N f(x + \sqrt{1 - k\Delta_l} Z^i)},$$

and, for $i \in \{1, \dots, N\}$, $Z^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(0, I)$. The exact Euler scheme is such that for $l \in \{0, 1, \dots\}$, $k \in \{0, 1, \dots, \Delta_l^{-1}\}$,

$$\widetilde{X}^{l}_{(k+1)\Delta_l} = \widetilde{X}^{l}_{k\Delta_l} + b(\widetilde{X}^{l}_{k\Delta_l}, k\Delta_l)\Delta_l + W_{(k+1)\Delta_l} - W_{k\Delta_l},$$

and note that the Brownian motions are all shared for both the approximated and exact Euler discretizations. Now, let $(l, t) \in \mathbb{N}_0 \times [0, 1]$ be given and define $\tau_t^l := \Delta_l \lfloor \frac{t}{\Delta_l} \rfloor$.

We have the following result. Associated technical details can be found in Appendix A.

Theorem 3.1. *Assume (A1). There exists $C < +\infty$ such that for any $(l, N) \in \mathbb{N}^2$ we have*

$$\mathbb{E}\left[\left\|\left(\widetilde{X}^{l,N}_1 - \widetilde{X}^{l-1,N}_1\right) - \left(\widetilde{X}^{l}_1 - \widetilde{X}^{l-1}_1\right)\right\|_2^2\right] \leq \frac{C\Delta_l}{N}.$$

Proof. We will consider for a fixed $t \in (0, 1]$ the quantity

$$\mathbb{E}\left[\left\|\left(\widetilde{X}^{l,N}_{\tau_t^l} - \widetilde{X}^{l-1,N}_{\tau_t^{l-1}}\right) - \left(\widetilde{X}^{l}_{\tau_t^l} - \widetilde{X}^{l-1}_{\tau_t^{l-1}}\right)\right\|_2^2\right]$$

and apply a Grönwall inequality type argument. Throughout all of our proofs C is a generic finite constant with value the can change upon each appearance, but will not depend upon (l, N, t). We have that for $t \in (0, 1]$

$$\mathbb{E}\left[\left\| \left(\widetilde{X}^{l,N}_{\tau_t^l} - \widetilde{X}^{l-1,N}_{\tau_t^{l-1}}\right) - \left(\widetilde{X}^{l}_{\tau_t^l} - \widetilde{X}^{l-1}_{\tau_t^{l-1}}\right) \right\|_2^2\right] = \mathbb{E}\left[\left\| T_1 + T_2 \right\|_2^2\right], \tag{10}$$

where we have defined

$$T_1 := \int_{\tau_t^{l-1}}^{\tau_t^l} \left(\hat{b}(\widetilde{X}^{l,N}_{\tau_s^l}, \tau_s^l) - b(\widetilde{X}^{l}_{\tau_s^l}, \tau_s^l)\right) ds,$$

$$T_2 := \int_0^{\tau_t^{l-1}} \left(\left(\hat{b}(\widetilde{X}^{l,N}_{\tau_s^l}, \tau_s^l) - \hat{b}(\widetilde{X}^{l-1,N}_{\tau_s^{l-1}}, \tau_s^{l-1})\right) - \left(b(\widetilde{X}^{l}_{\tau_s^l}, \tau_s^l) - b(\widetilde{X}^{l-1}_{\tau_s^{l-1}}, \tau_s^{l-1})\right)\right) ds.$$

By the C_2 and Jensen inequality, we have that

$$\mathbb{E}\left[\left\|T_1\right\|_2^2\right] \leq C\left(\tau_t^l - \tau_t^{l-1}\right) \int_{\tau_t^{l-1}}^{\tau_t^l} \mathbb{E}\left[\left\|\left(\hat{b}(\widetilde{X}^{l,N}_{\tau_s^l}, \tau_s^l) - b(\widetilde{X}^{l}_{\tau_s^l}, \tau_s^l)\right)\right\|_2^2\right] ds,$$

$$\mathbb{E}\left[\left\|T_2\right\|_2^2\right] \leq C\,\mathbb{E}\left[\left\| \int_0^{\tau_t^{l-1}} \left(\left(\hat{b}(\widetilde{X}^{l,N}_{\tau_s^l}, \tau_s^l) - \hat{b}(\widetilde{X}^{l-1,N}_{\tau_s^{l-1}}, \tau_s^{l-1})\right) - \left(b(\widetilde{X}^{l}_{\tau_s^l}, \tau_s^l) - b(\widetilde{X}^{l-1}_{\tau_s^{l-1}}, \tau_s^{l-1})\right)\right) ds \right\|_2^2\right].$$

We first treat the upper-bound for $\mathbb{E}\left[\left\|T_1\right\|_2^2\right]$. We have that

$$\int_{\tau_t^{l-1}}^{\tau_t^l} \mathbb{E}\left[\left\|\left(\hat{b}(\widetilde{X}^{l,N}_{\tau_s^l},\tau_s^l)-b(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)\right)\right\|_2^2\right] ds \leq$$
$$C\left\{\int_{\tau_t^{l-1}}^{\tau_t^l} \mathbb{E}\left[\left\|\left(\hat{b}(\widetilde{X}^{l,N}_{\tau_s^l},\tau_s^l)-\hat{b}(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)\right)\right\|_2^2\right] ds + \int_{\tau_t^{l-1}}^{\tau_t^l} \mathbb{E}\left[\left\|\left(\hat{b}(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)-b(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)\right)\right\|_2^2\right] ds\right\}.$$

Lemma A.1 followed by Lemma A.3 give

$$\mathbb{E}\left[\left\|\left(\hat{b}(\widetilde{X}^{l,N}_{\tau_s^l},\tau_s^l)-\hat{b}(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)\right)\right\|_2^2\right] \leq \frac{C}{N},$$

with C independent of s. Also, application of Lemma A.2 gives

$$\mathbb{E}\left[\left\|\left(\hat{b}(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)-b(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)\right)\right\|_2^2\right] \leq \frac{C}{N}.$$

Thus,

$$\mathbb{E}\left[\left\|T_1\right\|_2^2\right] \leq \frac{C\Delta_l^2}{N}. \tag{11}$$

We now turn to the upper-bound for $\mathbb{E}\left[\left\|T_2\right\|_2^2\right]$. Adding and subtracting $\hat{b}(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)-\hat{b}(\widetilde{X}^{l-1}_{\tau_s^{l-1}},\tau_s^{l-1})$, followed by use of C_2 and Jensen inequality, gives

$$\mathbb{E}\left[\left\|T_2\right\|_2^2\right] \leq C\left\{\mathbb{E}\left[\left\|\int_0^{\tau_t^{l-1}}\left(\left(\hat{b}(\widetilde{X}^{l,N}_{\tau_s^l},\tau_s^l)-\hat{b}(\widetilde{X}^{l-1,N}_{\tau_s^{l-1}},\tau_s^{l-1})\right)-\left(\hat{b}(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)-\hat{b}(\widetilde{X}^{l-1}_{\tau_s^{l-1}},\tau_s^{l-1})\right)\right)ds\right\|_2^2\right]\right.$$
$$\left.+\ \mathbb{E}\left[\left\|\int_0^{\tau_t^{l-1}}\left(\left(\hat{b}(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)-\hat{b}(\widetilde{X}^{l-1}_{\tau_s^{l-1}},\tau_s^{l-1})\right)-\left(b(\widetilde{X}^{l}_{\tau_s^l},\tau_s^l)-b(\widetilde{X}^{l-1}_{\tau_s^{l-1}},\tau_s^{l-1})\right)\right)ds\right\|_2^2\right]\right\}.$$

For the first expectation in the above right-hand term one can use Lemma A.6 and for the second expectation one can apply Lemma A.9, to yield the following upper-bound

$$\mathbb{E}\left[\left\|T_2\right\|_2^2\right] \leq C\left\{\frac{\Delta_l}{N}+\int_0^t \mathbb{E}\left[\left\|\left(\widetilde{X}^{l,N}_{\tau_s^l}-\widetilde{X}^{l-1,N}_{\tau_s^{l-1}}\right)-\left(\widetilde{X}^{l}_{\tau_s^l}-\widetilde{X}^{l-1}_{\tau_s^{l-1}}\right)\right\|_2^2\right] ds\right\}.$$

The proof is now concluded by applying Grönwall's inequality and setting $t=1$. □

We will also use the notation that for a differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$, $\nabla_k\psi(x) = (\partial\psi/\partial x_k)(x)$, $k \in \{1,\ldots,d\}$.

