Multilevel Sequential Monte Carlo Samplers

Alexandros Beskos[1], Ajay Jasra[2], Kody Law[3], Raul Tempone[4] & Yan Zhou[2]

[1]*Department of Statistical Science, University College London, London, WC1E 6BT, UK,*
[2]*Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG,*
[3]*Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, 37934 TN, USA*
& [4]*Center for Uncertainty Quantification in Computational Science & Engineering, King Abdullah University of Science and Technology, Thuwal, 23955-6900, KSA.*

Abstract

In this article we consider the approximation of expectations w.r.t. probability distributions associated to the solution of partial differential equations (PDEs); this scenario appears routinely in Bayesian inverse problems. In practice, one often has to solve the associated PDE numerically, using, for instance finite element methods which depends on the step-size level h_L. In addition, the expectation cannot be computed analytically and one often resorts to Monte Carlo methods. In the context of this problem, it is known that the introduction of the multilevel Monte Carlo (MLMC) method can reduce the amount of computational effort to estimate expectations, for a given level of error. This is achieved via a telescoping identity associated to a Monte Carlo approximation of a sequence of probability distributions with discretisation levels $\infty > h_0 > h_1 \cdots > h_L$. In many practical problems of interest, one cannot achieve an i.i.d. sampling of the associated sequence of probability distributions. A sequential Monte Carlo (SMC) version of the MLMC method is introduced to deal with this problem. It is shown that under appropriate assumptions, the attractive property of a reduction of the amount of computational effort to estimate expectations, for a given level of error, can be maintained within the SMC context, that is, relative to exact sampling and Monte Carlo for the distribution at the finest level h_L. The approach is numerically illustrated on a Bayesian inverse problem.

Keywords: Multilevel Monte Carlo, Sequential Monte Carlo, Bayesian Inverse Problems.
AMS subject classification: 65C30, 65Y20.

1. Introduction

Consider a sequence of probability measures $\{\eta_l\}_{l\geq 0}$ on a common measurable space $(E, \mathcal{E})$; we assume that the probabilities have common dominating finite-measure du and write the densities w.r.t. du as $\eta_l = \eta_l(u)$. In particular, for some known $\gamma_l : E \to \mathbb{R}_+$, we let

$$\eta_l(u) = \frac{\gamma_l(u)}{Z_l} \tag{1}$$

where the normalizing constant $Z_l = \int_E \gamma_l(u)du$ may be unknown. The context of interest is when the sequence of densities is associated to an 'accuracy' parameter h_l, with $h_l \to 0$ as $l \to \infty$ with $\infty > h_0 > h_1 > \cdots > h_\infty = 0$. This set-up is relevant to the context of discretised numerical approximations of continuum fields, as we will explain below. The objective is to compute:

$$\mathbb{E}_{\eta_\infty}[g(U)] := \int_E g(u)\eta_\infty(u)du$$

for potentially many measurable $\eta_\infty-$integrable functions $g : E \to \mathbb{R}$. In practice one cannot treat $h_l = 0$ and must consider these distributions with $h_l > 0$.

Problems involving numerical approximations of continuum fields are discretized before being solved numerically. Finer-resolution solutions are more expensive to compute than coarser ones. Such discretizations naturally give rise to hierarchies of resolutions via the use of nested meshes. Successive solution at refined meshes can be utilized to mitigate the number of necessary solves for the finest resolutions. For Monte Carlo methods, as in the context above, a telescoping sum of associated differences at successive refinement levels can be utilized. As we will now explain, this idea can be leveraged so one can obtain a method with smaller computational effort to reach a pre-determined error than applying a *standard Monte Carlo method immediately at the finest resolution* [14].

Multi Level Monte Carlo (MLMC) [14] (see also [15]) methods are such that one typically sets an error threshold for a target expectation, and then sets out to attain an estimator with the prescribed error utilizing an optimal allocation of Monte Carlo resources. Within the context of [14, 16], the continuum problem is a stochastic differential equation (SDE) or PDE with random coefficients, and the target quantity is an expectation of a functional, say $g : E \to \mathbb{R}$, of the parameter of interest $U \in E$, over an ideal measure $U \sim \eta_\infty$ that avoids discretisation. The levels are a hierarchy of refined approximations of the function-space, specified in terms of a small resolution parameter say h_l, for $0 \leq l \leq L$, thus giving rise to a corresponding sequence of approximate laws η_l. The method uses the telescopic sum

$$\mathbb{E}_{\eta_L}[g(U)] = \mathbb{E}_{\eta_0}[g(U)] + \sum_{l=1}^{L} \{\mathbb{E}_{\eta_l}[g(U)] - \mathbb{E}_{\eta_{l-1}}[g(U)]\}$$

and proceeds by coupling the consecutive probability distributions η_{l-1}, η_l. Thus, the expectations are estimated via the standard unbiased Monte Carlo averages

$$Y_l^{N_l} = \sum_{i=1}^{N_l} \{g(U_l^i) - g(U_{l-1}^i)\} N_l^{-1}$$

where $\{U_{l-1}^i, U_l^i\}$ are i.i.d. samples, with marginal laws η_{l-1}, η_l, respectively, carefully constructed on a joint probability space. This is repeated independently for $0 \leq l \leq L$. The overall multilevel estimator will be

$$\hat{Y}_{L,\text{Multi}} = \sum_{l=0}^{L} Y_l^{N_l} \ , \tag{2}$$

under the convention that $g(U_{-1}^{(i)}) = 0$. A simple error analysis gives that the mean squared error (MSE) is

$$\mathbb{E}\{\hat{Y}_{L,\text{Multi}} - \mathbb{E}_{\eta_\infty}[g(U)]\}^2 = \underbrace{\mathbb{E}\{\hat{Y}_{L,\text{Multi}} - \mathbb{E}_{\eta_L}[g(U)]\}^2}_{\text{variance}} + \underbrace{\{\mathbb{E}_{\eta_L}[g(U)] - \mathbb{E}_{\eta_\infty}[g(U)]\}^2}_{\text{bias}} \ . \tag{3}$$

One can now optimally allocate $N_0, N_1, \ldots, N_L$ to minimize the variance term $\sum_{l=0}^{L} V_l/N_l$ for fixed computational cost $\sum_{l=0}^{L} C_l N_l$, where V_l is the variance of $[g(U_l^{(i)}) - g(U_{l-1}^{(i)})]$ and C_l the computational cost for its realisation. Using Lagrange multipliers for the above constrained optimisation, we get the optimal allocation of resources $N_l \propto \sqrt{V_l/C_l}$. In more detail, the typical chronology is that one targets an MSE, say $\mathcal{O}(\epsilon^2)$, then (i) given a characterisation of the bias as an order of h_l,

one determines $h_l = M^{-l}$, $l = 0, 1, \ldots, L$, for some integer $M > 1$, and chooses a horizon L such that the bias is $\mathcal{O}(\epsilon^2)$ and (ii) given a characterisation of V_l, C_l as some orders of h_l, one optimizes the required samples $N_0, \ldots N_L$ needed to give variance $\mathcal{O}(\epsilon^2)$. Thus, a specification of the bias, variance and computational costs as functions of h_l is needed.

As a prototypical example of the above setting [14], consider the case $U = X(T)$ with $X(T)$ being the terminal position of the solution X of a SDE and η_l is the distribution of $X(T)$ under the consideration of a numerical approximation with time-step $\Delta t_l = h_l$. The laws η_{l-1}, η_l can be coupled via use of the same driving Brownian path. Invoking the relevant error analysis for SDE models, one can obtain (for $U \sim \eta_\infty$, $U_l \sim \eta_l$, and defined on the common probability space):

(i) weak error $|\mathbb{E}[g(U_l) - g(U)]| = \mathcal{O}(h_l^\alpha)$, providing the bias $\mathcal{O}(h_l^\alpha)$,

(ii) strong error, $\mathbb{E}|g(U_l) - g(U)|^2 = \mathcal{O}(h_l^\beta)$, giving the variance $V_l = \mathcal{O}(h_l^\beta)$,

(iii) computational cost for a realisation of $g(U_l) - g(U_{l-1})$, $C_l = \mathcal{O}(h_l^{-\zeta})$,

for some constants α, β, ζ related to the details of the discretisation method. The standard Euler Marayuma method for solution of SDE gives the orders $\alpha = \beta = \zeta = 1$.

Assuming a general context, given such rates for bias, V_l and C_l, one proceeds as follows. Set $h_l = M^{-(l+k)}$, for some integer $M > 1$. Then, targeting an error tolerance of ϵ and letting $h_L^\alpha = M^{-L\alpha} = \mathcal{O}(\epsilon)$, one has $L = \log(\epsilon^{-1})/(\alpha \log(M)) + \mathcal{O}(1)$, as in [14]. Using the optimal allocation $N_l \propto \sqrt{V_l/C_l}$, one finds that $N_l \propto h_l^{(\beta+\zeta)/2}$. Taking under consideration a target error of size $\mathcal{O}(\epsilon)$, one sets $N_l \propto \epsilon^{-2} h_l^{(\beta+\zeta)/2} K_L$, with K_L chosen to control the total error for increasing L. Thus, for the resulted estimator in (2)-(3), we have:

$$\text{Variance} = \sum_{l=0}^{L} V_l N_l^{-1} = \epsilon^2 K_L^{-1} \sum_{l=0}^{L} h_l^{(\beta-\zeta)/2} \; ;$$

$$\text{Comp. Cost} = \sum_{l=0}^{L} N_l C_l = K_L^2 \epsilon^{-2} \; .$$

To have a variance of $\mathcal{O}(\epsilon^2)$, one sets $K_L = \sum_{l=0}^{L} h_l^{(\beta-\zeta)/2}$, so K_L may or may not depend on ϵ depending on whether this sum converges or not (recalling that $L = \mathcal{O}(|\log(\epsilon)|)$). In the case of Euler-Marayuma, for example, $\beta = \zeta$, $K_L = L$, and the cost is $\mathcal{O}(\log(\epsilon)^2 \epsilon^{-2})$, versus $\mathcal{O}(\epsilon^{-3})$ using a single level with mesh-size $h_L = \mathcal{O}(\epsilon)$. If $\beta > \zeta$, corresponding for instance to the Milstein method, then the cost is $\mathcal{O}(\epsilon^{-2})$. The latter is the cost of obtaining the given level of error for a scalar random variable, and is therefore optimal. The worst scenario is when $\beta < \zeta$. In this case it is sufficient to set $K_L = h_L^{(\beta-\zeta)/2}$ to make the variance $\mathcal{O}(\epsilon^2)$, and then the number of samples on the finest level is given by $N_L = h_L^{\beta-2\alpha}$ whereas the total algorithmic cost is $\mathcal{O}(\epsilon^{-(\zeta/\alpha+\delta)})$, where $\delta = 2 - \beta/\alpha \geq 0$. In this case, one can choose the largest value for the bias, $\alpha = \beta/2$, so that $N_L = 1$ and the total cost, $\mathcal{O}(\epsilon^{-\zeta/\alpha})$, is dominated by this single sample. See [14] for more details.

It is important to note that the realizations $U_l^{(i)}$, $U_{l-1}^{(i)}$ for a given increment must be coupled to obtain decaying variances V_l. In the case of an SDE driven by Brownian motion one can simply simulate the driving noise on level l and then upscale it to level $l-1$ by summing elements of the finer path [14]. For the case of a PDE forward model relying on uncertain input the scenario is quite similar [6].