Proposition 3.1. *Assume (A1). Then for any $\varphi \in \mathcal{B}_b(\mathbb{R}^d)$ with $(\partial\varphi/\partial x_k) \in \mathrm{Lip}(\mathbb{R}^d) \cap \mathcal{B}_b(\mathbb{R}^d)$, $k \in \{1,\ldots,d\}$ there exists a $C < +\infty$ such that for any $(l,N) \in \mathbb{N}^2$ we have*

$$\mathbb{E}\left[\left(\left(\varphi(\widetilde{X}_1^{l,N})-\varphi(\widetilde{X}_1^{l-1,N})\right)-\left(\varphi(\widetilde{X}_1^{l})-\varphi(\widetilde{X}_1^{l-1})\right)\right)^2\right] \leq \frac{C\Delta_l}{N}.$$

Proof. For any $(x,y,u,v) \in \mathbb{R}^{4d}$ we have the representation

$$\left(\varphi(x)-\varphi(y)\right)-\left(\varphi(u)-\varphi(v)\right) = T_1+T_2,$$

where we have defined

$$T_1 = \sum_{k=1}^{d} \int_0^1 (\nabla_k \varphi)\big(y + \lambda(x-y)\big) \times \big((x-y) - (u-v)\big)_k \, d\lambda,$$

$$T_2 = \sum_{k=1}^{d} \int_0^1 \big\{(\nabla_k \varphi)\big(y + \lambda(x-y)\big) - (\nabla_k \varphi)\big(v + \lambda(u-v)\big)\big\} \times (u-v)_k \, d\lambda.$$

Via the C_2-inequality, it suffices now to bound the second moments of T_1 and T_2, when evaluated at $(x, y, u, v) = (\widetilde{X}_1^{l,N}, \widetilde{X}_1^{l-1,N}, \widetilde{X}_1^{l}, \widetilde{X}_1^{l-1})$. For T_1, since $\nabla_k \varphi$ is bounded, one has the upper-bound

$$\mathbb{E}\left[\left\|\big(\widetilde{X}_1^{l,N} - \widetilde{X}_1^{l-1,N}\big) - \big(\widetilde{X}_1^{l} - \widetilde{X}_1^{l-1}\big)\right\|_2^2\right]$$

that is bounded by $C\Delta_l/N$ via Theorem 3.1. For T_2, the Lipschitz property of $\nabla_k \varphi$ provides the upper-bound

$$C \times \mathbb{E}\left[\left(\big\|\widetilde{X}_1^{l-1,N} - \widetilde{X}_1^{l-1}\big\|_2^2 + \big\|\big(\widetilde{X}_1^{l,N} - \widetilde{X}_1^{l-1,N}\big) - \big(\widetilde{X}_1^{l} - \widetilde{X}_1^{l-1}\big)\big\|_2^2\right) \cdot \big\|\widetilde{X}_1^{l} - \widetilde{X}_1^{l-1}\big\|_2^2\right].$$

For the term

$$\mathbb{E}\left[\big\|\widetilde{X}_1^{l-1,N} - \widetilde{X}_1^{l-1}\big\|_2^2 \cdot \big\|\widetilde{X}_1^{l} - \widetilde{X}_1^{l-1}\big\|_2^2\right]$$

one can use Cauchy-Schwarz, Lemma A.3 and the convergence of the Euler approximation, to yield a bound of $C\Delta_l^2/N$. For the term

$$\mathbb{E}\left[\big\|\big(\widetilde{X}_1^{l,N} - \widetilde{X}_1^{l-1,N}\big) - \big(\widetilde{X}_1^{l} - \widetilde{X}_1^{l-1}\big)\big\|_2^2 \cdot \big\|\widetilde{X}_1^{l} - \widetilde{X}_1^{l-1}\big\|_2^2\right]$$

one can use a similar argument to yield a bound of $C\Delta_l^2/N$. This concludes the proof. □

Remark 3.1. *We expect that the bound in Theorem 3.1 (and hence Proposition 3.1) can be made sharper to $\mathcal{O}(\Delta_l^2/N)$. However, even with this result, one could not obtain finite variance and finite expected cost, as the rate of convergence of Monte Carlo estimators in $\mathbb{L}_2$ is $\mathcal{O}(N^{-1})$. That is, to ensure that the variance and expected cost are simultaneously finite, one needs any positive probability mass function $\mathbb{P}_{R,P}$ on $\mathbb{N}_0^2$ such that*

$$\sum_{(l,p)\in\mathbb{N}_0^2} \frac{\Delta_l^2}{N_p \mathbb{P}_{R,P}(l,p)} < \infty,$$

$$\sum_{(l,p)\in\mathbb{N}_0^2} \Delta_l^{-1} N_p \mathbb{P}_{R,P}(l,p) < \infty.$$

There is no $\mathbb{P}_{R,P}$ where the above conditions are satisfied simultaneously. If both inequalities were true, a straightforward application of the Cauchy-Schwartz inequality would give $\sum_{(l,p)\in\mathbb{N}_0^2} \sqrt{\Delta_l} < \infty$, which cannot hold. One possible way to address this could be to increase the rate in N by using Quasi-Monte Carlo estimators, which is something that we leave to future work. It may also be possible to sharpen the rate associated to the estimator as it is, in terms of N, using the ideas in [2]. However, the context of [2] is much simpler than that considered here and we expect that such a task to be rather arduous.

4 Numerical Results

4.1 MSE-to-Cost Rates

We now seek to verify Theorem 3.1 by computing the MSE-to-cost rates. Given $\varphi(x) = x$, we estimate the expectation $\pi(\varphi)$ by first running Algorithm 1 with both fixed and unfixed Gaussians to return the estimator in (4), which we denote by $\pi^{M,MC}(\varphi)$, second a multilevel estimator defined by the following collapsing sum identity, with Gaussians $\{Z_i\}_{i=1}^N$ are fixed for both levels l & $l-1$,

$$\pi^{M,ML}(\varphi) := \frac{1}{M}\sum_{i=1}^{M}\sum_{l=L_*}^{L}\frac{1}{N_l}\left\{\varphi(\widetilde{X}_1^{l,N_l}(i)) - \varphi(\widetilde{X}_1^{l-1,N_l}(i))\right\},$$

with the convention that $\varphi(\widetilde{X}_1^{-1,N_l}(i)) = 0$. Here $L_* \in \mathbb{N}_0$ is a starting level of discretization, L is the target level and $N_l = \mathcal{O}((L+1)2^{2L-l})$ (see e.g. [13] for more details on the choice of N_l). Finally, we run the unbiased estimator which returns

$$\pi^{M,UB}(\varphi) := \frac{1}{M}\sum_{i=1}^{M}\widehat{\pi(\varphi)}(i),$$

where the sequence $\{\widehat{\pi(\varphi)}(i)\}_{i=1}^M$ is the output of Algorithm 2. The MSE is the average of 100 independent simulations of each algorithm given by

$$MSE = \frac{1}{100}\sum_{j=1}^{100}\left[\left[\pi^{M,\cdot}(\varphi)\right]_j - \pi(\varphi)\right]^2,$$

where $\pi(\varphi)$ is the reference expectation.

4.1.1 Simulation Settings

In practice, one has to truncate the values of P and L in Algorithm 2. In all models below, we set $\mathbb{P}_R(L) = 2^{-1.5L}\mathbb{I}_{\{L_*,\cdots,L_{\max}\}}(L)$, where $L_*, L_{\max} \in \mathbb{N}_0$ and $L_* < L_{\max}$. Given L sampled from $\mathbb{P}_R$, we sample P from $\mathbb{P}_{P|L}(P|L) = g(P|L)\,\mathbb{I}_{\{P_*,\cdots,P_{\max}\}}(P)$ and set $N = N_0 2^P$ for some $N_0 \in \mathbb{N}$, where

$$g(P|L) = \begin{cases} 2^{4-P} & \text{if} \quad P \in \{P_*, \cdots, 4 \wedge (L_{\max} - L)\} \\ 2^{-P}\ P\ [\log_2(P)]^2 & \text{if} \quad P > 4 \end{cases}$$

This is the same choice as in [14]. We consider four different probability densities described below with $\varphi(x) = x$. For a given MSE $= \epsilon^2 > 0$, the cost to compute $\pi^M(\varphi)$ using Algorithm 1 is

$$\mathcal{C}_{\text{Single}} := NM2^L,$$

where it is assumed that the cost to simulate the SDE in (2) at a discretization level L is 2^L. As shown in Proposition 2.1, in order to have an MSE of order $\mathcal{O}(\epsilon^2)$, N, M and L must be chosen such that $N = \mathcal{O}(\epsilon^{-2})$, $M = \mathcal{O}(\epsilon^{-2})$ and $L = \mathcal{O}(|\log_2(\epsilon)|)$. The cost of the multilevel algorithm is

$$\mathcal{C}_{\text{ML}} := M\sum_{l=L_*}^{L} N_l\ 2^l,$$

where $N_l = \mathcal{O}((L - L_* + 1)2^{2L-l})$ and M & L as before. The expected cost of the proposed method in Algorithm 2 is

$$\overline{\mathcal{C}_{\text{Ub}}} := \frac{1}{100}\sum_{j=1}^{100}\mathcal{C}_j, \qquad \mathcal{C}_j = \sum_{i=1}^{M}\text{cost}^i,$$

where

$$\text{cost}^i = \begin{cases} N_0\ 2^{L_*} & \text{if} \quad L^i = L_*,\ P^i = P_* \\ N_0\ (2^{L^i} + 2^{L^i-1}) & \text{if} \quad L^i > L_*,\ P^i = P_* \\ (N_{P^i} + N_{P^i-1})\ 2^{L_*} & \text{if} \quad L^i = L_*,\ P^i > P_* \\ (N_{P^i} + N_{P^i-1})\ (2^{L^i} + 2^{L^i-1}) & \text{if} \quad L^i > L_*,\ P^i > P_* \end{cases},$$

with $L^i \sim \mathbb{P}_R$, and $N_{P^i} = N_0\ 2^{P^i}$, with $P^i \sim \mathbb{P}_{P|L}$. The values of P_*, L_*, $P_{\max}$, $L_{\max}$ and N_0 for the models below are described in Table 1.

Model	P_*	$P_{\max}$	L_*	$L_{\max}$	N_0
One-dimensional Gaussian	2	15	1	8	10
Two-dimensional Gaussian mixture	1	9	2	8	10
Bayesian logistic regression	1	9	3	9	10
Double-well potential	1	9	6	12	8

Table 1: The choices of the different parameters in Algorithm 2 for each model.

4.1.2 Models

(a) **One-Dimensional Gaussian Distribution:**
In the first example, we take $\pi(x) = \phi(x; 1, 2)$, a one-dimensional normal Gaussian density with mean 1 (which is the reference expectation) and variance 2. Clearly the true reference is $\pi(\varphi) = 1$.

(b) **Two-Dimensional Gaussian Mixture Distribution:**
We consider a two-dimensional Gaussian mixture distribution with density given by

$$\pi(x) = \frac{1}{16}\sum_{i=1}^{16}\phi(x; \mu_i, 0.03\ I_2),$$

where I_2 is the 2×2 identity matrix and $\mu = (\mu_1, \cdots, \mu_{16}) = \{-1, -0.5, 0.5, 1\}\times\{-1, -0.5, 0.5, 1\}$ where $\mu_i \in \mathbb{R}^2$. The reference expectation in this example is $\pi(\varphi) = \frac{1}{16}\sum_{i=1}^{16}\mu_i$.