The present work will focus on the case of an inverse problem with fixed-dimensional input. Indeed the difficulty arises here because we only know how to *evaluate* (up-to a constant) the

target density at any given level, and cannot directly obtain independent samples from it. The contribution of this work is then as follows. If one has a sequence of densities $\{\eta_l\}_{l\geq 0}$ for which one expects the application of MLMC to be beneficial, but cannot obtain exact samples from the couples η_{l-1}, η_l, then, we introduce a methodology where a version of the MLMC identity can be leveraged. Moreover, we show for our inverse problem and relative to the *standard (exact sampling) Monte Carlo method at the finest resolution*, there is a reduction in the amount of work, for the same level of error. There exist alternatives to solving such problems, for example one can review the recent works [16, 18] which use Markov chain Monte Carlo (MCMC) methods. In this article a more natural and powerful formulation is considered, related with the use of Sequential Monte Carlo approaches.

Sequential Monte Carlo (SMC) methods are amongst the most widely used computational techniques in statistics, engineering, physics, finance and many other disciplines. In particular SMC samplers [9] are designed to approximate a sequence $\{\eta_l\}_{l\geq 0}$ of probability distributions on a common space, whose densities are only known up-to a normalising constant. The method uses $N \geq 1$ samples (or particles) that are generated in parallel, and are propagated with importance sampling (often) via MCMC and resampling methods. Several convergence results, as N grows, have been proved (see e.g. [3, 7, 8, 12]). SMC samplers have also recently been proven to be stable in certain high-dimensional contexts [1]. Current state of the art for the analysis of SMC algorithms include the work of [3, 4, 7, 8, 12]. In this work, the method of SMC samplers is perfectly designed to approximate the sequence of distributions, but as we will see, implementing the standard telescoping identity of MLMC requires some ingenuity. In addition, in order to consider the benefit of using SMC, one must analyze the variance of the estimate; in such scenarios this is not a trivial extension of the convergence analysis previously mentioned. In particular, one must very precisely consider the auto-covariance of the SMC approximations and consider the rate of decrease of this quantity as the time-lag between SMC approximations increases. Such a precise analysis does not appear to exist in the literature. We note that our work, whilst presented in the context of PDEs, is not restricted to such scenarios and, indeed can be applied in almost any other similar context (that is, a sequence of distributions on a common space, with increasing computational costs associated to the evaluation of the densities which in some sense converge to a given density); however, the potential benefit of doing so, may not be obvious in general.

This article is structured as follows. In Section 2 the ML identity and SMC algorithm are given. In Section 3 our main complexity result is given under assumptions and their implications are discussed. In Section 4 we give a context where the assumptions of our theoretical results can be verified. In Section 5 our approach is numerically demonstrated on a Bayesian inverse problem. In Section 6 we consider extensions to our work. Section 3 and the Appendix provide the proofs of our main theorem.

2. Sequential Monte Carlo Methods

2.1. Notations

Let $(E, \mathcal{E})$ be a measurable space. The notation $\mathcal{B}_b(E)$ denotes the class of bounded and measurable real-valued functions. The supremum norm is written as $\|f\|_\infty = \sup_{u\in E} |f(u)|$ and $\mathcal{P}(E)$ is the set of probability measures on $(E, \mathcal{E})$. We will consider non-negative operators $K : E \times \mathcal{E} \to \mathbb{R}_+$ such that for each $u \in E$ the mapping $A \mapsto K(u, A)$ is a finite non-negative measure on $\mathcal{E}$ and for each $A \in \mathcal{E}$ the function $u \mapsto K(u, A)$ is measurable; the kernel K is Markovian if $K(u, dv)$ is a probability measure for every $u \in E$. For a finite measure μ on $(E, \mathcal{E})$, and a

real-valued, measurable $f : E \to \mathbb{R}$, we define the operations:

$$\mu K : A \mapsto \int K(u, A)\, \mu(du) \; ; \quad Kf : u \mapsto \int f(v)\, K(u, dv).$$

We also write $\mu(f) = \int f(u)\mu(du)$. In addition $\| \cdot \|_r$, $r \geq 1$, denotes the L_r-norm, where the expectation is w.r.t. the law of the appropriate simulated algorithm.

2.2. Algorithm

As described in Section 1, the context of interest is when a sequence of densities $\{\eta_l\}_{l\geq 0}$, as in (1), are associated to an 'accuracy' parameter h_l, with $h_l \to 0$ as $l \to \infty$, such that $\infty > h_0 > h_1 \cdots > h_\infty = 0$. Recall that $\eta_l(u) = \gamma_l(u)/Z_l$, where for any fixed and finite l, $\gamma_l : E \to \mathbb{R}_+$ can be evaluated and $Z_l = \int_E \gamma_l(u)du$ typically cannot be. η_∞ is too expensive or impossible to calculate. We choose a maximum level $L \geq 1$ and we will estimate

$$\mathbb{E}_{\eta_L}[g(U)] := \int_E g(u)\eta_L(u)du \; .$$

By the standard telescoping identity used in MLMC, one has

$$\begin{aligned}\mathbb{E}_{\eta_L}[g(U)] &= \mathbb{E}_{\eta_0}[g(U)] + \sum_{l=1}^{L} \Big\{ \mathbb{E}_{\eta_l}[g(U)] - \mathbb{E}_{\eta_{l-1}}[g(U)] \Big\} \\ &= \mathbb{E}_{\eta_0}[g(U)] + \sum_{l=1}^{L} \mathbb{E}_{\eta_{l-1}}\Big[\Big(\frac{\gamma_l(U) Z_{l-1}}{\gamma_{l-1}(U) Z_l} - 1 \Big) g(U) \Big] \; . \end{aligned} \tag{4}$$

We remark that our approach is simply transferring the problem of sampling from a sequence of suitably defined couplings to that of sampling from the original sequence of densities and replacing coupling with importance sampling. It should be noted that the importance sampling densities, that is the η_{l-1} on the R.H.S. of (4), are well chosen, in that one should be able to find a method which yields estimators whose computational cost to achieve a reasonable variance is not prohibitively high. As noted one assumes that one cannot sample from the $\{\eta_l\}_{l\geq 1}$. If we used importance sampling but with proposals for which one can sample exactly, then it is often the case that, for instance, the computational effort to control the variance of the estimate is exponential in the dimension of the problem [2]. However, for the methodology to be introduced below, it can be polynomial in the dimension of the problem; see [1].

Suppose now that one applies an SMC sampler [9] to obtain a collection of samples (particles) that sequentially approximate $\eta_0, \eta_1, \ldots, \eta_L$. We consider the case when one initializes the population of particles by sampling i.i.d. from η_0, then at every step resamples and applies a MCMC kernel to mutate the particles. We denote by $(U_0^{1:N_0}, \ldots, U_{L-1}^{1:N_{L-1}})$, with $+\infty > N_0 \geq N_1 \geq \cdots N_{L-1} \geq 1$, the samples after mutation; one resamples $U_l^{1:N_l}$ according to the weights $G_l(U_l^i) = (\gamma_{l+1}/\gamma_l)(U_l^i)$, for indices $l \in \{0, \ldots, L-1\}$. We will denote by $\{M_l\}_{1\leq l\leq L-1}$ the sequence of MCMC kernels used at stages $1, \ldots, L-1$, such that $\eta_l M_l = \eta_l$. The algorithm is summarised in Figure 1. For $\varphi : E \to \mathbb{R}$, $l \in \{1, \ldots, L\}$, we have the following estimator of $\mathbb{E}_{\eta_{l-1}}[\varphi(U)]$:

$$\eta_{l-1}^{N_{l-1}}(\varphi) = \frac{1}{N_{l-1}} \sum_{i=1}^{N_{l-1}} \varphi(U_{l-1}^i) \; .$$

We define

$$\eta_{l-1}^{N_{l-1}}(G_{l-1}M_l(du_l)) = \frac{1}{N_{l-1}}\sum_{i=1}^{N_{l-1}} G_{l-1}(U_{l-1}^i)M_l(U_{l-1}^i, du_l) \ .$$

The joint probability distribution for the SMC algorithm is

$$\prod_{i=1}^{N_0}\eta_0(du_0^i)\prod_{l=1}^{L-1}\prod_{i=1}^{N_l}\frac{\eta_{l-1}^{N_{l-1}}(G_{l-1}M_l(du_l^i))}{\eta_{l-1}^{N_{l-1}}(G_{l-1})} \ .$$

If one considers one more step in the above procedure, that would deliver samples $\{U_L^i\}_{i=1}^{N_L}$, a standard SMC sampler estimate of the quantity of interest in (4) is $\eta_L^N(g)$; the earlier samples are discarded. Within a multilevel context, a consistent SMC estimate of (4) is

$$\widehat{Y} = \eta_0^{N_0}(g) + \sum_{l=1}^{L}\left\{\frac{\eta_{l-1}^{N_{l-1}}(gG_{l-1})}{\eta_{l-1}^{N_{l-1}}(G_{l-1})} - \eta_{l-1}^{N_{l-1}}(g)\right\} , \tag{5}$$

and this will be proven to be superior than the standard one (i.i.d. sampling from η_L), under assumptions.

There are two important structural differences within the MLSMC context, compared to the standard ML implementation of [14], sketched in Section 1:

i) the $L+1$ terms in (5) are *not* unbiased estimates of the differences $\mathbb{E}_{\eta_l}[g(U)] - \mathbb{E}_{\eta_{l-1}}[g(U)]$, so the relevant MSE error decomposition here is:

$$\mathbb{E}\big[\{\widehat{Y} - \mathbb{E}_{\eta_\infty}[g(U)]\}^2\big] \leq 2\,\mathbb{E}\big[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2\big] + 2\,\{\mathbb{E}_{\eta_L}[g(U)] - \mathbb{E}_{\eta_\infty}[g(U)]\}^2 \ . \tag{6}$$

ii) the same $L+1$ estimates are *not* independent. Hence a substantially more complex error analysis will be required to characterise $\mathbb{E}[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2]$. In Section 3, we will obtain an expression for this discrepancy, which will be more involved than the standard $\sum_{l=0}^{L} V_l/N_l$, but will still allow for a relevant constrained optimisation to determine the optimal allocation of particle sizes N_l along the levels.

Given an appropriate classification of both terms on the R.H.S. of (6) as an order of the tolerance for a Bayesian Inverse Problem (to be described in Section 4), one can specify a level L, and optimal Monte-Carlo sample sizes N_l so that the MSE of $\widehat{Y}$ is $\mathcal{O}(\epsilon^2)$ at a reduced computational cost.

3. Development of Multilevel SMC

3.1. Main Result

We will now obtain an analytical result that controls the error term $\mathbb{E}[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2]$ in expression (6). This is of general significance for the development of MLSMC in various contexts. Then, we will look in detail at an inverse problem context (developed in Section 4) and fully investigate the MLSMC method.

For any $l \in \{0, \ldots, L\}$ and $\varphi \in \mathcal{B}_b(E)$ we write: $\eta_l(\varphi) := \int_E \varphi(u)\eta_l(u)du$. We introduce the following assumptions, which will be verifiable in some contexts. They are rather strong, but could be relaxed at condsiderable increase in the complexity of the arguments, which will ultimately provide the same information. In addition, the assumptions are standard in the literature of SMC methods; see [7, 8].

0. Sample $U_0^1, \dots U_0^{N_0}$ i.i.d. from η_0 and compute $G_0(u_0^i)$ for each sample $i \in \{1, \dots, N_0\}$: Set $l = 0$.

1. Sample $\check{U}_l^1, \dots, \check{U}_l^{N_{l+1}}$ with replacement from $u_l^{1:N_l}$ with selection probabilities $\{G_l(u_l^1)/\sum_{j=1}^{N_l} G_l(u_l^j), \dots, (G_l(u_l^{N_l})/\sum_{j=1}^{N_l} G_l(u_l^j)\}$.

2. Sample $U_{l+1}^i | \check{u}_l^i$ from $M_{l+1}(\check{u}_l^i, \cdot)$ and compute $G_{l+1}(u_{l+1}^i)$ for each sample $i \in \{1, \dots, N_{l+1}\}$.