(c) **Bayesian Logistic Regression:**
Next we consider the binary logistic regression in which the binary observations $\{Y_i\}_{i=1}^n$ are conditionally independent Bernoulli random variables such that $Y_i \in \{0, 1\}$ and

$$\mathbb{P}(Y_i = 1|X_i = x_i, \beta) = \rho(\beta^T x_i),$$

where $\rho : \mathbb{R} \to (0, 1)$ defined by $\rho(w) = e^w/(1 + e^w)$ is the logistic function and X_i and β in $\mathbb{R}^d$ are the covariates and the unknown regression coefficients, respectively. The prior density for the parameter β is a multivariate normal Gaussian given by

$$pr(\beta) = \phi(\beta; 0, \Sigma_\beta)$$

where Σ_β is defined through its inverse $\Sigma_\beta^{-1} = \frac{1}{n}(\sum_{i=1}^n X_i X_i^T)$. The covarites vectors $\{X_i\}_{i=1}^n$ are sampled independently from $\mathcal{U}\{-1,1\}^d$ which are then standardized. The density of the posterior distribution of β is given by

$$\pi\left(\beta|\{X_i = x_i, Y_i = y_i\}_{i=1}^n\right) \propto \exp\left(\sum_{i=1}^n \left[y_i \beta^T x_i - \log(1+\exp(\beta^T x_i))\right] - \frac{1}{2}\beta^T \Sigma_\beta^{-1} \beta\right).$$

We set $d = 5$, $n = 100$ and sample the binary observations $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} \text{Bernoulli}(\rho(\beta_*^T x_i))$, where $\beta_* = \mathbf{1}_d$, a vector of ones. We take the reference expectation to be the mean of 100 simulations of a random-walk Metropolis-Hastings MCMC with 10^7 samples and a burn-in of 10^3.

(d) **Double-Well Potential:**
Finally, we consider sampling from $\pi(x) = \exp(-U(x))$, where U is the double-well potential given by

$$U(x) = \frac{1}{4}\|x\|_2^4 - \frac{1}{2}\|x\|_2^2, \qquad x \in \mathbb{R}^d.$$

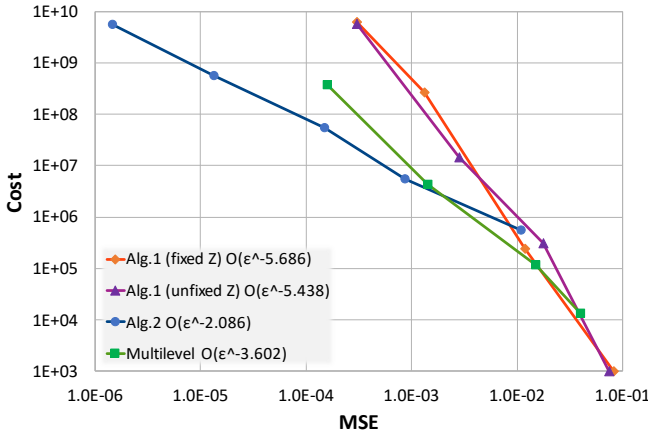
The double-well potential is one of several quartic potentials of substantial importance in quantum mechanics and quantum field theory [15] for the investigation of different physical phenomena or mathematical features.
In this example we test the algorithms in a high dimensional setting where we set $d = 30$. The true reference expectation is $\pi(\varphi) = 0$.

4.1.3 Results

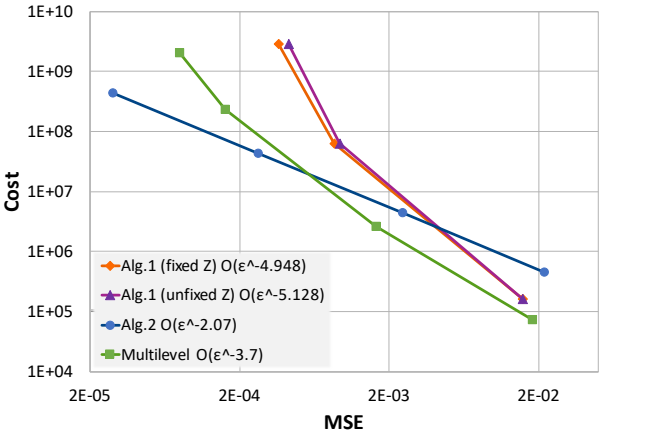
In Figure 2, we plot the MSE against the cost obtained by running the original SFS method presented in Algorithm 1, where the Gaussians Z_k^j, $j \in \{1, \cdots, N\}$, in step 1b.i., either fixed for all $k \in \{0, 1, \cdots, \Delta_l^{-1} - 1\}$ or sampled for each k, the multilevel method with fixed Gaussians and finally the unbiased method presented in Algorithm 2. From the plots, we observe that to obtain an MSE of order $\mathcal{O}(\epsilon^2)$, for some $\epsilon > 0$, the cost of Algorithm 2 is of order $\mathcal{O}(\epsilon^{-2})$ in all models. The cost order is much higher in the case of Algorithm 1 with both fixed and unfixed Gaussians. In both situations the cost is of order $\mathcal{O}(\epsilon^{-(5+\alpha)})$, $\alpha > 0$, for models (a), (b) & (d), and $\mathcal{O}(\epsilon^{-4})$ for model (c). The cost order for the multilevel algorithm with fixed Gaussians is almost the same for models (a) & (b) $\mathcal{O}(\epsilon^{-3.7})$, but notably smaller for model (c) where it is of order $\mathcal{O}(\epsilon^{-3.13})$ and higher for model (d) where it is of order $\mathcal{O}(\epsilon^{-4})$. In all examples we note that the unbiased method of Algorithm 2 is more efficient for lower MSE followed by the multilevel implementation. It is worth noting that we are only comparing the theoretical costs obtained from the formulae in subsubsection 4.1.1, not the machine computational time. The algorithm presented here is embarrassingly parallelizable over M, making it much faster on multi-core workstations.

4.2 Bayesian Elliptic Inverse Problem

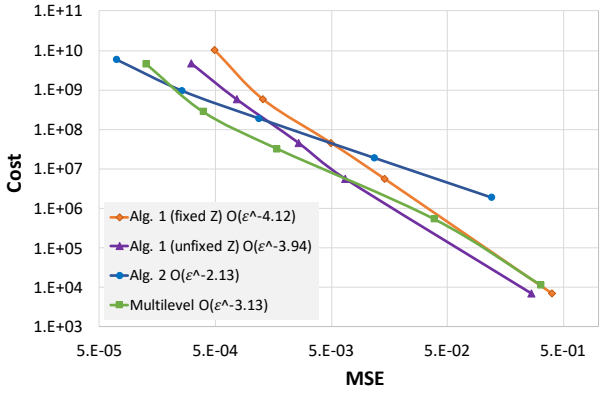
The objective of this section is to compare between the original algorithm and the proposed here in the context of Bayesian inverse problems. We consider estimates of expectations w.r.t. the posterior measure on some unknown field of interest using a Bayesian statistics approach of inverse problems that arise from the confluence of partial differential equations and observational data. However, due to floating-point precision limitations and high variance associated with computing



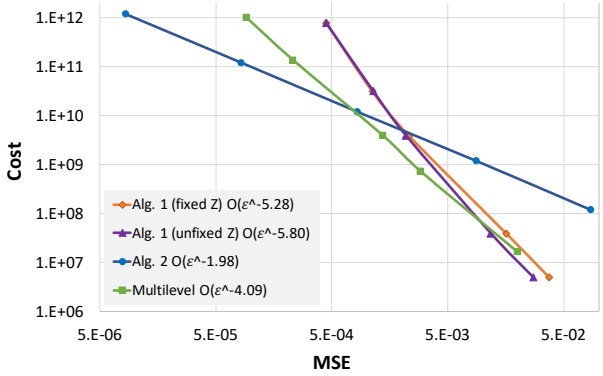
(a) One-dimensional Gaussian density



(b) Two-dimensional mixture of Gaussian densities



(c) Bayesian logistic regression



(d) Double-well potential

Figure 2: MSE versus cost (as computed in subsubsection 4.1.1) of running Algorithm 1 with both fixed (orange line) and unfixed (purple line) Gaussians, a multilevel algorithm (green line) and the unbiased estimation Algorithm 2 (blue line).

f, the unbiased method as presented in Algorithm 2 will not work well in general. Typically, in some cases, especially for high-dimensional distributions with densities that can be expressed as $\pi(x) \propto \exp(-U(x))$, the variable f, which can be written as $f(x) \propto \exp(-U(x) + \frac{1}{2}\|x\|^2)$, will register zero values in fixed point arithmetic, resulting in an infinite drift. Even with the common log-sum-exp trick, the values of $f(x + \sqrt{1-t}Z)$, $Z \sim \phi(z)$, may not be within the representable range of the computing machine. As a result, we advise using an alternative approach provided in Algorithms 3–4. We should highlight however that this algorithm will not be mathematically investigated in this study; instead, this will be the subject of future research.