3. Set $l = l + 1$. If $l = L$ stop, otherwise return to the start of Step 1.

Figure 1: The SMC algorithm.

(**A1**) There exist $0 < \underline{C} < \overline{C} < +\infty$ such that

$$\begin{aligned} \sup_{l \geq 1} \sup_{u \in E} G_l(u) &\leq \overline{C} \; ; \\ \inf_{l \geq 1} \inf_{u \in E} G_l(u) &\geq \underline{C} \; . \end{aligned}$$

(**A2**) There exists a $\rho \in (0,1)$ such that for any $l \geq 1$, $(u, v) \in E^2$, $A \in \mathcal{E}$:

$$\int_A M_l(u, du') \geq \rho \int_A M_l(v, dv') \; .$$

Theorem 3.1. *Assume (A1-2). There exist $C < +\infty$ and $\kappa \in (0,1)$ such that for any $g \in \mathcal{B}_b(E)$, with $\|g\|_\infty = 1$,*

$$\mathbb{E}\big[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2\big] \leq C \left(\frac{1}{N_0} + \sum_{l=1}^{L} \frac{\| \frac{Z_{l-1}}{Z_l} G_{l-1} - 1 \|_\infty^2}{N_{l-1}} + \sum_{1 \leq l < q \leq L} \| \tfrac{Z_{l-1}}{Z_l} G_{l-1} - 1 \|_\infty \| \tfrac{Z_{q-1}}{Z_q} G_{q-1} - 1 \|_\infty \big\{ \tfrac{\kappa^{q-l}}{N_{l-1}} + \tfrac{1}{N_{l-1}^{1/2} N_{q-1}} \big\} \right) .$$

3.2. Proof of Theorem 3.1

The following notations are adopted; this will substantially simplify subsequent expressions:

$$\begin{aligned} Y_{l-1}^{N_{l-1}} &= \frac{\eta_{l-1}^{N_{l-1}}(gG_{l-1})}{\eta_{l-1}^{N_{l-1}}(G_{l-1})} - \eta_{l-1}^{N_{l-1}}(g) \; , \\ Y_{l-1} &= \frac{\eta_{l-1}(gG_{l-1})}{\eta_{l-1}(G_{l-1})} - \eta_{l-1}(g) \quad \big(\equiv \eta_l(g) - \eta_{l-1}(g) \big) \; , \qquad (7) \\ \overline{\varphi_l}(u) &= \big(\tfrac{Z_{l-1}}{Z_l} G_{l-1}(u) - 1 \big) \; , \\ \widetilde{\varphi}_l(u) &= g(u) \overline{\varphi_l}(u) \; , \\ A_n(\varphi, N) &= \eta_n^N(\varphi G_n)/\eta_n^N(G_n) \; , \quad \varphi \in \mathcal{B}_b(E) \; , \quad 0 \leq n \leq L-1 \; , \qquad (8) \\ \overline{A}_n(\varphi, N) &= A_n(\varphi, N) - \frac{\eta_n(\varphi G_n)}{\eta_n(G_n)} \; . \qquad (9) \end{aligned}$$

Throughout this Section, C is a constant whose value may change, but does not depend on any time parameters of the Feynman-Kac formula, nor N_l. The proof of Theorem 3.1 follows from several technical lemmas which are now given and supported by further results in the Appendix; the proof of the theorem is at the end of this subsection. It is useful to observe that $Z_l/Z_{l-1} = \eta_{l-1}(G_{l-1})$, $\eta_{l-1}(\overline{\varphi}_l) = 0$ and $|A_n(\varphi, N)| \leq |\varphi|_\infty$ with probability 1 as the conditional L_1-norm of functional φ over a discrete distribution. We will make repeated use of the following identity which follows from these observations upon adding and subtracting $\eta_{l-1}^{N_{l-1}}(\frac{Z_{l-1}}{Z_l} g(\cdot) G_{l-1}(\cdot))$:

$$Y_{l-1}^{N_{l-1}} - Y_{l-1} = A_{l-1}(g, N_{l-1}) \{\eta_{l-1} - \eta_{l-1}^{N_{l-1}}\}(\overline{\varphi_l}) + \{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\widetilde{\varphi}_l) \ . \tag{10}$$

Lemma 3.1. *Assume (A1-2). There exists a $C < +\infty$ such that for any $l \geq 1$:*

$$\|Y_{l-1}^{N_{l-1}} - Y_{l-1}\|_2^2 \leq \frac{C \, \|\frac{Z_{l-1}}{Z_l} G_{l-1} - 1\|_\infty^2}{N_{l-1}} \ .$$

Proof. From (10) and the C_2-inequality we obtain:

$$\begin{aligned} \|Y_{l-1}^{N_{l-1}} - Y_{l-1}\|_2^2 &\leq 2 \, \|A_{l-1}(g, N_{l-1}) \{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\overline{\varphi_l})\|_2^2 + 2 \, \|\{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\widetilde{\varphi}_l)\|_2^2 \\ &\leq 2 \, \|\{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\overline{\varphi_l})\|_2^2 + 2 \, \|\{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\widetilde{\varphi}_l)\|_2^2 \end{aligned}$$

By [7, Theorem 7.4.4] we have that both L_2-norms are upper bounded by $\frac{C\|\frac{Z_{l-1}}{Z_l} G_{l-1} - 1\|_\infty^2}{2N_{l-1}}$. This completes the proof. □

By the C_2-inequality and standard properties of i.i.d. random variables one has:

$$\mathbb{E}\big[\{\widehat{Y} - \mathbb{E}_{\eta_L}[g(U)]\}^2\big] = \mathbb{E}\Big[\Big\{\sum_{l=1}^{N} (Y_{l-1}^{N_{l-1}} - Y_{l-1})\Big\}^2\Big] \leq \frac{C}{N_0} + 2\,\mathbb{E}\Big[\Big\{\sum_{l=2}^{N} (Y_{l-1}^{N_{l-1}} - Y_{l-1})\Big\}^2\Big] \ .$$

We have that:

$$\mathbb{E}\Big[\Big\{\sum_{l=2}^{N} (Y_{l-1}^{N_{l-1}} - Y_{l-1})\Big\}^2\Big] = \mathbb{E}\Big[\sum_{l=2}^{N} (Y_{l-1}^{N_{l-1}} - Y_{l-1})^2\Big] + 2 \sum_{2 \leq l < q \leq L} \mathbb{E}\big[(Y_{l-1}^{N_{l-1}} - Y_{l-1})(Y_{q-1}^{N_{q-1}} - Y_{q-1})\big]$$

Lemma 3.1 gives that:

$$\mathbb{E}\Big[\sum_{l=2}^{N} (Y_{l-1}^{N_{l-1}} - Y_{l-1})^2\Big] \leq C \sum_{l=2}^{L} \frac{\|\frac{Z_{l-1}}{Z_l} G_{l-1} - 1\|_\infty^2}{N_{l-1}}$$

thus it remains to treat the cross-interaction terms. Using the decomposition in (10), we obtain

$$\begin{aligned} \sum_{2 \leq l < q \leq L} \mathbb{E}\big[(Y_{l-1}^{N_{l-1}} - Y_{l-1})(Y_{q-1}^{N_{q-1}} - Y_{q-1})\big] = \\ = \sum_{2 \leq l < q \leq L} \mathbb{E}\left[A_{l-1}(g, N) A_{q-1}(g, N) \{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\overline{\varphi_l}) \{\eta_{q-1}^{N_{q-1}} - \eta_{q-1}\}(\overline{\varphi_q})\right] \\ + \sum_{2 \leq l < q \leq L} \mathbb{E}\left[A_{l-1}(g, N) \{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\overline{\varphi_l}) \{\eta_{q-1}^{N_{q-1}} - \eta_{q-1}\}(\widetilde{\varphi_q})\right] \\ + \sum_{2 \leq l < q \leq L} \mathbb{E}\left[A_{q-1}(g, N) \{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\widetilde{\varphi}_l) \{\eta_{q-1}^{N_{q-1}} - \eta_{q-1}\}(\overline{\varphi_q})\right] \\ + \sum_{2 \leq l < q \leq L} \mathbb{E}\left[\{\eta_{l-1}^{N_{l-1}} - \eta_{l-1}\}(\widetilde{\varphi}_l) \{\eta_{q-1}^{N_{q-1}} - \eta_{q-1}\}(\widetilde{\varphi_q})\right] \ . \end{aligned}$$

We will now apply Proposition Appendix A.1 to the relevant terms in the sum, to yield the upper-bound:

$$C \sum_{1\leq l<q\leq L} \|\widetilde{\varphi}_l\|_\infty \|\widetilde{\varphi_q}\|_\infty \Big\{ \frac{\kappa^{q-l}}{N_{l-1}} + \frac{1}{N_{l-1}^{1/2} N_{q-1}} \Big\} .$$

From here one can conclude the proof of Theorem 3.1.

3.3. MLSMC Variance Analysis

This section considers the specification of parameters for the MLSMC algorithm after consideration of Theorem 3.1. Recall that in the simpler SDE setting of [14] one must work with the strong error estimate $\mathbb{E}|g(U_l)-g(U)|^2 = \mathcal{O}(h_l^\beta)$ and the deduced variance $V_l = \mathrm{Var}[g(U_l)-g(U_{l-1})] = \mathcal{O}(h_l^\beta)$. From Theorem 3.1, a similar role within MLSMC is taken by:

$$V_l := \| \tfrac{Z_{l-1}}{Z_l} G_{l-1} - 1 \|_\infty^2 . \tag{11}$$

We assume that in the given context one can obtain that $V_l = \mathcal{O}(h_l^\beta)$ for some appropriate rate constant $\beta \geq 1$ (see Proposition 4.1 and equation (26) later on). Recall that we have $h_l = M^{-l}$, for some integer $M > 1$ and we assume a bias of $\mathcal{O}(h_L^\alpha)$. Thus, targeting an error tolerance of ϵ, we have $h_L^\alpha = M^{-L} = \mathcal{O}(\epsilon)$, so that $L = \log(\epsilon^{-1})/(\alpha \log(M)) + \mathcal{O}(1)$. Now, to optimally allocate $N_0, N_1, \ldots, N_L$, one proceeds along the lines outlined in the Introduction under consideration of Theorem 3.1. Notice that $\sum_{q=l+1}^L \kappa^{q-l} \leq \frac{1}{1-\kappa}$ and V_q is smaller than V_l (in terms of the obtained upper bounds), so the upper bound in Theorem 3.1 can be bounded by:

$$\frac{1}{N_0} + \sum_{l=1}^L \left(\frac{h_l^\beta}{N_l} + \Big(\frac{h_l^\beta}{N_l}\Big)^{1/2} \sum_{q=l+1}^L \frac{h_q^{\beta/2}}{N_q} \right) . \tag{12}$$

We also assume a computational cost proportional to $\sum_{l=0}^L N_l h_l^{-\zeta}$, for some rate $\zeta \geq 1$, with the resampling cost considered to to be negligible for practical purposes compared to the cost of the calculating the importance weights (as it is the case for the inverse problems we focus upon later). As with standard MLMC in [14], we need to find $N_0, \ldots, N_L$ that optimize (12) given a fixed computational cost $\sum_{l=0}^L N_l h_l^{-\zeta}$. Such a constrained optimization with the complicated error bound in (12), results in the need to solve a quartic equation as a function of V_l and C_l. Instead, one can *assume* that the second term on the R.H.S. of (12) is negligible, solve the constrained optimization ignoring that term, and then check that the effect of that term for the given choice of $\{N_l\}_{l=0}^{L-1}$ is smaller than $\mathcal{O}(\epsilon^2)$. Following this approach gives a constrained optimisation problem identical to the simple case of [14], with solution $N_l \propto \sqrt{V_l/C_l} = \mathcal{O}(h_l^{(\beta+\zeta)/2})$. One works as in Section 1, and selects:

$$N_l \propto \epsilon^{-2} h_l^{(\beta+\zeta)/2} K_L \; ; \quad K_L \eqsim \sum_{l=0}^L h_l^{(\beta-\zeta)/2} .$$