4.2.1 Model

Consider a Bayesian inverse problem involving inference of the log-permeability coefficient of a 2D elliptic PDE in a bounded and open region $\Omega \subset \mathbb{R}^2$ with convex boundary $\partial\Omega \in C^0$, given noisy measurements of (some components of, or functions of) the associated solution field. In particular,

we consider the elliptic PDE

$$\begin{aligned} -\nabla \cdot (K(u)\nabla p) &= F \quad \text{on } \Omega \\ p &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{12}$$

where p is the forward state, e.g. the pressure field in a porous media flow, $K(u)$ is a scalar function that represents the permeability filed e.g. of a subsurface rock, and F represents the external force field. This represents a simplified model in groundwater flow. It is important to assume that $K(u)$ is positive in order to have a well-posed problem [23], and hence we write $K(u) = \exp(u)$ and consider the problem of determining u given a set of noisy observations of the pressure in the interior of Ω. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces. Then the inverse problem is to determine $u \in X$, given the data

$$y = \mathcal{G}(u) + \eta \quad \in Y, \tag{13}$$

where $\mathcal{G} : X \to Y$ is the observation operator and $\eta \in \mathcal{N}(0, \Gamma)$ for some trace class, positive, self-adjoint operator Γ on Y. The Bayesian approach to this inverse problem is to find a posterior probability measure μ^y on X such that Bayes' rule holds

$$\frac{d\mu^y}{d\mu_0}(u) \propto l(u; y),$$

where $l(u; y)$ is the measurement likelihood and μ_0 is the prior, which here is assumed to be Gaussian of the form $\mu_0 = \mathcal{N}(0, \mathcal{C})$, with $\mathcal{C}$ a trace class, positive, self-adjoint operator on X. The likelihood can be obtained from (13) as

$$l(u; y) = \exp\left(-\frac{1}{2}\|\Gamma^{-1/2}(y - \mathcal{G}(u))\|_Y^2\right).$$

Note that if $\mathcal{G}$ is linear and the prior is Gaussian, then the posterior μ^y will be Gaussian as well, which is the case here as the differential operator associated with the above PDE is linear.

Clearly, for computer implementation, one needs to discretize the prior, the likelihood, and hence the posterior. A standard finite element method (FEM) with linear triangular elements is employed to solve the forward problem in (12). The induced mesh consists of right triangles in the domain $\Omega = [0, 1] \times [0, 1]$. We denote by $\widehat{N}$ the set of nodes (vertices) in the mesh and $\hat{n}$ the number of nodes.

We assume the following additive noise-corrupted pointwise observation model

$$y_j = p(x_j) + \eta_j, \qquad j = 1, \cdots, J,$$

where J is the total number of observation locations, $\{x_j\}_{j=1}^J$ the set of nodes of at which the pressure p is observed, $\eta = (\eta_1, \cdots, \eta_J)$ a Gaussian noise distributed according to $\mathcal{N}(0, \Gamma)$, and y_j the actual noise-corrupted observation at node $x_j \in \widehat{N}$. For simplicity Γ is assumed to equal $\sigma_y^2 I_J$. A finite-dimensional approximation of the prior is given by $\widehat{\mu}_0 = \mathcal{N}(0, C)$, with the entries of the covariance matrix are given by

$$C_{ij}(x_i, x_j) = \sigma \exp\left(-\|x_i - x_j\|_2/\alpha\right),$$

where $x_i, x_j \in \widehat{N}$, $i, j = 1, \cdots, \hat{n}$, $\|\cdot\|_2$ is the Euclidean distance and $\sigma, \alpha > 0$ are hyperparameters.

Algorithm 3 Alternative Coupling for SFS

1. Input: $N \in \mathbb{N}$ the number of samples used to approximate b and a level $l \in \mathbb{N}$ of discretization.

2. For $k = 0, 1, \cdots, \Delta_l^{-1} - 1$, generate the increments of Brownian motion, $W^l_{(k+1)\Delta_l} - W^l_{k\Delta_l}$, used for the Euler approximation at level l. Concatenate the increments to generate $W^{l-1}_{(k+1)\Delta_{l-1}} - W^{l-1}_{k\Delta_{l-1}}$ for $k = 0, 1, \cdots, \Delta_{l-1}^{-1} - 1$. Set $X_0^l = X_0^{l-1} = 0$.

3. For $k \in \{0, 1, \ldots, \Delta_{l-1}^{-1} - 1\}$ perform the following:

   - Generate N samples from $\pi_{X^{l-1}_{k\Delta_{l-1}}, k\Delta_{l-1}}$ and compute

   $$\hat{b}(X^{l-1}_{k\Delta_{l-1}}, k\Delta_{l-1}) = \frac{1}{\sqrt{1 - k\Delta_{l-1}}} \frac{1}{N} \sum_{i=1}^{N} Z^i.$$

   - Compute:

   $$X^{l-1}_{(k+1)\Delta_{l-1}} = X^{l-1}_{k\Delta_{l-1}} + \hat{b}(X^{l-1}_{k\Delta_{l-1}}, k\Delta_{l-1})\Delta_{l-1} + W^{l-1}_{(k+1)\Delta_{l-1}} - W^{l-1}_{(k+1)\Delta_{l-1}}.$$

   - Then perform the following for $m = 0, 1$:
     - Compute the weights and normalize

     $$w^i_{[2(k-1)+m]\Delta_l} \propto \frac{f\left(X^l_{[2(k-1)+m]\Delta_l} + \sqrt{1 - [2(k-1)+m]\Delta_l} Z^i\right)}{f\left(X^{l-1}_{(k+1)\Delta_{l-1}} + \sqrt{1 - [2(k-1)+m]\Delta_l} Z^i\right)} \tag{14}$$

     then compute

     $$\hat{b}\left(X^l_{[2(k-1)+m]\Delta_l}, [2(k-1)+m]\Delta_l\right) = \sum_{i=1}^{N} w^i_{[2(k-1)+m]\Delta_l} \, Z^i. \tag{15}$$

     - Compute:

     $$\begin{aligned} X^l_{[2(k-1)+m+1]\Delta_l} =& X^l_{[2(k-1)+m]\Delta_l} + \hat{b}\left(X^l_{[2(k-1)+m]\Delta_l}, [2(k-1)+m]\Delta_l\right) \\ &+ W^l_{[2(k-1)+m+1]\Delta_l} - W^l_{[2(k-1)+m]\Delta_l}. \end{aligned}$$

     where the increments of the Brownian motion are concatenated from 2..

4. Return $X_1^l[N]$ and $X_1^{l-1}[N]$.

4.2.2 Alternative Approach to Algorithm 2

Define a probability density for any fixed $(x, t) \in \mathbb{R}^d \times [0, 1]$ as

$$\pi_{x,t}(z) = \frac{f(x + \sqrt{1-t}z)\phi(z)}{\int_{\mathbb{R}^d} f(x + \sqrt{1-t}z)\phi(z)dz}. \tag{16}$$

Algorithm 4 Alternative Unbiased Estimator of $\pi(\varphi)$.

Input: number of replicates, $M \in \mathbb{N}$; sequence $(N_p)_{p\in\mathbb{N}_0}$ and two positive probability mass functions, $\mathbb{P}_R$ and $\mathbb{P}_P$, on $\mathbb{N}_0$, $\mathbb{P}_R$ and $\mathbb{P}_P$.

1. Repeat for $i \in \{1, 2, \ldots, M\}$:

   a. Sample $L^i \sim \mathbb{P}_R$ and $P^i \sim \mathbb{P}_P$.

   b. If $L^i = 0$, generate $X_1^{L^i}[N_{P^i}]$ and $X_1^{L^i}[N_{P^i-1}]$ from recursion (2) using the same Wiener increments with $l = 0$ and $\hat{b}$ as in (17).

   c. Otherwise:

      i. Run Algorithm 3 to return $X_1^{L^i}[N_{P^i}]$ and $X_1^{L^i-1}[N_{P^i}]$.

      ii. Run Algorithm 3 to return $X_1^{L^i}[N_{P^i-1}]$ and $X_1^{L^i-1}[N_{P^i-1}]$ using the same Wiener increments as in (i).

   c. Set:

$$\widehat{\pi(\varphi)}(i) = \frac{\left(\varphi(X_1^{L^i}[N_{P^i}]) - \varphi(X_1^{L^i-1}[N_{P^i}])\right) - \left(\varphi(X_1^{L^i}[N_{P^i-1}]) - \varphi(X_1^{L^i-1}[N_{P^i-1}])\right)}{\mathbb{P}_R(L^i)\mathbb{P}_P(P^i)}.$$

   Apply the conventions:
   If $L^i = 0$ then set $\varphi(X_1^{L^i-1}[N_{P^i}]) = \varphi(X_1^{L^i-1}[N_{P^i-1}]) = 0$.
   If $P^i = 0$ then set $\varphi(X_1^{L^i}[N_{P^i-1}]) = \varphi(X_1^{L^i-1}[N_{P^i-1}]) = 0$.

2. Return $\widehat{\pi(\varphi)}(i)$, $i \in \{1, 2, \ldots, M\}$.

Then we have that

$$b(x,t) = \frac{1}{\sqrt{1-t}}\mathbb{E}_{\pi_{x,t}}[Z].$$

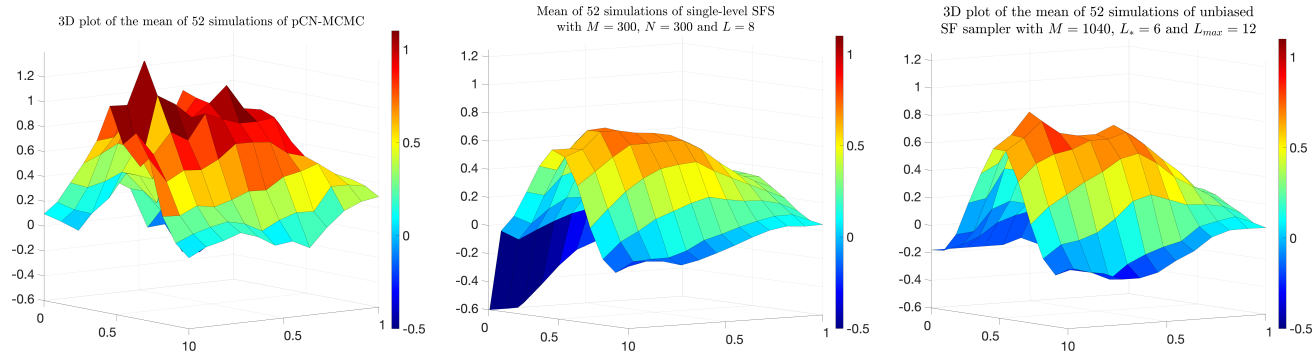
Consider the discretized SDE in (2) with $\hat{b}$ replaced by

$$\hat{b}(\widetilde{X}^l_{k\Delta_l}, k\Delta_l) = \frac{1}{\sqrt{1-k\Delta_l}}\frac{1}{N}\sum_{i=1}^{N} Z^i, \tag{17}$$