Then returning to (12) one can check that indeed the extra summand is smaller than $\mathcal{O}(\epsilon^2)$ for the above choice of N_l. Notice that: (i) $h_q^{\beta/2}/N_q = \mathcal{O}(\epsilon^2 h_l^{-\zeta/2}/K_L)$, and the sum $\sum_{q=l+1}^L h_l^{-\zeta/2}$ is dominated by $h_L^{-\zeta/2} = \mathcal{O}(\epsilon^{-\zeta/(2\alpha)})$; (ii) we have $(h_l^\beta/N_l)^{1/2} \propto \epsilon/K_L^{1/2} h_l^{(\beta-\zeta)/4}$. Therefore,

$$\sum_{l=1}^L \left(\Big(\frac{h_l^\beta}{N_l}\Big)^{1/2} \sum_{q=l+1}^L \frac{h_q^{\beta/2}}{N_q} \right) = \mathcal{O}\Big(\epsilon^2 \epsilon^{1-\zeta/(2\alpha)} \sum_{l=0}^L h_l^{(\beta-\zeta)/4} / K_L^{3/2} \Big) = \mathcal{O}(\epsilon^2 \epsilon^{1-\zeta/(2\alpha)}) .$$

Thus, when $\zeta \leq 2\alpha$, the overall mean squared error is still $\mathcal{O}(\epsilon^2)$. In the inverse problem context of Section 4, we will establish that $\beta = 2$, $\alpha = \beta/2$. Also, in many cases (depending on the chosen PDE solver) we have $\zeta = d$ (d is the spatio-temporal dimension of the underlying continuum).

Remark 3.1. *If one were able to perform i.i.d. sampling from η_L, a computational effort proportional to $Nh_L^{-\zeta}$ is used, with N the number of simulated samples. To make the overall error (bias squared plus variance) of using i.i.d. sampling $\mathcal{O}(\epsilon^2)$ then one must take $N = \mathcal{O}(\epsilon^{-2})$, as the variance of the MC estimate is $[\eta_L(g^2) - \eta_L(g)^2]/N = \mathcal{O}(N^{-1})$, independently of L. This is a computational cost of $\mathcal{O}(\epsilon^{-2-\zeta})$ is used which is far worse than MLSMC samplers in most cases of practical interest. One notes that, of course, i.i.d. sampling is assumed not to be possible in the first instance.*

4. Bayesian Inverse Problem

A context will now be introduced in which the results are of interest and the assumptions can be satisfied. In particular, the ubiquitous problem of inferring the diffusion coefficient for an elliptic PDE will be considered. Solution of this problem is of broad interest across science and engineering applications. The diffusion coefficient $u(x)$ may represent for example thermal, electrical, or hydraulic conductivity of a material in the case that the solution $p(x)$ is temperature, electric potential, or pressure, respectively. An external forcing $f(x)$ on the right-hand side represents sources and/or sinks.

We begin with another round of notations. Introduce the Gelfand triple $V := H^1(D) \subset L^2(D) \subset H^{-1}(D) =: V^*$, where the domain D will be understood. Furthermore, denote by $\langle \cdot, \cdot \rangle, \| \cdot \|$ the inner product and norm on L^2, with superscripts to denote the corresponding inner product and norm on the Hilbert spaces V and V^*. Denote the finite dimensional Euclidean inner product and norms as $\langle \cdot, \cdot \rangle, | \cdot |$, with the latter also representing size of a set and absolute value, and denote weighted norms by adding a subscript as $\langle \cdot, \cdot \rangle_A := \langle A^{-\frac{1}{2}} \cdot, A^{-\frac{1}{2}} \cdot \rangle$, with corresponding norms $| \cdot |_A$ or $\| \cdot \|_A$ for Euclidean and L^2 spaces, respectively (for symmetric, positive definite A with $A^{\frac{1}{2}}$ being the unique symmetric square root). In the following, the generic constant C will be used for the right-hand side of inequalities as necessary, its precise value actually changing between usage.

Let $D \subset \mathbb{R}^d$ with $\partial D \in C^1$ convex. For $f \in V^*$, consider the following PDE on D:

$$-\nabla \cdot (\widehat{u} \nabla p) = f \ , \quad \text{on } D \ , \tag{13}$$

$$p = 0 \ , \quad \text{on } \partial D \ , \tag{14}$$

where:

$$\widehat{u}(x) = \bar{u}(x) + \sum_{k=1}^{K} u_k \sigma_k \phi_k(x) \ . \tag{15}$$

Define $u = \{u_k\}_{k=1}^K$, with $u_k \sim U[-1, 1]$ i.i.d. (the uniform distribution on $[-1, 1]$). This determines the prior distribution for u. Assume that $\bar{u}, \phi_k \in C^\infty$ for all k and that $\|\phi_k\|_\infty = 1$. In particular, assume $\{\sigma_k\}_{k=1}^K$ decay[1] with k. The state space is $E = \prod_{k=1}^K [-1, 1]$. It is important that the

[1]If $K \to \infty$ it is important that they decay with a suitable rate in order to ensure u lives almost surely in an appropriate sequence-space, or equivalently $\widehat{u}$ lives in the appropriate function-space. However, here we down-weight higher frequencies as necessary only to induce certain smoothness properties, while actually for a given value of $u \in E$ the resulting permeability $\widehat{u} \in \widehat{E} \subset C^\infty(D) \subset C(D) \subset L^\infty(D) \subset L^p(D)$ for all $p \geq 1$.

following property holds:

$$\inf_x \widehat{u}(x) \geq \inf_x \bar{u}(x) - \sum_{k=1}^{K} \sigma_k \geq u_* > 0$$

so that the operator on the left-hand side of (13) is uniformly elliptic. Let $p(\cdot; u)$ denote the weak solution of (13) for parameter value u. Define the following vector-valued function

$$\mathcal{G}(p) = [g_1(p), \cdots, g_M(p)]^\top ,$$

where g_m are elements of the dual space V^* for $m = 1, \ldots, M$. It is assumed that the data take the form

$$y = \mathcal{G}(p) + \xi , \quad \xi \sim N(0, \Gamma) , \quad \xi \perp u , \tag{16}$$

where $N(0, \Gamma)$ denotes the Gaussian random variable with mean 0 and covariance Γ, and $\perp$ denotes independence. The unnormalized density then is given by:

$$\gamma(u) = e^{-\Phi[\mathcal{G}(p(\cdot;u))]} ; \quad \Phi(\mathcal{G}) = \tfrac{1}{2} |\mathcal{G} - y|^2_\Gamma .$$

Consider the triangulated domains $\{D^l\}_{l=1}^{\infty}$ approximating D, where l indexes the number of nodes $N(l)$, so that we have $D^1 \subset \cdots \subset D^l \subset D^\infty := D$, with sufficiently regular triangles. Consider a finite element discretization on D^l consisting of H^1 functions $\{\psi_\ell\}_{\ell=1}^{N(l)}$. In particular, continuous piecewise linear hat functions will be considered here, the explicit form of which will be given in section 5.1. Denote the corresponding space of functions of the form $\varphi = \sum_{\ell=1}^{N(l)} v_\ell \psi_\ell^l$ by V^l, and notice that $V^1 \subset V^2 \subset \cdots \subset V^l \subset V$. By making the further Assumption 7 of [16] that the weak solution $p(\cdot; u)$ of (13)-(14) for parameter value u is in the space $W = H^2 \cap H_0^1 \subset V$, one obtains a well-defined finite element approximation $p^l(\cdot; u)$ of $p(\cdot; u)$. Thus, the sequence of distributions of interest in this context is:

$$\eta_l(u) = \frac{\gamma_l(u)}{Z_l} = \frac{e^{-\Phi[\mathcal{G}(p^l(\cdot;u))]}}{\int_E e^{-\Phi[\mathcal{G}(p^l(\cdot;u))]} du} , \quad l = 0, 1, \ldots, L.$$

4.1. Error Estimates

Notice one can take the inner product of (13) with the solution $p \in V$, and perform integration by parts on the right-hand side, in order to obtain $\langle \widehat{u} \nabla p, \nabla p \rangle = \langle f, p \rangle$. Therefore

$$u_* \|p\|_V^2 = u_* \langle \nabla p, \nabla p \rangle \leq \langle \widehat{u} \nabla p, \nabla p \rangle = \langle f, p \rangle \leq \|f\|_{V^*} \|p\|_V . \tag{17}$$

So the following bound holds in V, uniformly over u:

$$\|p(\cdot; u)\|_V \leq \frac{\|f\|_{V^*}}{u_*} . \tag{18}$$

Notice that:

$$|\mathcal{G}(p) - \mathcal{G}(p')| = \Big(\sum_{m=1}^{M} \langle g_m, p - p' \rangle^2 \Big)^{1/2} \leq \|p - p'\|_V \sum_{m=1}^{M} \|g_m\|_{V^*} = C \|p - p'\|_V . \tag{19}$$

So the following uniform bound also holds:

$$|\mathcal{G}(p(\cdot; u))| \leq C \frac{\|f\|_{V^*}}{u_*} .$$

The uniform bound on $\mathcal{G}$ provides the Lipschitz bound

$$|\Phi(\mathcal{G}) - \Phi(\mathcal{G}')| \leq C|\mathcal{G} - \mathcal{G}'|, \tag{20}$$

obtained as follows:

$$\begin{aligned} |\Phi(\mathcal{G}) - \Phi(\mathcal{G}')| =& \frac{1}{2}\left| |\mathcal{G} - y|_\Gamma^2 - |\mathcal{G}' - y|_\Gamma^2 \right| \\ &= \left| |\mathcal{G}|_\Gamma^2 - |\mathcal{G}'|_\Gamma^2 + 2\langle \mathcal{G}' - \mathcal{G}, y\rangle_\Gamma \right| \\ &\leq \left(|\mathcal{G}| + |\mathcal{G}'| + 2|y|\right) |\Gamma^{-1}||\mathcal{G} - \mathcal{G}'| \ , \end{aligned}$$

Setting $\mathcal{G}' = 0$ gives the boundedness of Φ.

Considering some sequence h_l indicating the maximum diameter of an individual element at level l, with $h_l \to 0$ (e.g. $h_l = 2^{-l}$), the following asymptotic bound holds for continuous piecewise linear hat functions [5][2]

$$\|p(\cdot; u) - p^l(\cdot; u)\|_V \leq C h_l \|p(\cdot; u)\|_W \ . \tag{21}$$

Furthermore, Proposition 29 of [16] provides a uniform bound based on the following decomposition of (13):

$$-\Delta p = \frac{1}{\widehat{u}} \left(f + \nabla \widehat{u} \cdot \nabla p\right) \ .$$

Thus, we have

$$\begin{aligned} \sup_u \|p(\cdot; u)\|_W &\leq C' \sup_u \|\Delta p(\cdot; u)\| \\ &\leq \frac{C'}{u_*} \sup_u \left(\|f\| + \|\widehat{u}\|_V \|p\|_V\right) \\ &\leq C\|f\| \ , \end{aligned} \tag{22}$$

where the first line holds by equivalence of norms, the second holds since $\widehat{u} \in C^\infty$, by the triangle inequality and Cauchy-Schwarz inequality, and the last line holds by (18) and the fact $\|f\|_{V^*} \leq c\|f\|$ for some c. The constant C depends on $u_*, \|\nabla \widehat{u}\|_\infty, C'$, and c . Note that $\|\widehat{u}\|_V \leq \|\nabla \widehat{u}\|_\infty \leq C'' < \infty$ by (15). Note that the bound (22) in (21) together with (18) provides a uniform bound over l for $\mathcal{G}^l$, defined by $\mathcal{G}^l : u \mapsto \mathcal{G}(p^l(\cdot; u))$, following the same argument as (19), which means that the Lipschitz bound in (20) holds here over different l as well.

Now, the following holds by (21), (22), (18), and the triangle inequality

$$\|p^l(\cdot; u) - p^{l-1}(\cdot; u)\|_V \leq C h_l \ . \tag{23}$$

Hence, from (19)

$$|\mathcal{G}^l(u) - \mathcal{G}^{l-1}(u)| = |\mathcal{G}(p^l(\cdot; u)) - \mathcal{G}(p^{l-1}(\cdot; u))| \leq C h_l \ , \tag{24}$$

where C is independent of the realization of u.

Recall from Section 3, $G_{l-1}(u) = \gamma_l(u)/\gamma_{l-1}(u)$ and that we seek to be able to verify (A1); the following result will help us to do that.

Proposition 4.1. *For $G_{l-1}(u) := \exp\{\Phi(\mathcal{G}^{l-1}(u)) - \Phi(\mathcal{G}^l(u))\}$ one has the following estimates, uniformly in u:*

$$1 - \mathcal{O}(h_l) = \underline{C}_l := e^{-Ch_l} \leq G_{l-1} = \exp\{\Phi(\mathcal{G}^{l-1}) - \Phi(\mathcal{G}^l)\} \leq e^{Ch_l} =: \overline{C}_l = 1 + \mathcal{O}(h_l). \tag{25}$$

[2]Higher order finite elements can yield stronger convergence rates, but will not be considered here in the interest of a more streamlined presentation. In fact, even this estimate is not sharp, but it is suitable to illustrate the theory.

Proof. In combination with (20), equation (24) gives the stated result. □

Proposition 4.2 (Bias)**.** *Let $g \in \mathcal{B}_b(E)$. Then*

$$|\mathbb{E}_{\eta_L}[g(U)] - \mathbb{E}_{\eta_\infty}[g(U)]| \leq Ch_L \ .$$

Proof. It follows from the same reasoning as in Proposition 4.1, upon observing that

$$\mathbb{E}_{\eta_L}[g(U)] - \mathbb{E}_{\eta_\infty}[g(U)] = \mathbb{E}_{\eta_\infty}\left[g(U)\left(\frac{d\eta_L}{d\eta_\infty} - 1\right)\right] \ .$$

□

4.2. Verification of Assumptions

Assumption (A1) is satisfied by letting

$$\underline{C} := \inf_{l\geq 1} \underline{C}_l \ ; \quad \overline{C} := \sup_{l\geq 1} \overline{C}_l \ .$$

Notice that the asymptotic bounds of Proposition 4.1 imply that $\underline{C}_l$ is increasing with l while $\overline{C}_l$ are decreasing with l. Therefore, these will actually be minimum and maximum over a sufficiently large set of low indices. We remark that as

$$\begin{aligned} \frac{Z_{l-1}}{Z_l}G_{l-1} - 1 &= \frac{Z_{l-1}}{Z_l}[G_{l-1} - 1] - \frac{Z_l - Z_{l-1}}{Z_l} \\ &= \frac{Z_{l-1}}{Z_l}[G_{l-1} - 1] - \frac{1}{Z_l}\int_E (G_{l-1}(u) - 1)\gamma_{l-1}(u)du \end{aligned}$$

then by the above verification of (A1) and Proposition 4.1, one does indeed have that

$$\|\tfrac{Z_{l-1}}{Z_l}G_{l-1} - 1\|_\infty^2 = \mathcal{O}(h_l^\beta) \tag{26}$$

as was claimed in Section 3.3.

Assumption (A2) can be shown to hold in this context, if a Gibbs sampler is constructed. Let θ be the uniform measure on $[-1, 1]$ and consider a probability measure π on $E := \prod_{i=1}^K [-1, 1]$ with density w.r.t. the measure $\bigotimes_{i=1}^K \theta(du_i)$:

$$\pi(u) = \frac{\exp\{-\Phi(u)\}}{\int_E \exp\{-\Phi(u)\}\bigotimes_{i=1}^K \theta(du_i)}$$

where it is assumed that $\forall u \in E$, $\Phi(u) \in [0, \Phi^*]$. This is the setting above, for all l, following from equations (20) and (24).

Let $k \in \mathbb{N}, k < K$ be given and consider a partition of $\{1, \ldots, K\}$ into k disjoint subsets $(a_i)_{i=1}^k$. For example $k = 2$ and a_1 and a_2 are the sets of (positive) odd and even numbers up to K, respectively.

One can consider the Gibbs sampler to generate from π, with kernel:

$$M(u, du') = \Big(\prod_{j=1}^k \pi(u'_{a_j}|u'_{a_1:a_{j-1}}, u_{a_{j+1}:a_k})\Big)\bigotimes_{i=1}^K \theta(du'_i)$$

with

$$\pi(u'_{a_j}|u'_{a_1:a_{j-1}}, u_{a_{j+1}:a_k}) = \frac{\pi(u'_{a_1:a_j}, u_{a_{j+1}:a_k})}{\int_{[-1,1]^{|\{a_j\}|}} \pi(u'_{a_1:a_j}, u_{a_{j+1}:a_k}) \bigotimes_{i\in(a_j)} \theta(du'_i)}.$$

One can, for example, perform rejection sampling on π using the prior as a proposal (and accepting with probability $\exp\{-\Phi(u)\}$) and we would still have a theoretical acceptance probability of

$$\int_E \exp\{-\Phi(u)\} \bigotimes_{i=1}^{K} \theta(du_i) \geq \exp\{-\Phi^*\}.$$

Sampling from the full conditionals will have a higher-acceptance probability and thus the Gibbs sampler is not an unreasonable algorithm.

Proposition 4.3. *For any $u, \tilde{u} \in E$*

$$M(\tilde{u}, du') \geq \exp\{-2\Phi^*(k-1)\} M(u, du').$$

Proof. Consider

$$\begin{aligned} \frac{\pi(u'_{a_j}|u'_{a_1:a_{j-1}}, u_{a_{j+1}:a_k})}{\pi(u'_{a_j}|u'_{a_1:a_{j-1}}, \tilde{u}_{a_{j+1}:a_k})} &= \frac{\pi(u'_{a_1:a_j}, u_{a_{j+1}:a_k})}{\pi(u'_{a_1:a_j}, \tilde{u}_{a_{j+1}:a_k})} \frac{\int_{[-1,1]^{|a_j|}} \pi(u'_{a_1:a_j}, \tilde{u}_{a_{j+1}:a_k}) \bigotimes_{i\in(a_j)} \theta(du'_i)}{\int_{[-1,1]^{|a_j|}} \pi(u'_{a_1:a_j}, u_{a_{j+1}:a_k}) \bigotimes_{i\in(a_j)} \theta(du'_i)} \\ &\leq \exp\{2\Phi^*\}. \end{aligned}$$

Thus, since

$$M(u, du') = \Big(\prod_{j=1}^{k} \pi(u'_{a_j}|u'_{a_1:a_{j-1}}, u_{a_{j+1}:a_k}) \Big) \bigotimes_{i=1}^{K} \theta(du'_i),$$

and

$$M(\tilde{u}, du') = \Big(\prod_{j=1}^{k} \pi(u'_{a_j}|u'_{a_1:a_{j-1}}, \tilde{u}_{a_{j+1}:a_k}) \Big) \bigotimes_{i=1}^{K} \theta(du'_i),$$

and the final element in each product is identical, it follows that

$$M(\tilde{u}, du') \geq \exp\{-2\Phi^*(k-1)\} M(u, du').$$

as was to be proved. □

5. Numerical Results

5.1. Set-Up

In this section a 1D version of the elliptic PDE problem in (13) is considered. Let $D = [0, 1]$ and consider $f(x) = 100x$. For the prior specification of u, we set $K = 50$, $\bar{u}(x) = 0.15$, and for $k > 0$, let $\sigma_k = (2/5)4^{-k}$, $\phi_k(x) = \sin(k\pi x)$ if k is odd and $\phi_k(x) = \cos(k\pi x)$ if k is even. The forward problem at resolution level l is solved using a finite element method with piecewise linear shape functions on a uniform mesh of width $h_l = 2^{-(l+k)}$, for some starting $k \geq 1$ (so that there are at least two grid-blocks in the finest, $l = 0$, case). Thus, on the l^{th} level the finite-element basis functions are $\{\psi_i^l\}_{i=1}^{2^{l+k}-1}$ defined as (for $x_i = i \cdot 2^{-(l+k)}$) [5]:

$$\psi_i^l(x) = \begin{cases} (1/h_l)[x - (x_i - h_l)] & if \quad x \in [x_i - h_l, x_i], \\ (1/h_l)[x_i + h_l - x] & if \quad x \in [x_i, x_i + h_l]. \end{cases}$$

The functional of interest g is taken as the solution of the forward problem at the midpoint of the domain, that is $g(u) = p(0.5; u)$. The observation operator is $\mathcal{G}(u) = [p(0.25; u), p(0.75; u)]^\top$, and the observational noise covariance is taken to be $\Gamma = 0.25^2 I$.

To solve the PDE, the ansatz $p^l(x) = \sum_{i=1}^{2^{l+k}-1} p_i^l \psi_i^l(x)$ is plugged into (13), and projected onto each basis element:

$$-\left\langle \nabla \cdot \left(\hat{u} \nabla \sum_{i=1}^{2^{l+k}} p_i^l \psi_i^l(x) \right), \psi_j^l(x) \right\rangle = \langle f, \psi_j^l \rangle \ ,$$

resulting in the following linear system:

$$\mathbf{A}^l(u)\mathbf{p}^l = \mathbf{f}^l,$$

where we introduce the matrix $\mathbf{A}^l(u)$ with entries $A_{ij}^l(u) = \langle \widehat{u} \nabla \psi_i^l, \nabla \psi_j^l \rangle$, and vectors $\mathbf{p}^l, \mathbf{f}^l$ with entries p_i^l and $f_i^l = \langle f, \psi_i^l \rangle$, respectively. Omitting the index l, the matrix is sparse and tridiagonal with

$$A_{(i-1)i}(u) = A_{i(i-1)}(u) = -(1/h^2) \int_{x_{i-1}}^{x_i} \widehat{u}(x) dx \ , \quad A_{ii} = (1/h^2) \left(\int_{x_{i-1}}^{x_i} \widehat{u}(x) dx + \int_{x_i}^{x_{i+1}} \widehat{u}(x) dx \right) \ ,$$

and zero otherwise. The elements f_i are computed analogously. The system can therefore be solved with cost $\mathcal{O}(2^{l+k})$, corresponding to a computational cost rate of $\zeta = 1$.

To get some understanding about the numerics and validate the theory, a number of results and figures will be generated. First, in subsection 5.2.1 the MSE as a function of cost is considered using the theoretical rate $\beta = 2$ (21) and $\alpha = \beta/2$ (following the error analysis in Section 4.1). The numbers N_l are optimally allocated with this β and the ζ above, using the formulae from Section 3.3. Observing the cost vs. error trend for a range of errors ϵ, we observe the appropriate scaling between computational cost and mean squared error (e.g. cost $\propto$ MSE^{-1} for MLSMC). Next, in subsection 5.2.2, the rate β is estimated, illustrating that the estimate (21) is not sharp.