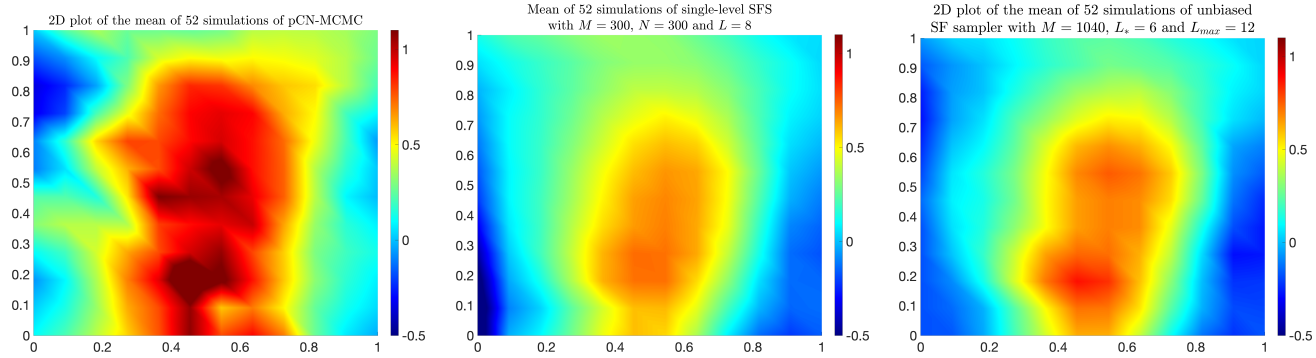
where $\{Z_i\}_{i=1}^N$ are samples generated from $\pi_{\widetilde{X}^l_{k\Delta_l}, k\Delta_l}$ (by using any sampling method, e.g. MCMC). Moreover, for any $(x, \tilde{x}, t) \in \mathbb{R}^d \times \mathbb{R}^d \times [0, 1)$ we can clearly write

$$b(x,t) = \frac{1}{\sqrt{1-t}} \frac{\mathbb{E}_{\pi_{\tilde{x},t}}\left[Z\frac{f(x+\sqrt{1-t}Z)}{f(\tilde{x}+\sqrt{1-t}Z)}\right]}{\mathbb{E}_{\pi_{\tilde{x},t}}\left[\frac{f(x+\sqrt{1-t}Z)}{f(\tilde{x}+\sqrt{1-t}Z)}\right]}. \tag{18}$$

This enables us to produce a coupled pair of (approximate) samples from a pair of Euler discretizations of (1) as explained in Algorithm 3. With the new identity in (18), we hope that computing the ratio $f(x+\sqrt{1-t}Z)/f(\tilde{x}+\sqrt{1-t}Z)$, $Z \sim \pi_{\tilde{x},t}$, on the log-scale will overcome the fixed point arithmetic issue discussed in the introduction of subsection 4.2. Basically, we estimate the drift in (2) using the identity in (17) at level zero and at the coarser level in the coupling step while using (14–15) at the finer level.



(a) The reference log-permeability field in 3D (b) Single-level SFS results in 3D (c) Unbiased SFS results in 3D



(d) The reference log-permeability field in 2D (e) Single-level SFS results in 2D (f) Unbiased SFS results in 2D

Figure 3: Results of running the single-level SFS in Algorithm 1 with $\hat{b}$ computed as in (17) and as in Algorithm 4. It took around 33.5 hours to run Algorithm 4, and for the same accuracy (RMSE for both is ~ 3.9), Algorithm 1 took around 168 hours.

4.2.3 Simulation Results

We set the number of nodes in the mesh to $\hat{n} = 12^2$ and consider observations generated from the solution to the PDE on a finer mesh at 81 random nodes with σ_y chosen such that a prescribed signal-to-noise ratio, defined as $\max\{p\}/\sigma_y$, is equal to 100. The true log-premeability field u_{truth} used to generate the data is defined by the superposition of two Gaussians with covariances $0.05I_2$ and $0.07I_2$, centered at $(0.5, 0.5)$ and $(0.5, 0.95)$, with weights $\{0.6, 0.2\}$ respectively. The external force field F is defined by a superposition of four weighted two-dimensional Gaussian bumps with covariance $0.03I_2$, centered at (0.3, 0.5), (0.4, 0.3), (0.6, 0.9), and (0.7, 0.1), and with weights $\{2, -3, 1, -4\}$, respectively. In the covariance matrix of the prior, we take $\sigma = 1.5$ and $\alpha = 0.9$. We ran 52 independent simulations of Algorithm 1 with $\hat{b}$ computed as in (17) with $L = 8$, $N = 300$ and $M = 300$. We also ran 52 independent simulations of Algorithm 4 with $L_* = 6$, $L_{\max} = 12$ and $M = 1040$. The probability mass functions $\mathbb{P}_R(L)$ and $\mathbb{P}_{P|L}(P|L)$ are the same as in subsubsection 4.1.1. These parameters are chosen such that both algorithms give almost the same MSE. We remark that the reference log-permeability field was computed by taking the mean of 52 simulations of a preconditioned Crank-Nicolson (pCN) MCMC [4] with 10^7 samples a burn-in period of 10^3. We also used the pCN-MCMC to sample from the density $\pi_{x,t}$ defined in (16). Both algorithms were run on a workstation with 52 cores. Whilst aiming for similar precision, the computing cost of Algorithm 1 was about a week, whereas for Algorithms 3-4 the cost was a day

and a half. The results are shown in Figure 3 and in this example we managed to obtain with Algorithms 3-4 more accurate estimates at a fraction of the computational cost.

Acknowledgements

AJ & HR were supported by KAUST baseline funding.

A Theoretical Results for the Proof of Theorem 3.1

The following Section contains a collection of Lemmata used to prove Theorem 3.1. Recall that C is a generic finite constant with value that could change upon each appearance, but will not depend upon (l, N) and as t is bounded by 1, not on t either. Recall, we use the notation that for a differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$, $\nabla_k \psi(x) = (\partial\psi/\partial x_k)(x)$, $k \in \{1, \dots, d\}$.

Lemma A.1. *Assume (A1). Then, there exists $C < \infty$ so that for any $(x, y, N, s) \in \mathbb{R}^{2d} \times \mathbb{N} \times [0, 1]$ we have, almost surely, that*

$$\left\|\hat{b}(x, s) - \hat{b}(y, s)\right\|_2 \leq C \left\|x - y\right\|_2.$$

Proof. It suffices to prove that, almost surely, for each $j \in \{1, \dots, d\}$:

$$\left|\left\{\hat{b}(x, s) - \hat{b}(y, s)\right\}_j\right| \leq C \left\|x - y\right\|_2. \tag{19}$$

We have the simple decomposition

$$\frac{A(x)}{B(x)} - \frac{A(y)}{B(y)} = \tfrac{A(x)}{B(x)B(y)} \cdot \big(B(y) - B(x)\big) + \tfrac{1}{B(y)} \cdot \big(A(x) - A(y)\big),$$

where we have set

$$A(x) := \tfrac{1}{N} \sum_{i=1}^{N} \nabla_j f(x + \sqrt{1 - s} Z^i), \quad B(x) := \frac{1}{N} \sum_{i=1}^{N} f(x + \sqrt{1 - s} Z^i).$$

By (A1) a. the terms

$$\frac{A(x)}{B(x)B(y)}, \quad \frac{1}{B(y)},$$

are uniformly upper-bounded by a deterministic constant, so we have

$$\left|\left\{\hat{b}(x, s) - \hat{b}(y, s)\right\}_j\right| \leq C \left(|B(y) - B(x)| + |A(y) - A(x)| \right).$$

Applying the triangular inequality multiple times and (A1) b. allows us to deduce the bound (19) and thus the proof is concluded. □

Lemma A.2. *Assume that (A1). Then, for any $p \in [1, \infty)$ there exists a $C < \infty$ such that for any $(l, N, t) \in \mathbb{N}_0 \times \mathbb{N} \times [0, 1]$*

$$\mathbb{E} \left\|\hat{b}(\widetilde{X}^l_{\tau^l_t}, \tau^l_t) - b(\widetilde{X}^l_{\tau^l_t}, \tau^l_t)\right\|_2^p \leq \frac{C}{N^{p/2}}.$$

Proof. We consider a single co-ordinate of the vector $\hat{b}(\widetilde{X}^l_{\tau^l_t}, \tau^l_t) - b(\widetilde{X}^l_{\tau^l_t}, \tau^l_t)$ and it suffices to bound

$$\mathbb{E}\left|\left\{\hat{b}(\widetilde{X}^l_{\tau^l_t}, \tau^l_t) - b(\widetilde{X}^l_{\tau^l_t}, \tau^l_t)\right\}_j\right|^p \tag{20}$$

for any $j \in \{1, \ldots, d\}$. We have the simple decomposition

$$\frac{A^N}{B^N} - \frac{A}{B} = \tfrac{A^N}{B \cdot B^N}(B - B^N) + \tfrac{1}{B}(A^N - A),$$

where we have defined

$$A^N := \tfrac{1}{N}\sum_{i=1}^N \nabla_j f(\widetilde{X}^l_{\tau^l_t} + \sqrt{1-\tau^l_t} Z^i), \quad B^N := \tfrac{1}{N}\sum_{i=1}^N f(\widetilde{X}^l_{\tau^l_t} + \sqrt{1-\tau^l_t} Z^i),$$
$$A := \mathbb{E}_\phi\left[\nabla_j f(\widetilde{X}^l_{\tau^l_t} + \sqrt{1-\tau^l_t} Z)\right], \quad B := \mathbb{E}_\phi\left[f(\widetilde{X}^l_{\tau^l_t} + \sqrt{1-\tau^l_t} Z)\right].$$