5.2. Results

The following setting is simulated. The sequence of step-sizes is given by $h_l = 2^{-(l+k)}$, $k = 3$. The data $\mathcal{G}(u)$ is simulated with a given $u_i \sim U[-1, 1]$ (i=1,...,50) and $h = 2^{-20}$. The observation variance and other algorithmic elements are as stated above. We will contrast the cost of two algorithms for different target errors. The first is (i) MLSMC as detailed above; the second is (ii) plain SMC: the same sequence of distributions as MLSMC, but using equal number of particles for a given L, and averaging only the samples at the last level. For both MLSMC and SMC algorithms, random walk MCMC kernels were used. The algorithm updated u in blocks of variables.

5.2.1. Algorithmic Performance with Diminishing MSE

Given the choices of $\alpha = 1$ and $\beta = 2$, the performance of the MLSMC algorithm is benchmarked by simulating samplers with different maximum levels L. The value of $\eta_\infty(g)$ was first estimated with the MLSMC algorithm targeting $\eta_{12}(g)$ (h^{-14}), with $N_L = 1000$. This sampler was realized 100 times and the average of the estimator is take as the ground truth. The standard deviation is much smaller than the smallest bias of subsequent simulations. Now, for each L the MLSMC and standard SMC samplers are each realized 100 times and the MSE with respect to this ground truth is reported. When updating $L \to L + 1$, the new bias is approximately a factor $2^{-\alpha}$ smaller than the previous one. Therefore the two sources of error in (6) can be roughly balanced by setting $N_l' = 2^{2\alpha} N_l$, for $l = 0, 1, \ldots, L$, and $N_{L+1}' = 2^{2\alpha - (\beta+\zeta)/2} N_L'$ for the MLSMC algorithm.

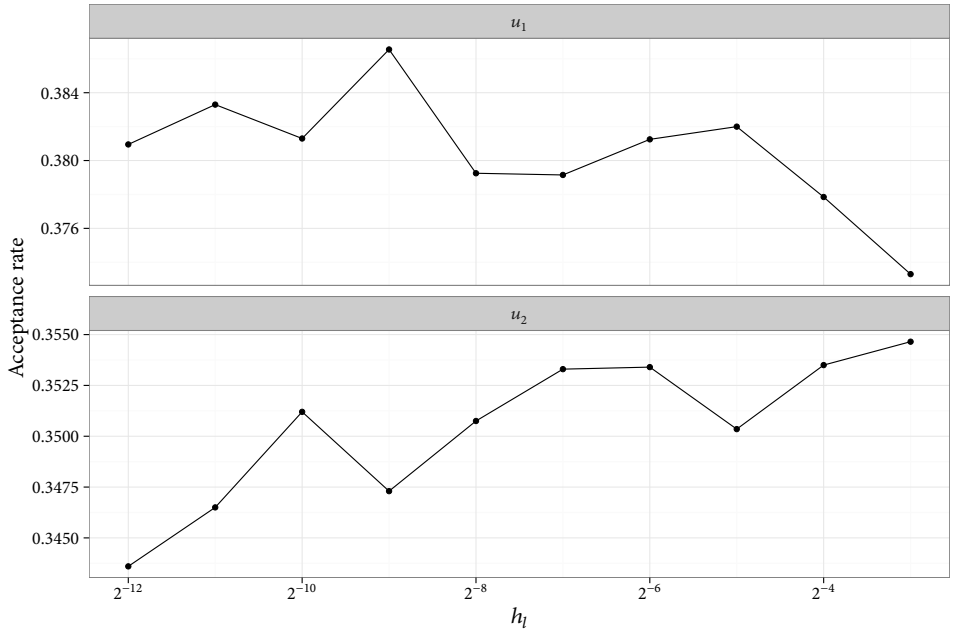

Figure 2: Acceptance rates of MCMC kernel.

For the SMC algorithm, the number of particles is updated to $N' = 2^{-2\alpha}N$. To check the effectiveness of the MCMC steps employed for dispersing the particles within the SMC methods, we show in Figure 2 the average (over the number of particles) acceptance probability of the first two parameters for each iteration when the MCMC was executed. The plot indicates reasonable performance of this particular aspect of the sequential algorithm.

The cost-vs-error plots for SMC and MLSMC are shown in Figure 3. Note that the bullets in the graph correspond to different choices of L (ranging from $L = 0$ to $L = 9$ for MLSMC and $L = 0$ to $L = 6$ for SMC). As mentioned above, the MSE data points are each estimated with 100 realizations of the given sampler. The fitted linear model of log Cost against log MSE has a gradient of -1.568 and -1.061 for SMC and MLSMC respectively. This verifies numerically the expected asymptotic behavior $\text{Cost} \propto \text{MSE}^{-1}$ for MLSMC, determined from the theory. Furthermore, the first rate indicates that the single level SMC performs similarly to the single level vanilla MC with asymptotic behavior $\text{cost} \propto \text{MSE}^{-3/2}$. The results clearly establish the potential improvements of MLSMC versus a standard SMC sampler. It is remarked that the MLSMC algorithm can be improved in many directions and this is subject to future work.

5.2.2. Numerical Estimation of Algorithmic Rates

In general a rate of convergence may not be a priori available, and so one may wish to empirically estimate the rate. In the case presented above the rate is not sharp, so the value of β can be estimated numerically. To this end the quantity $\|p^l(\cdot; u) - p^{l-1}(\cdot; u)\|_V^2$ is computed over increasing levels l, using a reference value of u. Figure 4 shows these computed values plotted against h_l on base-2 logarithmic scales. A fit of a linear model gives rate $\beta = 2.009$, and a similar experiment gives $\alpha = 1.015$. This is consistent with the rate $\beta = 2$ and $\alpha = \beta/2$ expected from the theoretical error analysis in Section 4.1 (and agrees also with other literature [5]) and used above in subsection 5.2.1.

An expensive MLSMC, using theoretical rates β but exceedingly larger N_L is executed to get some results over the algorithmic variabilty. The simulations are repeated 100 times. The

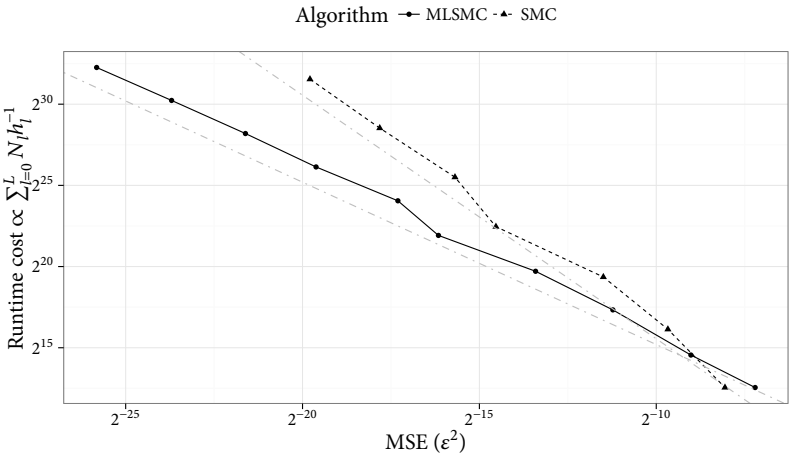

Figure 3: Computational cost against mean squared error against.

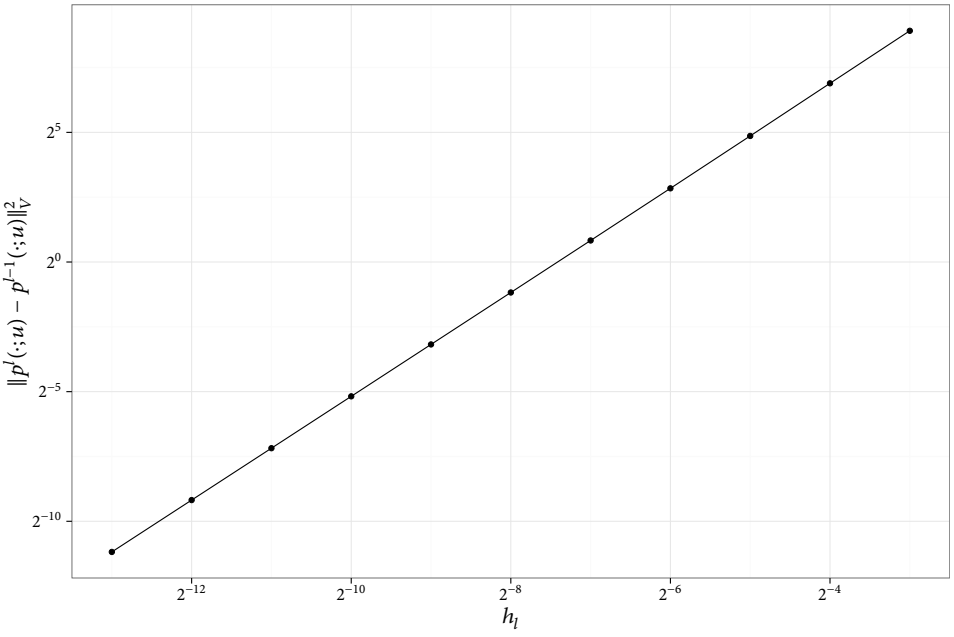

Figure 4: Variance rate estimate using H^1-norms.

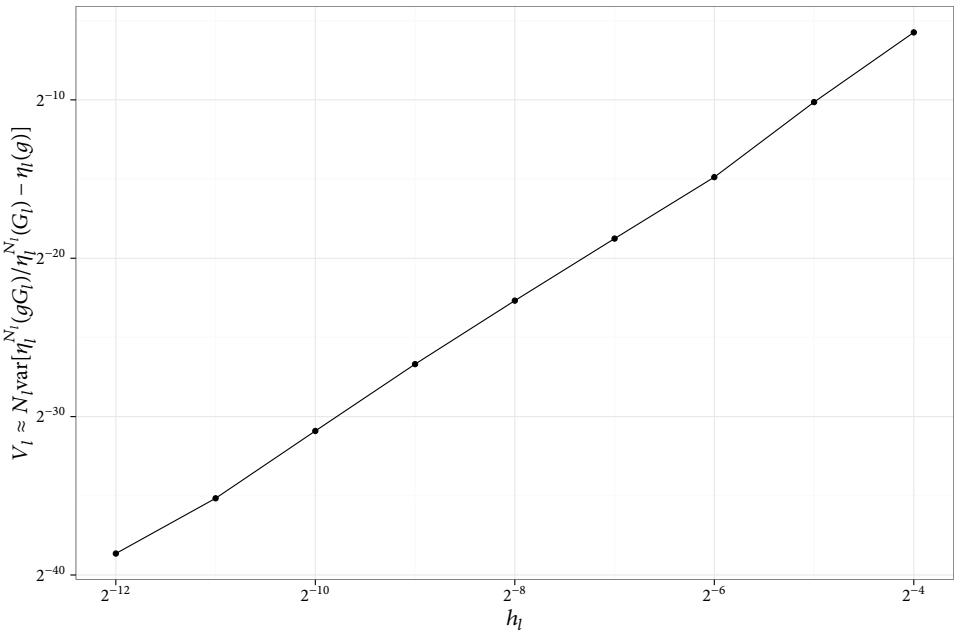

Figure 5: Empirical variance rate estimates.

estimated variance of $\eta_l^{N_l}(gG_l)/\eta_l^{N_l}(G_l) - \eta_l^{N_l}(g)$ multiplied by the sample size N_l, as a proxy of V_l, is plotted in Figure 5 against h_l on the same scales as before. The estimate of the rate now is $\beta = 4.111$. In this case the numerical estimate is much stronger than the theoretical rate used here. In fact, under suitable regularity conditions one may theoretically obtain the rate $\beta = 4$ with a stronger $L^2(D)$ bound on $\|p(\cdot; u) - p^l(\cdot; u)\|$, which follows from an Aubin-Nitsche duality argument [13]. Nonetheless, the objective of the present work is to illustrate the theory and not to really optimize the implementation. In fact, similar results as presented above are obtained using either rate, presumably owing to the fact that $\beta = 2 > \zeta$, which is already the optimal relationship of β and ζ and hence already provides the optimal asymptotic behavior of MSE$\propto$cost^{-1}. In case an optimal β induces a change in the relationship between β and ζ, one may expect a change in asymptotic behavior of cost vs. MSE, which justifies such empirical rate estimation.

6. Extensions

This work has considered an MLSMC sampler for a Bayesian inverse problem. Many extensions are possible. One extension concerns new ML estimators of the normalising constant Z_L as well as a complexity investigation. Another is consideration of the removal of discretisation bias in the spirit of [20]. A third extension is the implementation on more challenging and advanced models. These extensions are considered in [11, 17] as they require considerable extra efforts to achieve.

Acknowledgements

AJ, KL & YZ were supported by an AcRF tier 2 grant: R-155-000-143-112. AJ is affiliated with the Risk Management Institute and the Center for Quantitative Finance at NUS. RT, KL & AJ were additionally supported by King Abdullah University of Science and Technology (KAUST). KL was further supported by ORNLDRD Strategic Hire grant. AB was supported by the Leverhulme Trust Prize. We thank the referees for their comments which have greatly improved the article.

Appendix A. Technical Results

Introduce the following notations. For $\varphi \in \mathcal{B}_b(E)$, $p \geq 0$ and $\eta \in \mathcal{P}(E)$

$$\Phi_p(\eta)(\varphi) = \frac{\eta(G_{p-1}M_p(\varphi))}{\eta(G_{p-1})}$$

where $M_p(\varphi)(u) = \int_E \varphi(v)M_p(u, dv)$. Define the operator $Q_{p+1}(u, dv) = G_p(u)M_{p+1}(u, dv)$ and denote $Q_{p,n}(\varphi) = Q_{p+1}(\cdots Q_n(\varphi))$ ($0 \leq p \leq n$, $Q_{n,n}$ is the identity operator). Also set

$$D_{p,n}(\varphi) = \frac{Q_{p,n}(\varphi - \eta_n(\varphi))}{\eta_p(Q_{p,n}(1))},$$

$D_{n,n}$ is the identity operator, and define the following

$$\begin{aligned} V_p^{N_p}(\varphi) &= \sqrt{N_p}\,[\eta_p^{N_p} - \Phi_p(\eta_{p-1}^{N_{p-1}})](\varphi)\ , \\ R_{p+1}^{N_p}(D_{p,n}(\varphi)) &= \frac{\eta_p^{N_p}(D_{p,n}(\varphi))}{\eta_p^{N_p}(G_p)}[\eta_p(G_p) - \eta_p^{N_p}(G_p)]\ , \end{aligned} \tag{A.1}$$

with the convention that $\Phi_0(\eta_{-1}^{N_{-1}}) \equiv \eta_0$. Working similarly to the derivation of [10, Eq. (6.2)], but now with varying number of particles, we have that for any $n \geq 0$

$$\begin{aligned} [\eta_n^{N_n} - \eta_n](\varphi) &= [\eta_n^{N_n} - \Phi_n(\eta_{n-1}^{N_{n-1}})](\varphi) + [\Phi_n(\eta_{n-1}^{N_{n-1}}) - \eta_n](\varphi) \\ &= \frac{V_n^{N_n}(\varphi)}{\sqrt{N_n}} + \frac{\eta_{n-1}^{N_{n-1}}(G_{n-1}M_n(\varphi))}{\eta_{n-1}^{N_{n-1}}(G_{n-1})} - \frac{\eta_{n-1}(G_{n-1}M_n(\varphi))}{\eta_{n-1}(G_{n-1})} \\ &= \frac{V_n^{N_n}(\varphi)}{\sqrt{N_n}} + R_n^{N_{n-1}}(D_{n-1,n}(\varphi)) + [\eta_{n-1}^{N_{n-1}} - \eta_{n-1}](D_{n-1,n}(\varphi)) \end{aligned}$$

where notice that $D_{n-2,n-1}D_{n-1,n} = D_{n-2,n}$. Thus, working iteratively we have that

$$[\eta_n^{N_n} - \eta_n](\varphi) = \sum_{p=0}^{n} \frac{V_p^{N_p}(D_{p,n}(\varphi))}{\sqrt{N_p}} + \sum_{p=0}^{n-1} R_{p+1}^{N_p}(D_{p,n}(\varphi)). \tag{A.2}$$

Throughout this Section C is a constant whose value may change, but does not depend on any time parameters of the Feynman-Kac formula, nor $(N_0, \ldots, N_{L-1})$.

Now a technical Lemma is introduced, which will contain results that are frequently used in the below calculations.

Lemma Appendix A.1. *Assume (A1-2). There exist $C < +\infty$, $\kappa \in (0, 1)$ such that for any $n \geq p \geq 0$, $q \geq s \geq 0$, $1 \leq r < +\infty$ and $\varphi_n, \varphi_q \in \mathcal{B}_b(E)$:*

i) $\|D_{p,n}(\varphi_n)\|_\infty \leq C\kappa^{n-p}\|\varphi_n\|_\infty$.

ii) $\|V_p^{N_p}(D_{p,n}(\varphi_n))\|_r \leq C\kappa^{n-p}\|\varphi_n\|_\infty$.

iii) $\|R_{p+1}^{N_p}(D_{p,n}(\varphi_n))\|_r \leq \frac{C\kappa^{n-p}\|\varphi_n\|_\infty}{N_p}$.

iv) $\|V_p^{N_p}(D_{p,n}(\varphi_n))V_s^{N_s}(D_{s,q}(\varphi_q))\|_1 \leq C\kappa^{n-p+q-s}\|\varphi_n\|_\infty\|\varphi_q\|_\infty$.

v) $\|V_p^{N_p}(D_{p,n}(\varphi_n))R_{s+1}^{N_s}(D_{s+1,q}(\varphi_q))\|_1 \leq \frac{C\kappa^{n-p+q-s}\|\varphi_n\|_\infty\|\varphi_q\|_\infty}{N_s}$.

vi) $\|R_{p+1}^{N_p}(D_{p+1,n}(\varphi_n))R_{s+1}^{N_s}(D_{s+1,q}(\varphi_q))\|_1 \leq \frac{C\kappa^{n-p+q-s}\|\varphi_n\|_\infty\|\varphi_q\|_\infty}{N_pN_s}$.

Proof. For (i). This follows from standard calculations in the analysis of Feynman-Kac formulae; see e.g. the proof of Proposition 2 in [19]. For (ii).This follows from [7, Lemma 7.3.3] and (i). For (iii). Recall (A.1) and note that $\eta_p(D_{p,n}(\varphi_n)) = 0$; then on application of Cauchy-Schwarz and assumption (A1) one has that

$$\|R_{p+1}^{N_p}(D_{p,n}(\varphi_n))\|_r \leq C\,\|\eta_p^{N_p}(D_{p,n}(\varphi_n))\|_{2r}\cdot\|\eta_p(G_p)-\eta_p^{N_p}(G_p)\|_{2r}\ .$$

The result follows from [7, Theorem 7.4.4] and (i). (iv) follows from Cauchy-Schwarz and (ii). (v) follows from Cauchy-Schwarz, (ii) and (iii). (vi) follows from Cauchy-Schwarz and (iii). □

Recall equations (8) and (9), and define the following terms,

$$\mathcal{V}_n(\varphi,N)=\sum_{p=0}^{n}\frac{V_p^{N_p}(D_{p,n}(\varphi))}{\sqrt{N_p}}\ ;\quad \mathcal{R}_n(\varphi,N)=\sum_{p=0}^{n-1}R_{p+1}^{N_p}(D_{p,n}(\varphi))\ .$$

Here we use a slight abuse of notation for N, representing $(N_0,\ldots,N_n)$ (or $(N_0,\ldots,N_{n-1})$).

Lemma Appendix A.2. *Assume (A1-2). There exist a $C<+\infty$, $\kappa\in(0,1)$ such that for any $n>q\geq 0$ and $\varphi_n,\varphi_q,g\in\mathcal{B}_b(E)$, $\|g\|_\infty=1$:*

i) $\left|\mathbb{E}\left[\,A_q(g,N_q)\mathcal{V}_n(\varphi_n,N)\mathcal{V}_q(\varphi_q,N)\,\right]\right| \leq \frac{C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\kappa^{n-q}}{N_q}$.

ii) $\left|\mathbb{E}\left[\,A_q(g,N_q)\mathcal{V}_n(\varphi_n,N)\mathcal{R}_q(\varphi_q,N)\,\right]\right| \leq \frac{C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\kappa^{n-q}}{N_q^{3/2}}$.

iii) $\left|\mathbb{E}\left[\,A_q(g,N_q)\mathcal{V}_q(\varphi_q,N)\mathcal{R}_n(\varphi_n,N)\,\right]\right| \leq \frac{C\|\varphi_n\|_\infty\|\varphi_q\|_\infty}{\sqrt{N_q}N_n}$.

iv) $\left|\mathbb{E}\left[\,A_q(g,N_q)\mathcal{R}_q(\varphi_q,N)\mathcal{R}_n(\varphi_n,N)\,\right]\right| \leq \frac{C\|\varphi_n\|_\infty\|\varphi_q\|_\infty}{N_nN_q}$.

Proof. Set $\mathcal{F}_q^N$ as the σ algebra generated by particle system up to time q.

i) We start by noting that $\mathbb{E}\left[\,V_p^{N_p}(D_{p,n}(\varphi_n))\mid\mathcal{F}_q^N\,\right]=0$ for any $q<p\leq n$, so that:

$$\mathbb{E}\left[\,A_q(g,N_q)\mathcal{V}_n(\varphi_n,N)\mathcal{V}_q(\varphi_q,N)\,\right]=\sum_{0\leq p,s\leq q}\mathbb{E}\left[\,\tfrac{1}{\sqrt{N_pN_s}}A_q(g,N_q)V_p^{N_p}(D_{p,n}(\varphi_n))V_s^{N_s}(D_{s,q}(\varphi_q))\,\right]\ .$$

As $|A_q(g,N_q)|\leq 1$, one can use Lemma Appendix A.1 (iv), to obtain the upper bound

$$\begin{aligned}\sum_{0\leq p,s\leq q}\mathbb{E}\left|\,\tfrac{1}{\sqrt{N_pN_s}}A_q(g,N_q)V_p^{N_p}(D_{p,n}(\varphi_n))V_s^{N_s}(D_{s,q}(\varphi_q))\,\right| &\leq C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\sum_{0\leq p,s\leq q}\frac{\kappa^{n-p+q-s}}{\sqrt{N_pN_s}}\\ &\leq \frac{C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\kappa^{n-q}}{N_q}\ .\end{aligned}$$

ii) Again using $\mathbb{E}\left[\,V_p^{N_p}(D_{p,n}(\varphi_n))\mid\mathcal{F}_q^N\,\right]=0$ for $q<p\leq n$, we have

$$\mathbb{E}\left[\,A_q(g,N_q)\mathcal{V}_n(\varphi_n,N)\mathcal{R}_q(\varphi_q,N)\,\right]=\sum_{0\leq p,s\leq q}\mathbb{E}\left[\,\tfrac{1}{\sqrt{N_p}}A_q(g,N_q)V_p^{N_p}(D_{p,n}(\varphi_n))R_{s+1}^{N_s}(D_{s,q}(\varphi_q))\,\right]\ .$$