Thus, using (A1) and the C_p-inequality we have the following upper-bound for (20)

$$C\left(\mathbb{E}\,|B - B^N|^p + \mathbb{E}\,|A^N - A|^p\right).$$

As $\widetilde{X}^l_{\tau^l_t}$ is independent of $Z^1, \ldots, Z^N$ we can use standard results for iid random variables to deduce that (20) is upper bounded by $C/N^{p/2}$ and the proof is now completed. □

Lemma A.3. *Assume that (A1). Then, for any $p \in [1, \infty)$ there exists a $C < \infty$ such that for any $(l, N, t) \in \mathbb{N}_0 \times \mathbb{N} \times [0, 1]$*

$$\mathbb{E}\left\|\widetilde{X}^{l,N}_{\tau^l_t} - \widetilde{X}^l_{\tau^l_t}\right\|_2^p \leq \frac{C}{N^{p/2}}. \tag{21}$$

Proof. We have that

$$\widetilde{X}^{l,N}_{\tau^l_t} - \widetilde{X}^l_{\tau^l_t} = \int_0^{\tau^l_t} \left(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^l_{\tau^l_s}, \tau^l_s)\right) ds + \int_0^{\tau^l_t} \left(\hat{b}(\widetilde{X}^l_{\tau^l_s}, \tau^l_s) - b(\widetilde{X}^l_{\tau^l_s}, \tau^l_s)\right) ds.$$

Therefore, it easily follows that $\mathbb{E}\left\|\widetilde{X}^{l,N}_{\tau^l_t} - \widetilde{X}^l_{\tau^l_t}\right\|_2^p$ is upper bounded by

$$C\int_0^{\tau^l_t} \left\{\mathbb{E}\left\|\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^l_{\tau^l_s}, \tau^l_s)\right\|_2^p + \mathbb{E}\left\|\hat{b}(\widetilde{X}^l_{\tau^l_s}, \tau^l_s) - b(\widetilde{X}^l_{\tau^l_s}, \tau^l_s)\right\|_2^p\right\} ds.$$

For the first term in the integral one can use Lemma A.1 and for the second one can use Lemma A.2. Thus, one can deduce the following upper bound for (21)

$$C\left(\int_0^t \mathbb{E}\left\|\widetilde{X}^{l,N}_{\tau^l_s} - \widetilde{X}^l_{\tau^l_s}\right\|_2^p ds + \frac{1}{N^{p/2}}\right).$$

So, the proof is concluded by applying Grönwall's inequality. □

Lemma A.4. *Assume that (A1). Then, there exists a $C < +\infty$ such that for any $(l, N, s) \in \mathbb{N}_0 \times \mathbb{N} \times [0, 1]$ we have*

$$\mathbb{E}\left\|\left(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l,N}_{\tau^{l-1}_s}, \tau^{l-1}_s)\right) - \left(\hat{b}(\widetilde{X}^l_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^l_{\tau^{l-1}_s}, \tau^{l-1}_s)\right)\right\|_2^2 \leq \frac{C\Delta_l}{N}.$$

Proof. We will consider one co-ordinate of the vector

$$T := \big(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l,N}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(\hat{b}(\widetilde{X}^{l}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)$$

as the argument is essentially the same across all co-ordinates. We have that for any $j \in \{1, \dots, d\}$

$$T_j = \Big(\frac{A^{l,N}}{B^{l,N}} - \frac{C^{l,N}}{D^{l,N}}\Big) - \Big(\frac{A^{l}}{B^{l}} - \frac{C^{l}}{D^{l}}\Big),$$

where we have defined

$$\begin{aligned}
A^{l,N} &:= \tfrac{1}{N}\sum_{i=1}^N \nabla_j f(\widetilde{X}^{l,N}_{\tau^l_s} + \sqrt{1-\tau^l_s}Z^i); \quad B^{l,N} := \tfrac{1}{N}\sum_{i=1}^N f(\widetilde{X}^{l,N}_{\tau^l_s} + \sqrt{1-\tau^l_s}Z^i),\\
C^{l,N} &:= \tfrac{1}{N}\sum_{i=1}^N \nabla_j f(\widetilde{X}^{l,N}_{\tau^{l-1}_s} + \sqrt{1-\tau^{l-1}_s}Z^i), \quad D^{l,N} := \tfrac{1}{N}\sum_{i=1}^N f(\widetilde{X}^{l,N}_{\tau^{l-1}_s} + \sqrt{1-\tau^{l-1}_s}Z^i),\\
A^{l} &:= \tfrac{1}{N}\sum_{i=1}^N \nabla_j f(\widetilde{X}^{l}_{\tau^l_s} + \sqrt{1-\tau^l_s}Z^i), \quad B^{l} := \tfrac{1}{N}\sum_{i=1}^N f(\widetilde{X}^{l}_{\tau^l_s} + \sqrt{1-\tau^l_s}Z^i),\\
C^{l} &:= \tfrac{1}{N}\sum_{i=1}^N \nabla_j f(\widetilde{X}^{l}_{\tau^{l-1}_s} + \sqrt{1-\tau^{l-1}_s}Z^i), \quad D^{l} := \tfrac{1}{N}\sum_{i=1}^N f(\widetilde{X}^{l}_{\tau^{l-1}_s} + \sqrt{1-\tau^{l-1}_s}Z^i).
\end{aligned}$$

We use [13, Lemma C.5.] stating that for reals (a, b, c, d) and non-zero reals (A, B, C, D)

$$\begin{aligned}
&\big(\tfrac{a}{A} - \tfrac{b}{B}\big) - \big(\tfrac{c}{C} - \tfrac{d}{D}\big) =\\
&\quad \tfrac{1}{A}\big((a-b) - (c-d)\big) - \tfrac{b}{AB}\big((A-B) - (C-D)\big) + \tfrac{1}{AC}(C-A)(c-d)\\
&\qquad - \tfrac{1}{AB}(b-d)(C-D) + \tfrac{d}{CBD}(B-D)(C-D) + \tfrac{d}{ACB}(A-C)(C-D). \qquad (22)
\end{aligned}$$

One can consider all six terms, individually, by using the C_2-inequality. However, as the terms

$$\tfrac{1}{A}\big((a-b) - (c-d)\big), \quad \tfrac{b}{AB}\big((A-B) - (C-D)\big),$$

are similar, we treat only the former. In addition, as the last four terms on the right side of (22) can be dealt with using similar calculations, we only deal with one of them. To that end, we seek to bound the two terms:

$$\begin{aligned}
T_{j,1} &:= \mathbb{E}\Big[\Big(\tfrac{1}{(B^{l,N})^2}\cdot\big((A^{l,N} - C^{l,N}) - (A^l - C^l)\big)^2\Big],\\
T_{j,2} &:= \mathbb{E}\Big[\tfrac{1}{(B^l)^2(B^{l,N})^2}\cdot(B^l - B^{l,N})^2\cdot(A^l - C^l)^2\Big].
\end{aligned}$$

For $T_{j,1}$ we easily obtain the upper-bound

$$C\,\mathbb{E}\Big|\big(\nabla_j f(Y^{l,N}_{\tau^l_s}) - \nabla_j f(Y^{l,N}_{\tau^{l-1}_s})\big) - \big(\nabla_j f(Y^{l}_{\tau^l_s}) - \nabla_j f(Y^{l}_{\tau^{l-1}_s})\big)\Big|^2$$

where we have set

$$\begin{aligned}
Y^{l,N}_{\tau^l_s} &:= \widetilde{X}^{l,N}_{\tau^l_s} + \sqrt{1-\tau^l_s}Z^1, \quad Y^{l}_{\tau^l_s} := \widetilde{X}^{l}_{\tau^l_s} + \sqrt{1-\tau^l_s}Z^1,\\
Y^{l,N}_{\tau^{l-1}_s} &:= \widetilde{X}^{l,N}_{\tau^{l-1}_s} + \sqrt{1-\tau^{l-1}_s}Z^1, \quad Y^{l}_{\tau^{l-1}_s} := \widetilde{X}^{l}_{\tau^{l-1}_s} + \sqrt{1-\tau^{l-1}_s}Z^1.
\end{aligned}$$

Note now that

$$\big(\nabla_j f(Y^{l,N}_{\tau^l_s}) - \nabla_j f(Y^{l,N}_{\tau^{l-1}_s})\big) - \big(\nabla_j f(Y^{l}_{\tau^l_s}) - \nabla_j f(Y^{l}_{\tau^{l-1}_s})\big) = \bar{T}_{j,1}(1) + \bar{T}_{j,1}(2),$$

for the terms

$$\bar{T}_{j,1}(1) := \sum_{k=1}^{d} \int_0^1 \nabla_k \nabla_j f(Y^{l,N}_{\tau^{l-1}_s} + \lambda(Y^{l,N}_{\tau^l_s} - Y^{l,N}_{\tau^{l-1}_s})) \cdot \big\{(Y^{l,N}_{\tau^l_s} - Y^{l,N}_{\tau^{l-1}_s}) - (Y^{l}_{\tau^l_s} - Y^{l}_{\tau^{l-1}_s})\big\}_k \, d\lambda,$$

$$\bar{T}_{j,1}(2) := \sum_{k=1}^{d} \int_0^1 \Big(\nabla_k \nabla_j f(Y^{l,N}_{\tau^{l-1}_s} + \lambda(Y^{l,N}_{\tau^l_s} - Y^{l,N}_{\tau^{l-1}_s})) - \nabla_k \nabla_j f(Y^{l}_{\tau^{l-1}_s} + \lambda(Y^{l}_{\tau^l_s} - Y^{l}_{\tau^{l-1}_s}))\Big) \cdot \{Y^{l}_{\tau^l_s} - Y^{l}_{\tau^{l-1}_s}\}_k \, d\lambda$$