By $|A_q(g,N_q)|\leq 1$ and Lemma Appendix A.1(v) we have the upper-bound

$$\sum_{0\leq p,s\leq q}\mathbb{E}\left|\tfrac{1}{\sqrt{N_p}}A_q(g,N)V_p^{N_p}(D_{p,n}(\varphi_n))R_{s+1}^{N_s}(D_{s,q}(\varphi_q))\right|\leq C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\sum_{0\leq p,s\leq q}\tfrac{\kappa^{n-p+q-s}}{\sqrt{N_p}N_s}$$
$$\leq \tfrac{C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\kappa^{n-q}}{N_q^{3/2}}\ .$$

iii) By $|A_q(g,N_q)|\leq 1$ and Lemma Appendix A.1(v), we have the upper bound

$$\left|\mathbb{E}\left[\,A_q(g,N_q)\mathcal{V}_q(\varphi_q,N)\mathcal{R}_n(\varphi_n,N)\,\right]\right|\leq\sum_{p=0}^{n-1}\sum_{s=0}^{q}\mathbb{E}\left|\tfrac{1}{\sqrt{N_s}}A_q(g,N_q)R_{p+1}^N(D_{p,n}(\varphi_n))V_s^N(D_{s,q}(\varphi_q))\right|$$
$$\leq C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\sum_{p=0}^{n-1}\sum_{s=0}^{q}\tfrac{\kappa^{n-p+q-s}}{\sqrt{N_s}N_p}\leq\tfrac{C\|\varphi_n\|_\infty\|\varphi_q\|_\infty}{\sqrt{N_q}N_n}\ .$$

iv) Working as in (iii), by $|A_q(g,N_q)|\leq 1$ and Lemma Appendix A.1(vi), we have

$$\left|\mathbb{E}\left[\,A_q(g,N_q)\mathcal{R}_q(\varphi_q,N)\mathcal{R}_n(\varphi_n,N)\,\right]\right|\leq\sum_{p=0}^{n-1}\sum_{s=0}^{q-1}\mathbb{E}\left|\,A_q(g,N_q)R_{p+1}^N(D_{p,n}(\varphi_n))R_{s+1}^N(D_{s,q}(\varphi_q))\,\right|$$
$$\leq C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\sum_{p=0}^{n-1}\sum_{s=0}^{q}\tfrac{\kappa^{n-p+q-s}}{N_pN_s}\leq\tfrac{C\|\varphi_n\|_\infty\|\varphi_q\|_\infty}{N_nN_q}\ .$$

□

Lemma Appendix A.3. *Assume (A1-2). There exists a $C<+\infty$, such that for any $n\geq 0$, $1\leq r<+\infty$ and $\varphi\in\mathcal{B}_b(E)$:*

$$\|\overline{A}_n(\varphi,N_n)\|_r\leq\tfrac{C\|\varphi\|_\infty}{\sqrt{N_n}}.$$

Proof. The result is standard, but we give the proof for completeness. We have

$$\|\overline{A}_n(\varphi,N_n)\|_r=\left\|\tfrac{\eta_n^{N_n}(\varphi G_n)-\eta_n(\varphi G_n)}{\eta_n^{N_n}(G_n)}+\tfrac{\eta_n(\varphi G_n)}{\eta_n^{N_n}(G_n)\eta_n(G_n)}[\eta_n(G_n)-\eta_n^{N_n}(G_n)]\right\|_r\ .$$

Application of Minkowski, (A1) and [7, Theorem 7.4.4] complete the proof. □

Proposition Appendix A.1. *Assume (A1-2). There exist a $C<+\infty$, $\kappa\in(0,1)$ such that for any $n>q\geq 0$ and $\varphi_n,\varphi_q,f,g\in\mathcal{B}_b(E)$, $\|g\|_\infty=\|f\|_\infty=1$:*

$$\left|\mathbb{E}\left[\,A_n(g,N_n)A_q(f,N_q)(\eta_n^{N_n}-\eta_n)(\varphi_n)(\eta_q^{N_q}-\eta_q)(\varphi_q)\,\right]\right|\leq C\|\varphi_n\|_\infty\|\varphi_q\|_\infty\left(\tfrac{\kappa^{n-q}}{N_q}+\tfrac{1}{N_q^{1/2}N_n}\right)$$

Proof. From the definition of A_n, $\overline{A}_n$ we have that

$$\mathbb{E}\left[\,A_n(g,N_n)A_q(f,N_q)(\eta_n^{N_n}-\eta_n)(\varphi_n)(\eta_q^{N_q}-\eta_q)(\varphi_q)\,\right]=$$
$$\Delta_{1,q,n}(f,g,\varphi_q,\varphi_n,N_q,N_n)+\tfrac{\eta_n(gG_n)}{\eta_n(G_n)}\,\Delta_{2,q,n}(f,\varphi_q,\varphi_n,N_q,N_n)$$

where we have defined

$$\Delta_{1,q,n}(f,g,\varphi_q,\varphi_n,N_q,N_n)=\mathbb{E}\left[\,\overline{A}_n(g,N_n)A_q(f,N_q)(\eta_n^{N_n}-\eta_n)(\varphi_n)(\eta_q^{N_q}-\eta_q)(\varphi_q)\,\right]\ ,$$
$$\Delta_{2,q,n}(f,\varphi_q,\varphi_n,N_q,N_n)=\mathbb{E}\left[\,A_q(f,N_q)(\eta_n^{N_n}-\eta_n)(\varphi_n)(\eta_q^{N_q}-\eta_q)(\varphi_q)\,\right]\ .$$

By Lemma Appendix A.2 and the fact that $\frac{\eta_n(|g|G_n)}{\eta_n(G_q)} \leq 1$, we have that

$$\left| \frac{\eta_n(gG_n)}{\eta_n(G_n)} \Delta_{2,q,n}(f, \varphi_q, \varphi_n, N_q, N_n) \right| \leq C\|\varphi_n\|_\infty \|\varphi_q\|_\infty \left(\frac{\kappa^{n-q}}{N_q} + \frac{\kappa^{n-q}}{N_q^{3/2}} + \frac{1}{\sqrt{N_q} N_n} + \frac{1}{N_n N_q} \right) .$$

Thus we concentrate on $\Delta_{1,q,n}(f, g, \varphi_n, \varphi_q, N_q, N_n)$. We have via (A.2):

$$\begin{aligned} &\Delta_{1,q,n}(f, g, \varphi_n, \varphi_q, N_q, N_n) = \\ &\quad = \mathbb{E}\left[\overline{A}_n(g, N_n) A_q(f, N_q) (\mathcal{V}_n(\varphi_n, N) + \mathcal{R}_n(\varphi_n, N))(\mathcal{V}_q(\varphi_q, N) + \mathcal{R}_q(\varphi_q, N)) \right] . \end{aligned}$$

We will deal with each of the 4 terms on the R.H.S. separately.

We start with $\mathbb{E}[\, \overline{A}_n(g, N_n) A_q(f, N_q) \mathcal{V}_n(\varphi_n, N) \mathcal{V}_q(\varphi_q, N)\,]$ and work as follows,

$$\begin{aligned} &\left| \mathbb{E}[\, \overline{A}_n(g, N_n) A_q(f, N_q) \mathcal{V}_n(\varphi_n, N) \mathcal{V}_q(\varphi_q, N)\,] \right| \\ &\quad = \left| \sum_{p=0}^{n} \sum_{s=0}^{q} \mathbb{E}\left[\tfrac{1}{\sqrt{N_p N_s}} \overline{A}_n(g, N_n) A_q(f, N_q) V_p^{N_p}(D_{p,n}(\varphi_n)) V_s^{N_s}(D_{s,q}(\varphi_q)) \right] \right| \\ &\quad \leq \sum_{p=0}^{n} \sum_{s=0}^{q-1} \tfrac{1}{\sqrt{N_p N_s}} \|\overline{A}_n(g, N)\|_3 \, \|V_p^{N_p}(D_{p,n}(\varphi_n))\|_3 \, \|V_s^{N_s}(D_{s,q}(\varphi_q))\|_3 \\ &\quad \leq C\|\varphi_n\|_\infty \|\varphi_q\|_\infty \tfrac{1}{\sqrt{N_n}} \sum_{p=0}^{n} \sum_{s=0}^{q} \tfrac{\kappa^{n-p+q-s}}{\sqrt{N_p N_s}} \\ &\quad \leq \tfrac{C\|\varphi_n\|_\infty \|\varphi_q\|_\infty}{N_n \sqrt{N_q}} . \end{aligned}$$

where for the third line we have used $|A_q(g, N_q)| \leq 1$ and two applications of Hölder's inequality; for the forth line we have used Lemma Appendix A.1(ii) and Lemma Appendix A.3.

Using very similar calculations one can obtain the upper bounds,

$$\begin{aligned} \left| \mathbb{E}[\, \overline{A}_n(g, N_n) A_q(f, N_q) \mathcal{V}_n(\varphi_n, N) \mathcal{R}_q(\varphi_q, N)\,] \right| &\leq \tfrac{C\|\varphi_n\|_\infty \|\varphi_q\|_\infty}{N_n N_q} , \\ \left| \mathbb{E}[\, \overline{A}_n(g, N_n) A_q(f, N_q) \mathcal{R}_n(\varphi_n, N) \mathcal{V}_q(\varphi_q, N)\,] \right| &\leq \tfrac{C\|\varphi_n\|_\infty \|\varphi_q\|_\infty}{N_n^{3/2} N_q^{1/2}} , \\ \left| \mathbb{E}[\, \overline{A}_n(g, N) A_q(f, N) \mathcal{R}_n(\varphi_n) \mathcal{R}_q(\varphi_q)\,] \right| &\leq \tfrac{C\|\varphi_n\|_\infty \|\varphi_q\|_\infty}{N_n^{3/2} N_q} . \end{aligned}$$

The proof is now complete. □

References

[1] Beskos, A., Crisan, D. & Jasra, A. (2014). On the stability of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.*, **24**, 1396–1445.

[2] Bickel, P., Li, B. & Bengtsson, T. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics*, B. Clarke & S. Ghosal, Eds, 318–329, IMS.

[3] Chan, H. P. & Lai, T. L. (2013). A general theory of particle filters in hidden Markov models and some applications. *Ann. Statist.*, **41**, 2877-2904.

[4] Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.*, **32**, 2385–2411.

[5] Ciarlet, P. G. (1978). *The Finite Element Method for Elliptic Problems*. Elsevier: Amsterdam.

[6] Cliffe, K. A., Giles, M. B., Scheichl, R. & Teckentrup, A. L. (2011). Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comp. Visual. Sci.*, **14**, 3–15.

[7] Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.

[8] Del Moral, P. (2013). *Mean Field Simulation for Monte Carlo Integration* Chapman & Hall: London.

[9] Del Moral, P., Doucet, A. & Jasra, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.

[10] Del Moral, P., Doucet, A. & Jasra, A. (2012). On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, **18**, 252–278.

[11] Del Moral, P., Jasra, A., Law, K.J.H. & Zhou, Y. (2016). Multilevel SMC samplers for normalizing constants. Technical Report, NUS.

[12] Douc, R. & Moulines, E. (2008). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Ann. Statist.*, **36**, 2344–2376.

[13] Ern, A. (2004). *Theory and Practice of Finite Elements*. Springer, New York.

[14] Giles, M. B. (2008). Multilevel Monte Carlo path simulation. *Op. Res.*, **56**, 607-617.

[15] Heinrich, S. (1998). Monte Carlo complexity of parametric integration. *J. Complex.*, **14**, 151–175.

[16] Hoang, V., Schwab, C. & Stuart, A. (2013). Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Prob.*, **29**, 085010.

[17] Jasra, A., Law, K.J.H. & Zhou, Y. (2016). Forward and Inverse Uncertainty Quantification using Multilevel Monte Carlo Algorithms for an Elliptic Nonlocal Equation. Work in progress.

[18] Ketelsen, C., Scheichl, R. & Teckentrup, A. L. (2013). A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. arXiv preprint arXiv:1303.7343.

[19] Martin, J. S., Jasra, A. & McCoy, E. (2013). Inference for a class of partially observed point process. *Ann. Inst. Stat. Math.*, **65**, 413–437.

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