Then to obtain the required bound for $T_{j,1}$, it suffices to bound the second moments of $\bar{T}_{j,1}(1)$ and $\bar{T}_{j,1}(2)$ respectively. For $\bar{T}_{j,1}(1)$ we have

$$\mathbb{E}\,\big|\,\bar{T}_{j,1}(1)\,\big|^2 \le C\,\mathbb{E}\,\big\|(\widetilde{X}^{l,N}_{\tau^l_s} - \widetilde{X}^{l,N}_{\tau^{l-1}_s}) - (\widetilde{X}^{l}_{\tau^l_s} - \widetilde{X}^{l}_{\tau^{l-1}_s})\big\|_2^2.$$

Also, since

$$\begin{aligned}(\widetilde{X}^{l,N}_{\tau^l_s} - \widetilde{X}^{l,N}_{\tau^{l-1}_s}) - (\widetilde{X}^{l}_{\tau^l_s} - \widetilde{X}^{l}_{\tau^{l-1}_s}) \\ = \int_{\tau^{l-1}_s}^{\tau^l_s} \Big(\big(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_u}, \tau^l_u) - \hat{b}(\widetilde{X}^{l}_{\tau^l_u}, \tau^l_u)\big) + \big(\hat{b}(\widetilde{X}^{l}_{\tau^l_u}, \tau^l_u) - b(\widetilde{X}^{l}_{\tau^l_u}, \tau^l_u)\big)\Big) du \end{aligned} \tag{23}$$

one can follow similar arguments that were used to deduce (11) to obtain

$$\mathbb{E}\,\big|\,\bar{T}_{j,1}(1)\,\big|^2 \le \frac{C\Delta_l^2}{N}.$$

Using (A1) it follows that

$$\mathbb{E}\,\big|\,\bar{T}_{j,1}(2)\,\big|^2 \le C\,\mathbb{E}\,\Bigg[\Big(\big\|Y^{l,N}_{\tau^{l-1}_s} - Y^{l}_{\tau^{l-1}_s}\big\|_2^2 + \big\|(\widetilde{X}^{l,N}_{\tau^l_s} - \widetilde{X}^{l,N}_{\tau^{l-1}_s}) - (\widetilde{X}^{l}_{\tau^l_s} - \widetilde{X}^{l}_{\tau^{l-1}_s})\big\|_2^2\Big) \times \big\|Y^{l}_{\tau^l_s} - Y^{l}_{\tau^{l-1}_s}\big\|_2^2\Bigg].$$

Using Cauchy-Schwarz, we obtain

$$\begin{aligned}&\mathbb{E}\,\big|\,\bar{T}_{j,1}(2)\,\big|^2 \\ &\le C\,\mathbb{E}\,\big[\,\|Y^{l}_{\tau^l_s} - Y^{l}_{\tau^{l-1}_s}\|_2^4\,\big]^{1/2} \\ &\qquad \times \Big(\mathbb{E}\,\big[\,\|Y^{l,N}_{\tau^{l-1}_s} - Y^{l}_{\tau^{l-1}_s}\|_2^4\,\big]^{1/2} + \mathbb{E}\,\big[\,\big\|(\widetilde{X}^{l,N}_{\tau^l_s} - \widetilde{X}^{l,N}_{\tau^{l-1}_s}) - (\widetilde{X}^{l}_{\tau^l_s} - \widetilde{X}^{l}_{\tau^{l-1}_s})\big\|_2^4\,\big]^{1/2}\Big) \\ &= C\,\mathbb{E}\,\big[\,\big\|\widetilde{X}^{l}_{\tau^l_s} - \widetilde{X}^{l}_{\tau^{l-1}_s} + \{\sqrt{1-\tau^l_s} - \sqrt{1-\tau^{l-1}_s}\}Z^1\big\|^4\,\big]^{1/2} \\ &\qquad \times \Big(\mathbb{E}\,\big[\,\big\|\widetilde{X}^{l,N}_{\tau^{l-1}_s} - \widetilde{X}^{l}_{\tau^{l-1}_s}\big\|_2^4\,\big]^{1/2} + \mathbb{E}\,\big[\,\big\|(\widetilde{X}^{l,N}_{\tau^l_s} - \widetilde{X}^{l,N}_{\tau^{l-1}_s}) - (\widetilde{X}^{l}_{\tau^l_s} - \widetilde{X}^{l}_{\tau^{l-1}_s})\big\|_2^4\,\big]^{1/2}\Big).\end{aligned}$$

For the first factor term in the above upper bound, using standard results on Euler discretizations and Gaussian random variables, we have

$$\mathbb{E}\left[\left\|(\widetilde{X}^l_{\tau^l_s}-\widetilde{X}^l_{\tau^{l-1}_s})+(\sqrt{1-\tau^l_s}-\sqrt{1-\tau^{l-1}_s})Z^1\right\|_2^4\right]^{1/2}\leq C\Delta_l.$$

Then using Lemma A.3 and the above arguments one obtains

$$\mathbb{E}\left[\bar{T}_{j,1}(2)^2\right]\leq\frac{C\Delta_l}{N}.$$

Thus we can conclude that

$$T_{j,1}\leq\frac{C\Delta_l}{N}.$$

For $T_{j,2}$, using (A1) and Cauchy-Schwarz it follows that

$$\begin{aligned}T_{j,2}\leq C\,\mathbb{E}\Big[\Big(f(\widetilde{X}^l_{\tau^l_s}+\sqrt{1-\tau^l_s}Z^1)-f(\widetilde{X}^{l,N}_{\tau^l_s}+\sqrt{1-\tau^l_s}Z^1)\Big)^4\Big]^{1/2}\\ \times\,\mathbb{E}\Big[\Big(\nabla_j f(\widetilde{X}^l_{\tau^l_s}+\sqrt{1-\tau^l_s}Z^1)-\nabla_j f(\widetilde{X}^l_{\tau^{l-1}_s}+\sqrt{1-\tau^{l-1}_s}Z^1)\Big)^4\Big]^{1/2}.\end{aligned}$$

For the first factor one can use (A1) b. and Lemma A.3, and for the second one can use (A1) b. and standard properties of Euler-discretizations to give

$$T_{j,2}\leq\frac{C\Delta_l}{N}.$$

This completes the proof. □

Lemma A.5. *Assume that (A1). Then, there exists a $C<+\infty$ such that for any $(l,N,s)\in\mathbb{N}^2\times[0,1]$ we have*

$$\begin{aligned}\mathbb{E}\left\|\left(\hat{b}(\widetilde{X}^{l,N}_{\tau^{l-1}_s},\tau^{l-1}_s)-\hat{b}(\widetilde{X}^{l-1,N}_{\tau^{l-1}_s},\tau^{l-1}_s)\right)-\left(\hat{b}(\widetilde{X}^l_{\tau^{l-1}_s},\tau^{l-1}_s)-\hat{b}(\widetilde{X}^{l-1}_{\tau^{l-1}_s},\tau^{l-1}_s)\right)\right\|_2^2\leq\\ C\left(\frac{\Delta_l^2}{N}+\mathbb{E}\left\|\left(\widetilde{X}^{l,N}_{\tau^l_s}-\widetilde{X}^{l-1,N}_{\tau^{l-1}_s}\right)-\left(\widetilde{X}^l_{\tau^l_s}-\widetilde{X}^{l-1}_{\tau^{l-1}_s}\right)\right\|_2^2\right).\end{aligned}$$

Proof. The proof is much the same as that of Lemma A.4. The only difference is that the term Δ_l^2/N occurs as one considers the processes at the same time instance and that one will obtain an additive term in the upper-bound of the type:

$$\mathbb{E}\left\|\left(\widetilde{X}^{l,N}_{\tau^{l-1}_s}-\widetilde{X}^{l-1,N}_{\tau^{l-1}_s}\right)-\left(\widetilde{X}^l_{\tau^{l-1}_s}-\widetilde{X}^{l-1}_{\tau^{l-1}_s}\right)\right\|_2^2.$$

This latter term is upper-bounded by

$$C\left(\mathbb{E}\left\|\left(\widetilde{X}^{l,N}_{\tau^l_s}-\widetilde{X}^{l,N}_{\tau^{l-1}_s}\right)-\left(\widetilde{X}^l_{\tau^l_s}-\widetilde{X}^l_{\tau^{l-1}_s}\right)\right\|_2^2+\mathbb{E}\left\|\left(\widetilde{X}^{l,N}_{\tau^l_s}-\widetilde{X}^{l-1,N}_{\tau^{l-1}_s}\right)-\left(\widetilde{X}^l_{\tau^l_s}-\widetilde{X}^{l-1}_{\tau^{l-1}_s}\right)\right\|_2^2\right).$$

The first-term above is easily proved to be $\mathcal{O}(\Delta_l^2/N)$ (see (23) and the subsequent argument) and this concludes the proof. □

Lemma A.6. *Assume (A1). Then there exists a $C < +\infty$ such that for any $(l, N, t) \in \mathbb{N}^2 \times [0, 1]$ we have*

$$\mathbb{E}\Big\| \int_0^t \Big(\big(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l-1,N}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(\hat{b}(\widetilde{X}^{l}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)\Big) ds \Big\|_2^2 \leq$$

$$C\Big(\frac{\Delta_l}{N} + \int_0^t \mathbb{E}\Big\| \big(\widetilde{X}^{l,N}_{\tau^l_s} - \widetilde{X}^{l-1,N}_{\tau^{l-1}_s}\big) - \big(\widetilde{X}^{l}_{\tau^l_s} - \widetilde{X}^{l-1}_{\tau^{l-1}_s}\big)\Big\|_2^2 ds \Big).$$

Proof. It is simple to establish that

$$\mathbb{E}\Big\| \int_0^t \Big(\big(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l-1,N}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(\hat{b}(\widetilde{X}^{l}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)\Big) ds \Big\|_2^2 \leq$$

$$C \int_0^t \mathbb{E}\Big\| \big(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l-1,N}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(\hat{b}(\widetilde{X}^{l}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)\Big\|_2^2 ds$$

so we focus on the term inside the integrand. We have

$$\mathbb{E}\Big\| \big(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l-1,N}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(\hat{b}(\widetilde{X}^{l}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)\Big\|_2^2 \leq$$

$$C\Bigg(\mathbb{E}\Big\| \big(\hat{b}(\widetilde{X}^{l,N}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l,N}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(\hat{b}(\widetilde{X}^{l}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)\Big\|_2^2 +$$

$$\mathbb{E}\Big\| \big(\hat{b}(\widetilde{X}^{l,N}_{\tau^{l-1}_s}, \tau^{l-1}_s) - \hat{b}(\widetilde{X}^{l-1,N}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(\hat{b}(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s) - \hat{b}(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)\Big\|_2^2 \Bigg).$$

So, the proof is concluded by applying Lemmata A.4-A.5. □

Lemma A.7. *Assume (A1). Then there exists a $C < +\infty$ such that for any $(l, N, s) \in \mathbb{N}^2 \times [0, 1]$ we have*

$$\mathbb{E}\Big\| \big(\hat{b}(\widetilde{X}^{l}_{\tau^l_s}, \tau^l_s) - \hat{b}(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(b(\widetilde{X}^{l}_{\tau^l_s}, \tau^l_s) - b(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)\Big\|_2^2 \leq \frac{C\Delta_l}{N}.$$

Proof. This can be proved using the same identity (equation (22)) as Lemma A.4. The subsequent calculations are much simpler than the proof of Lemma A.4 and are hence omitted. □

Lemma A.8. *Assume (A1). Then there exists a $C < +\infty$ such that for any $(l, N, t) \in \mathbb{N}^2 \times [0, 1]$ we have*

$$\mathbb{E}\Big\| \big(\hat{b}(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s) - \hat{b}(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(b(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s) - b(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big)\Big\|_2^2 \leq \frac{C\Delta_l^2}{N}.$$

Proof. We will consider one co-ordinate of the vector

$$T := \big(\hat{b}(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s) - \hat{b}(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big) - \big(b(\widetilde{X}^{l}_{\tau^{l-1}_s}, \tau^{l-1}_s) - b(\widetilde{X}^{l-1}_{\tau^{l-1}_s}, \tau^{l-1}_s)\big),$$

as the argument is essentially the same across all co-ordinates. We have that for any $j \in \{1, \ldots, d\}$

$$T_j = \Big(\frac{C^l}{D^l} - \frac{E^l}{F^l}\Big) - \Big(\frac{C}{D} - \frac{E}{F}\Big),$$

where we have defined

$$C^l := \frac{1}{N}\sum_{i=1}^N \nabla_j f(\widetilde{X}^l_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z^i); \quad D^l := \frac{1}{N}\sum_{i=1}^N f(\widetilde{X}^l_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z^i),$$

$$E^l := \frac{1}{N}\sum_{i=1}^N \nabla_j f(\widetilde{X}^{l-1}_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z^i); \quad F^l := \frac{1}{N}\sum_{i=1}^N f(\widetilde{X}^{l-1}_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z^i),$$

$$C := \mathbb{E}_\phi\left[\nabla_j f(\widetilde{X}^l_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z)\right], \quad D := \mathbb{E}_\phi\left[f(\widetilde{X}^l_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z)\right],$$

$$E := \mathbb{E}_\phi\left[\nabla_j f(\widetilde{X}^{l-1}_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z)\right], \quad F := \mathbb{E}_\phi\left[f(\widetilde{X}^{l-1}_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z)\right].$$

Again, we use (22) and just give a proof for two terms

$$T_{j,1} := \mathbb{E}\left[\frac{1}{(D^l)^2}\big((C^l - E^l) - (C-E)\big)^2\right],$$
$$T_{j,2} := \mathbb{E}\left[\frac{1}{(D^l)^2 D^2}(D - D^l)^2(C-E)^2\right].$$

For $T_{j,1}$ applying (A1) a. and using the fact that $(\widetilde{X}^l_{\tau_s^{l-1}}, \widetilde{X}^{l-1}_{\tau_s^{l-1}})$ are independent of the iid $Z^1, \ldots, Z^N$, we have

$$T_{j,1} \leq \frac{C}{N}\mathbb{E}\left|\nabla_j f(\widetilde{X}^l_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z^1) - \nabla_j f(\widetilde{X}^{l-1}_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z^1)\right|^2.$$

Then using (A1) b. along with standard results for strong errors of diffusions we have

$$T_{j,1} \leq \frac{C\Delta_l^2}{N}.$$

For $T_{j,2}$ applying (A1) a. and the Cauchy-Schwarz inequality

$$T_{j,2} \leq C\,\mathbb{E}\Big[\Big(\mathbb{E}_\phi\left[f(\widetilde{X}^l_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z)\right] - \frac{1}{N}\sum_{i=1}^N f(\widetilde{X}^l_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z^i)\Big)^4\Big]^{1/2}$$
$$\times\,\mathbb{E}\Big[\Big(\mathbb{E}_\phi\left[\nabla_j f(\widetilde{X}^l_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z) - \nabla_j f(\widetilde{X}^{l-1}_{\tau_s^{l-1}} + \sqrt{1-\tau_s^{l-1}}Z)\right]\Big)^4\Big]^{1/2}.$$

For the first term on the right-hand-side one can use standard results for iid random variables and for the second (A1) a. along with standard results for strong errors of diffusions to yield

$$T_{j,2} \leq \frac{C\Delta_l^2}{N}.$$

From here, one can complete the proof fairly easily. □

Lemma A.9. *Assume (A1). Then there exists a $C < +\infty$ such that for any $(l, N, t) \in \mathbb{N}^2 \times [0,1]$ we have*

$$\mathbb{E}\left\|\int_0^t \Big(\big(\hat{b}(\widetilde{X}^l_{\tau_s^l}, \tau_s^l) - \hat{b}(\widetilde{X}^{l-1}_{\tau_s^{l-1}}, \tau_s^{l-1})\big) - \big(b(\widetilde{X}^l_{\tau_s^l}, \tau_s^l) - b(\widetilde{X}^{l-1}_{\tau_s^{l-1}}, \tau_s^{l-1})\big)\Big)ds\right\|_2^2 \leq \frac{C\Delta_l}{N}.$$

Proof. The proof is essentially the same as that for Lemma A.6, except one must use Lemmata A.7-A.8 instead of Lemmata A.4-A.5; therefore the proof is omitted. □

References

[1] Bernton, E., Heng, J., Doucet, A., & Jacob, P.E. (2019) Schrodinger Bridge Samplers, arXiv preprint, arXiv:2106.01357.

[2] Blanchet, J., Glynn, P., & Pei, Y. (2019). Unbiased multilevel Monte Carlo: Stochastic Optimization, Steady-state Simulation, Quantiles, and Other Applications. arXiv preprint.

[3] De Bortoli, V., Thornton, J., Heng, J., & Doucet, A. (2021) Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling, In *Proc. NeurIPS 2021.*

[4] Cotter, S. L., Roberts, G. O., Stuart, A. M. & White, D. (2013). MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.* **28**(3):424–446.

[5] Dai Pra, P. (1991). A stochastic control approach to reciprocal diffusion processes. *Appl. Math. Optim.*, **23**(1), 313–329.

[6] Ermak, D. L. (1975). A computer simulation of charged particles in solution. I. Technique and Equilibrium Properties. *J. Chem. Phys.*, **62**, 4189–4196.

[7] Föllmer, H. (1988). Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV-XVII, 1986–87*, pages 101–203. Springer.

[8] Heng, J., Houssineau, J. & Jasra, A. (2021). On unbiased score estimation for partially observed diffusions. arXiv preprint.

[9] Heng, J., Jasra, A., Law, K. J. H., & Tarakanov, A. (2021). On unbiased estimation for discretized models. arXiv preprint.

[10] Huang, J., Jiao, Y., Kang, L., Liao, X., Liu, J. & Liu, Y. (2021). Schrödinger-Föllmer sampler: Sampling without ergodicity, arXiv preprint arXiv: 2106.10880.

[11] Jamison, B. (1975). The Markov processes of Schrödinger, *Z. Wahrsch. Vern. Gebiete* 32, 323–331

[12] Jacob, P., O'Leary, J. & Atachade, Y. (2020). Unbiased Markov chain Monte Carlo with couplings (with discussion). *J. R. Statist. Soc. Ser. B*, **82**, 543–600.

[13] Jasra, A., Kamatani, K., Law, K. J. H., & Zhou, Y. (2017). Multilevel particle filter. *SIAM J. Numer. Anal.*, **55**, 3068–3096.

[14] Jasra, A., Law, K. J. H, & Yu, F. (2022). Unbiased filtering of diffusions. *Adv. Appl. Probab.* (to appear).

[15] Jelic, V. & and Marsiglio, F. (2012) The double-well potential in quantum mechanics: a simple, numerically exact formulation. *Eur. J. Phys.* 33 1651.

[16] Jiao, Y., Kang, L. , Liu, Y. & Zhou, Y. (2021). Convergence analysis of the Schrödinger-Föllmer sampler without convexity. arXiv preprint.

[17] Kloeden, P. E., & Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, Heidelberg.

[18] McLeish, D. (2011). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Meth. Appl.*, **17**, 301–315.

[19] Parisi, G. (1981). Correlation functions and computer simulations. *Nuclear Phys. B*, **180**, 378–384.

[20] Rhee, C. H. & Glynn, P. (2015). Unbiased estimation with square root convergence for SDE models. *Op. Res.*, **63**, 1026–1043.

[21] Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer: New York.

[22] Schrödinger, E. (1931). Über die Umkehrung der Naturgesetze. *Sitzung ber Preuss. Akad. Wissen., Berlin Phys. Math.*, 144.

[23] Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica*, **19**, 451–559.